

SPRINGER  
REFERENCE

Hartmut Grassl  
William H. K. Lee  
*Section Editors*

Robert A. Meyers  
*Editor-in-Chief*

VOLUME 1

# Extreme Environmental Events

## Complexity in Forecasting and Early Warning

Selected entries from the Encyclopedia  
of Complexity and Systems Science

 Springer

Extreme Environmental Events  
Complexity in Forecasting and Early Warning

---

This book consists of selections from the  
*Encyclopedia of Complexity and Systems Science*  
edited by Robert A. Meyers,  
published by Springer New York in 2009.

Robert A. Meyers (Ed.)

# Extreme Environmental Events

Complexity in Forecasting and Early Warning

With 661 Figures and 51 Tables

**ROBERT A. MEYERS**, Ph. D.  
Editor-in-Chief  
RAMTECH LIMITED  
122 Escalle Lane  
Larkspur, CA 94939  
USA  
robert.meyers@ramtechlimited.org

Library of Congress Control Number: 2010938141

**ISBN: 978-1-4419-7695-6**

This publication is available also as:

Print publication under ISBN: 978-1-4419-7694-9 and

Print and electronic bundle under ISBN 978-1-4419-7696-3

© 2011 SpringerScience+Business Media, LLC.

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC., 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

This book consists of selections from the *Encyclopedia of Complexity and Systems Science* edited by Robert A. Meyers, published by Springer New York in 2009.

springer.com

Printed on acid free paper

# Preface

Extreme Environmental Events is an authoritative single source for understanding and applying the basic tenets of complexity and systems theory as well as the tools and measures for analyzing complex systems in understanding, predicting, monitoring and evaluating major phenomena and natural disasters affecting life on earth. These phenomena are earthquakes, tsunamis, volcanoes, climate change and weather. Early warning, damage, and the immediate response of human populations to these phenomena are also covered from a complexity (nonlinear) viewpoint. The content is presented in 61 articles and 1250 pages written by 110 of the world's experts in each field.

Extreme Environmental Events is written for an audience of advanced university undergraduate and graduate students, professors, and professionals in a wide range of fields including earth sciences, climatology, sociology, mathematics, physics and engineering. Each article was selected and peer reviewed by one of our Section Editors with advice and consultation provided by our Board Members and Editor-in-Chief. This level of coordination assures that the reader can have a level of confidence in the relevance and accuracy of the information far exceeding that generally found on the World Wide Web or any print publication. Accessibility is also a priority and for this reason each article includes a glossary of important terms and a concise definition of the subject. A list of the 61 articles and authors is presented on pages XV through XVII and a listing of the articles by section is shown on pages VII to VIII. A summary, perspective and roadmap for the articles on earthquakes, tsunamis and volcanoes is presented on pages 68 to 78. Also, a summary, perspective and roadmap for the articles on climate modeling, global warming and weather prediction is presented on pages 66 to 67.

Complex systems are systems that comprise many interacting parts with the ability to generate a new quality of collective behavior through self-organization, e.g. the spontaneous formation of temporal, spatial or functional structures. They are therefore adaptive as they evolve and may contain self-driving feedback loops. Thus, complex systems are much more than a sum of their parts. Complex systems are often characterized as having extreme sensitivity to initial conditions as well as emergent behavior that are not readily predictable or even completely deterministic. The conclusion is that a reductionist (bottom-up) approach is often an incomplete description of a phenomenon. This recognition, that the collective behavior of the whole system cannot be simply inferred from the understanding of the behavior of the individual components, has led to many new concepts and sophisticated mathematical and modeling tools for application to extreme environmental phenomena. These tools include fractals, cellular automata, solitons game theory, network theory and statistical physics.

## Acknowledgments

I wish to thank Springer management David Packer, Lydia Mueller and Sylvia Blago for selecting this project for publication and providing valuable advise and counsel.

Robert A. Meyers  
Editor in Chief  
Larkspur, California  
August 2010

# Sections

## **Climate Modeling, Global Warming and Weather Prediction,**

**Section Editors: Hartmut Grassl, Brian Dangerfield, and Marilda Sotomayor**

Abrupt Climate Change Modeling  
Climate Change and Agriculture  
Climate Change and Human Health  
Climate Change, Economic Costs of  
Climate Modeling, Global Warming and Weather Prediction, Introduction to  
Cryosphere Models  
Dynamic Games with an Application to Climate Change Models  
Regional Climate Models: Linking Global Climate Change to Local Impacts  
Single Column Modeling of Atmospheric Boundary Layers  
and the Complex Interactions with the Land Surface  
System Dynamics Models of Environment, Energy and Climate Change

## **Complexity in Earthquakes, Tsunamis, and Volcanoes, and Forecast,**

**Section Editors: William H. K. Lee, Daniel ben-Avraham, Shlomo Havlin,  
Mohamed A. Helal, Muhammad Sahimi, and M. Cristina Marchetti**

Brittle Tectonics: A Non-linear Dynamical System  
Complexity in Earthquakes, Tsunamis, and Volcanoes, and Forecast, Introduction to  
Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of  
Earthquake Clusters over Multi-dimensional Space, Visualization of  
Earthquake Damage: Detection and Early Warning in Man-Made Structures  
Earthquake Early Warning System in Southern Italy  
Earthquake Engineering, Non-linear Problems in  
Earthquake Forecasting and Verification  
Earthquake Location, Direct, Global-Search Methods  
Earthquake Magnitude  
Earthquake Monitoring and Early Warning Systems  
Earthquake Networks, Complex  
Earthquake Nucleation Process  
Earthquake Occurrence and Mechanisms, Stochastic Models for  
Earthquake Scaling Laws  
Earthquake Source: Asymmetry and Rotation Effects  
Earthquake Source Parameters, Rapid Estimates for Tsunami Warning  
Earthquakes, Dynamic Triggering of  
Earthquakes, Electromagnetic Signals of  
Earth's Crust and Upper Mantle, Dynamics of Solid-Liquid Systems in  
Fractals in Geology and Geophysics  
Geo-Complexity and Earthquake Prediction  
GPS: Applications in Crustal Deformation Monitoring

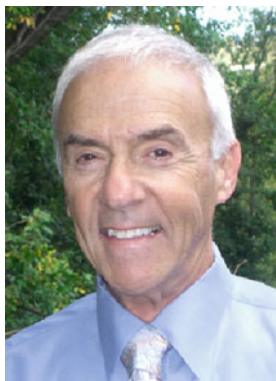
Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures  
Infrasound from Earthquakes, Tsunamis and Volcanoes  
Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of Percolation, and Faults and Fractures in Rock  
Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation  
Seismic Wave Propagation in Media with Complex Geometries, Simulation of Seismic Waves in Heterogeneous Earth, Scattering of  
Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates  
Space Seismicity, Statistical Physics Approaches to  
Slug Flow: Modeling in a Conduit and Associated Elastic Radiation  
Solitons, Tsunamis and Oceanographical Applications of  
Submarine Landslides and Slow Earthquakes: Monitoring Motion with GPS and Seafloor Geodesy  
Tomography, Seismic  
Tsunami Earthquakes  
Tsunami Forecasting and Warning  
Tsunami Inundation, Modeling of  
Tsunamis, Inverse Problem of  
Volcanic Eruptions, Explosive: Experimental Insights  
Volcanic Eruptions: Cyclicity During Lava Dome Growth  
Volcanic Eruptions: Stochastic Models of Occurrence Patterns  
Volcanic Hazards and Early Warning  
Volcano Seismic Signals, Source Quantification of  
Volcanoes, Non-linear Processes in  
Wedge Mechanics: Relation With Subduction Zone Earthquakes and Tsunamis

**Complexity in Human Reaction to Extreme Events,  
Section Editors: Boris Kerner and Andrzej Nowak**

Evacuation as a Communication and Social Phenomenon  
Evacuation Dynamics: Empirical Results, Modeling and Applications  
Extreme Events in Socio-economic and Political Complex Systems, Predictability of  
Pedestrian, Crowd and Evacuation Dynamics



## About the Editor-in-Chief



### **Robert A. Meyers**

President: RAMTECH Limited  
Manager, Chemical Process Technology, TRW Inc.  
Post-doctoral Fellow: California Institute of Technology  
Ph. D. Chemistry, University of California at Los Angeles  
B. A., Chemistry, California State University, San Diego

### **Biography**

Dr. Meyers has worked with more than 25 Nobel laureates during his career.

### **Research**

Dr. Meyers was Manager of Chemical Technology at TRW (now Northrop Grumman) in Redondo Beach, CA and is now President of RAMTECH Limited. He is co-inventor of the Gravimelt process for desulfurization and demineralization of coal for air pollution and water pollution control. Dr. Meyers is the inventor of and was project manager for the DOE-sponsored Magnetohydrodynamics Seed Regeneration Project which has resulted in the construction and successful operation of a pilot plant for production of potassium formate, a chemical utilized for plasma electricity generation and air pollution control. Dr. Meyers managed the pilot-scale DoE project for determining the hydrodynamics of synthetic fuels. He is a co-inventor of several thermo-oxidative stable polymers which have achieved commercial success as the GE PEI, Upjohn Polyimides and Rhone-Polenc bismaleimide resins. He has also managed projects for photochemistry, chemical lasers, flue gas scrubbing, oil shale analysis and refining, petroleum analysis and refining, global change measurement from space satellites, analysis and mitigation (carbon dioxide and ozone), hydrometallurgical refining, soil and hazardous waste remediation, novel polymers synthesis, modeling of the economics of space transportation systems, space rigidizable structures and chemiluminescence-based devices.

He is a senior member of the American Institute of Chemical Engineers, member of the American Physical Society, member of the American Chemical Society and serves on the UCLA Chemistry Department Advisory Board. He was a member of the joint USA-Russia working group on air pollution control and the EPA-sponsored Waste Reduction Institute for Scientists and Engineers.

Dr. Meyers has more than 20 patents and 50 technical papers. He has published in primary literature journals including *Science* and the *Journal of the American Chemical Society*, and is listed in *Who's Who in America* and *Who's Who in the World*. Dr. Meyers' scientific achievements have been reviewed in feature articles in the popular press in publications such as *The New York Times Science Supplement* and *The Wall Street Journal* as well as more specialized publications such as *Chemical Engineering and Coal Age*. A public service film was produced by the Environmental Protection Agency of Dr. Meyers' chemical desulfurization invention for air pollution control.

### **Scientific Books**

Dr. Meyers is the author or Editor-in-Chief of 12 technical books one of which won the Association of American Publishers Award as the best book in technology and engineering.

### **Encyclopedias**

Dr. Meyers conceived and has served as Editor-in-Chief of the Academic Press (now Elsevier) *Encyclopedia of Physical Science and Technology*. This is an 18-volume publication of 780 twenty-page articles written to an audience of university students and practicing professionals. This encyclopedia, first published in 1987, was very successful, and because of this, was revised and reissued in 1992 as a second edition. The Third Edition was published in 2001 and is now on-line. Dr. Meyers has completed two editions of the *Encyclopedia of Molecular Cell Biology and Molecular Medicine* for Wiley VCH publishers (1995 and 2004). These cover molecular and cellular level genetics, biochemistry, pharmacology, diseases and structure determination as well as cell biology. His eight-volume *Encyclopedia of Environmental Analysis and Remediation* was published in 1998 by John Wiley & Sons and his 15-volume *Encyclopedia of Analytical Chemistry* was published in 2000, also by John Wiley & Sons, all of which are available on-line.

## Editorial Board Members



PROFESSOR BENOIT B. MANDELBROT  
Sterling Professor Emeritus of Mathematical Sciences at  
Yale University  
1993 Wolf Prize for Physics and the  
2003 Japan Prize for Science and Technology  
*Current interests include:* seeking a measure of order in  
physical, mathematical or social phenomena that are  
characterized by abundant data but wild variability.



RICHARD E. STEARNS  
1993 Turing Award for foundations  
of computational complexity  
*Current interests include:* computational complexity,  
automata theory, analysis of algorithms, and game theory.



MARIO J. MOLINA  
1995 Nobel Prize in Chemistry for atmospheric  
chemistry, particularly the formation and decomposition  
of ozone  
*Current interests include:* atmospheric chemical processes,  
and science-policy issues related to urban and regional air  
pollution and to global change.



STEPHEN WOLFRAM  
Founder and CEO, Wolfram Research  
Creator, Mathematica®  
Author, *A New Kind of Science*



JOSEPH P. S. KUNG  
Professor  
Department of Mathematics  
University of North Texas



WILLIAM H. K. LEE  
Scientist Emeritus  
US Geological Survey  
Menlo Park, CA 94025, USA

## Section Editors

### Climate Modeling, Global Warming and Weather Prediction



HARTMUT GRASSL  
Professor emeritus, Hamburg University  
Former Director of the Max Planck Institute of  
Meteorology, Hamburg  
Former Director World Climate Research Program  
1994–1999

### Complexity in Earthquakes, Tsunamis and Volcanoes and Forecast



WILLIAM H. K. LEE  
Scientist Emeritus, US Geological Survey, Menlo Park

## Contributing Section Editors



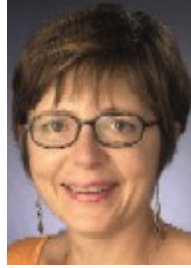
DANIEL BEN-AVRAHAM  
Professor  
Department of Physics  
Clarkson University



BRIAN DANGERFIELD  
Professor of Systems Modelling & Executive Editor  
System Dynamics Review Centre for OR & Applied  
Statistics Salford Business School  
Faculty of Business, Law & the Built Environment  
University of Salford



SHLOMO HAVLIN  
Professor  
Department of Physics  
Bar Ilan University



M. CRISTINA MARCHETTI  
William R. Kenan, Jr. Professor of Physics  
Physics Department  
Syracuse University



MOHAMED A. HELAL  
Professor  
Department of Mathematics  
Faculty of Science  
University of Cairo



ANDRZEJ NOWAK  
Director of the Center for Complex Systems  
University of Warsaw  
Assistant Professor, Psychology Department  
Florida Atlantic University



BORIS KERNER  
Head of "Traffic"  
DaimlerChrysler AG



MUHAMMAD SAHIMI  
Professor of Chemical Engineering  
and Materials Science  
University of Southern California



MARILDA SOTOMAYOR  
Professor  
Department of Economics  
University of São Paulo, Brazil  
Department of Economics  
Brown University, Providence

# Table of Contents

<b>Abrupt Climate Change Modeling</b>	
<i>Gerrit Lohmann</i> . . . . .	1
<b>Brittle Tectonics: A Non-linear Dynamical System</b>	
<i>Christopher H. Scholz</i> . . . . .	22
<b>Climate Change and Agriculture</b>	
<i>Cynthia Rosenzweig</i> . . . . .	31
<b>Climate Change and Human Health</b>	
<i>Hartmut Grassl</i> . . . . .	42
<b>Climate Change, Economic Costs of</b>	
<i>Richard S. J. Tol</i> . . . . .	52
<b>Climate Modeling, Global Warming and Weather Prediction, Introduction to</b>	
<i>Hartmut Grassl</i> . . . . .	66
<b>Complexity in Earthquakes, Tsunamis, and Volcanoes, and Forecast, Introduction to</b>	
<i>William H. K. Lee</i> . . . . .	68
<b>Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of</b>	
<i>Rowena Lohman</i> . . . . .	79
<b>Cryosphere Models</b>	
<i>Roger G. Barry</i> . . . . .	95
<b>Dynamic Games with an Application to Climate Change Models</b>	
<i>Prajit K. Dutta</i> . . . . .	109
<b>Earthquake Clusters over Multi-dimensional Space, Visualization of</b>	
<i>David A. Yuen, Witold Dzwinel, Yehuda Ben-Zion, Ben Kadlec</i> . . . . .	126
<b>Earthquake Damage: Detection and Early Warning in Man-Made Structures</b>	
<i>Maria I. Todorovska</i> . . . . .	150
<b>Earthquake Early Warning System in Southern Italy</b>	
<i>Aldo Zollo, Giovanni Iannaccone, Vincenzo Convertito, Luca Elia, Iunio Iervolino, Maria Lancieri, Anthony Lomax, Claudio Martino, Claudio Satriano, Emanuel Weber, Paolo Gasparini</i> . . . . .	175
<b>Earthquake Engineering, Non-linear Problems in</b>	
<i>Mihailo D. Trifunac</i> . . . . .	201
<b>Earthquake Forecasting and Verification</b>	
<i>James R. Holliday, John B. Rundle, Donald L. Turcotte</i> . . . . .	218
<b>Earthquake Location, Direct, Global-Search Methods</b>	
<i>Anthony Lomax, Alberto Michelini, Andrew Curtis</i> . . . . .	230
<b>Earthquake Magnitude</b>	
<i>Peter Bormann, Joachim Saul</i> . . . . .	255

<b>Earthquake Monitoring and Early Warning Systems</b>	
<i>William H. K. Lee, Yih-Min Wu</i> . . . . .	278
<b>Earthquake Networks, Complex</b>	
<i>Sumiyoshi Abe, Norikazu Suzuki</i> . . . . .	312
<b>Earthquake Nucleation Process</b>	
<i>Yoshihisa Iio</i> . . . . .	320
<b>Earthquake Occurrence and Mechanisms, Stochastic Models for</b>	
<i>David Vere-Jones</i> . . . . .	338
<b>Earthquake Scaling Laws</b>	
<i>Raul Madariaga</i> . . . . .	364
<b>Earthquake Source: Asymmetry and Rotation Effects</b>	
<i>Roman Teisseyre</i> . . . . .	383
<b>Earthquake Source Parameters, Rapid Estimates for Tsunami Warning</b>	
<i>Barry Hirshorn, Stuart Weinstein</i> . . . . .	406
<b>Earthquakes, Dynamic Triggering of</b>	
<i>Stephanie G. Prejean, David P. Hill</i> . . . . .	425
<b>Earthquakes, Electromagnetic Signals of</b>	
<i>Seiya Uyeda, Masashi Kamogawa, Toshiyasu Nagao</i> . . . . .	447
<b>Earth's Crust and Upper Mantle, Dynamics of Solid-Liquid Systems in</b>	
<i>Yasuko Takei</i> . . . . .	462
<b>Evacuation as a Communication and Social Phenomenon</b>	
<i>Douglas Goudie</i> . . . . .	484
<b>Evacuation Dynamics: Empirical Results, Modeling and Applications</b>	
<i>Andreas Schadschneider, Wolfram Klingsch, Hubert Klüpfel, Tobias Kretz, Christian Rogsch, Armin Seyfried</i> . . . . .	517
<b>Extreme Events in Socio-economic and Political Complex Systems, Predictability of</b>	
<i>Vladimir Keilis-Borok, Alexandre Soloviev, Allan Lichtman</i> . . . . .	551
<b>Fractals in Geology and Geophysics</b>	
<i>Donald L. Turcotte</i> . . . . .	568
<b>Geo-complexity and Earthquake Prediction</b>	
<i>Vladimir Keilis-Borok, Andrei Gabrielov, Alexandre Soloviev</i> . . . . .	573
<b>GPS: Applications in Crustal Deformation Monitoring</b>	
<i>Jessica Murray-Moraleda</i> . . . . .	589
<b>Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures</b>	
<i>P. Martin Mai</i> . . . . .	623
<b>Infrasound from Earthquakes, Tsunamis and Volcanoes</b>	
<i>Milton Garces, Alexis Le Pichon</i> . . . . .	663
<b>Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of</b>	
<i>Karin A. Dahmen, Yehuda Ben-Zion</i> . . . . .	680
<b>Pedestrian, Crowd and Evacuation Dynamics</b>	
<i>Dirk Helbing, Anders Johansson</i> . . . . .	697
<b>Percolation, and Faults and Fractures in Rock</b>	
<i>Pierre M. Adler, Jean-François Thovert, Valeri V. Mourzenko</i> . . . . .	717
<b>Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation</b>	
<i>Mie Ichihara, Takeshi Nishimura</i> . . . . .	731
<b>Regional Climate Models: Linking Global Climate Change to Local Impacts</b>	
<i>Daniela Jacob</i> . . . . .	753



<b>Seismic Wave Propagation in Media with Complex Geometries, Simulation of</b> <i>Heiner Igel, Martin Käser, Marco Stupazzini</i>	765
<b>Seismic Waves in Heterogeneous Earth, Scattering of</b> <i>Haruo Sato</i>	788
<b>Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space</b> <i>Gert Zöller, Sebastian Hainzl, Yehuda Ben-Zion, Matthias Holschneider</i>	805
<b>Seismicity, Statistical Physics Approaches to</b> <i>Didier Sornette, Maximilian J. Werner</i>	825
<b>Single Column Modeling of Atmospheric Boundary Layers and the Complex Interactions with the Land Surface</b> <i>Albert A. M. Holtslag, Gert-Jan Steeneveld</i>	844
<b>Slug Flow: Modeling in a Conduit and Associated Elastic Radiation</b> <i>Luca D'Auria, Marcello Martini</i>	858
<b>Solitons, Tsunamis and Oceanographical Applications of</b> <i>M. Lakshmanan</i>	873
<b>Submarine Landslides and Slow Earthquakes: Monitoring Motion with GPS and Seafloor Geodesy</b> <i>Benjamin A. Brooks, James H. Foster, Jeffrey J. McGuire, Mark Behn</i>	889
<b>System Dynamics Models of Environment, Energy and Climate Change</b> <i>Andrew Ford</i>	908
<b>Tomography, Seismic</b> <i>Jose Pujol</i>	928
<b>Tsunami Earthquakes</b> <i>Jascha Polet, H. Kanamori</i>	967
<b>Tsunami Forecasting and Warning</b> <i>Osamu Kamigaichi</i>	982
<b>Tsunami Inundation, Modeling of</b> <i>Patrick J. Lynett</i>	1008
<b>Tsunamis, Inverse Problem of</b> <i>Kenji Satake</i>	1022
<b>Volcanic Eruptions, Explosive: Experimental Insights</b> <i>Stephen J. Lane, Michael R. James</i>	1035
<b>Volcanic Eruptions: Cyclicity During Lava Dome Growth</b> <i>Oleg Melnik, R. Stephen J. Sparks, Antonio Costa, Alexei A. Barmin</i>	1082
<b>Volcanic Eruptions: Stochastic Models of Occurrence Patterns</b> <i>Mark S. Bebbington</i>	1104
<b>Volcanic Hazards and Early Warning</b> <i>Robert I. Tilling</i>	1135
<b>Volcano Seismic Signals, Source Quantification of</b> <i>Hiroyuki Kumagai</i>	1146
<b>Volcanoes, Non-linear Processes in</b> <i>Bernard Chouet</i>	1179
<b>Wedge Mechanics: Relation with Subduction Zone Earthquakes and Tsunamis</b> <i>Kelin Wang, Yan Hu, Jiangheng He</i>	1207
<b>List of Glossary Terms</b>	1219
<b>Index</b>	1225

# Contributors

ABE, SUMIYOSHI  
Mie University  
Tsu  
Japan  
Institut Supérieur des Matériaux et Mécaniques  
Le Mans  
France

ADLER, PIERRE M.  
UPMC-Sisyphé  
Paris  
France

BARMIN, ALEXEI A.  
Moscow State University  
Moscow  
Russia

BARRY, ROGER G.  
University of Colorado  
Boulder  
USA

BEBBINGTON, MARK S.  
Massey University  
Palmerston North  
New Zealand

BEHN, MARK  
Woods Hole Oceanographic Institution  
Woods Hole  
USA

BEN-ZION, YEHUDA  
University of Southern California  
Los Angeles  
USA

BORMANN, PETER  
GeoForschungsZentrum Potsdam  
Potsdam  
Germany

BROOKS, BENJAMIN A.  
University of Hawaii  
Honolulu  
USA

CHOUET, BERNARD  
US Geological Survey  
Menlo Park  
USA

CONVERTITO, VINCENZO  
Istituto Nazionale di Geofisica e Vulcanologia  
(RISSC-Lab)  
Napoli  
Italy

COSTA, ANTONIO  
University of Bristol  
Bristol  
UK  
Istituto Nazionale di Geofisica e Vulcanologia  
Naples  
Italy

CURTIS, ANDREW  
The University of Edinburgh  
Edinburgh  
United Kingdom

DAHMEN, KARIN A.  
University of Illinois at Urbana-Champaign  
Urbana  
USA

D'AURIA, LUCA  
Istituto Nazionale di Geofisica e Vulcanologia,  
Sezione di Napoli  
Naples  
Italy

DUTTA, PRAJIT K.  
Columbia University  
New York  
USA

DZWINEL, WITOLD  
AGH University of Sci. and Technol.  
Kraków  
Poland

ELIA, LUCA  
Istituto Nazionale di Geofisica e Vulcanologia  
(RISSC-Lab)  
Napoli  
Italy

FORD, ANDREW  
Washington State University, Pullman  
Washington  
USA

FOSTER, JAMES H.  
University of Hawaii  
Honolulu  
USA

GABRIELOV, ANDREI  
Purdue University  
West Lafayette  
USA

GARCES, MILTON  
HIGP, SOEST, University of Hawaii, Manoa  
Kailua–Kona  
USA

GASPARINI, PAOLO  
Università di Napoli “Federico II” (RISSC-Lab)  
Napoli  
Italy

GOUDIE, DOUGLAS  
James Cook University  
Townsville  
Australia

GRASSL, HARTMUT  
Max Planck Institute for Meteorology  
Hamburg  
Germany

HAINZL, SEBASTIAN  
GFZ German Research Centre for Geosciences  
Potsdam  
Germany

HE, JIANGHENG  
Geological Survey of Canada  
Sidney  
Canada

HELBING, DIRK  
ETH Zurich  
Zurich  
Switzerland  
Collegium Budapest  
Budapest  
Hungary

HILL, DAVID P.  
Volcano Hazards Program  
Menlo Park  
USA

HIRSHORN, BARRY  
NOAA/NWS/Pacific Tsunami Warning Center  
Ewa Beach  
USA

HOLLIDAY, JAMES R.  
University of California  
Davis  
USA

HOLSCHNEIDER, MATTHIAS  
University of Potsdam  
Potsdam  
Germany

HOLTSLAG, ALBERT A. M.  
Wageningen University  
Wageningen  
The Netherlands

HU, YAN  
University of Victoria  
Victoria  
Canada

IANNACCONE, GIOVANNI  
Istituto Nazionale di Geofisica e Vulcanologia  
(RISSC-Lab)  
Napoli  
Italy

ICHIHARA, MIE  
University of Tokyo  
Tokyo  
Japan

IERVOLINO, IUNIO  
Università di Napoli “Federico II”  
Napoli  
Italy

IGEL, HEINER  
Ludwig-Maximilians-University  
Munich  
Germany

IIO, YOSHIHISA  
Kyoto University  
Kyoto  
Japan

JACOB, DANIELA  
Max-Planck-Institute for Meteorology  
Hamburg  
Germany

JAMES, MICHAEL R.  
Lancaster University  
Lancaster  
UK

JOHANSSON, ANDERS  
ETH Zurich  
Zurich  
Switzerland

KADLEC, BEN  
University of Colorado  
Boulder  
USA

KAMIGAICHI, OSAMU  
Japan Meteorological Agency  
Tokyo  
Japan

KAMOGAWA, MASASHI  
Tokyo Gakugei University  
Koganei-shi  
Japan

KANAMORI, H.  
Caltech  
Pasadena  
USA

KÄSER, MARTIN  
Ludwig-Maximilians-University  
Munich  
Germany

KEILIS-BOROK, VLADIMIR  
University of California  
Los Angeles  
USA  
Russian Academy of Sciences  
Moscow  
Russia

KLINGSCH, WOLFRAM  
University of Wuppertal  
Wuppertal  
Germany

KLPFEL, HUBERT  
TraffGo HT GmbH  
Duisburg  
Germany

KRETZ, TOBIAS  
PTV Planung Transport Verkehr AG  
Karlsruhe  
Germany

KUMAGAI, HIROYUKI  
National Research Institute for Earth Science and Disaster  
Prevention  
Tsukuba  
Japan  
IAVCEI/IASPEI Joint Commission on Volcano  
Seismology  
Tsukuba  
Japan

LAKSHMANAN, M.  
Bharathidasan University  
Tiruchirapalli  
India

LANCIERI, MARIA  
Istituto Nazionale di Geofisica e Vulcanologia  
(RISSC-Lab)  
Napoli  
Italy

LANE, STEPHEN J.  
Lancaster University  
Lancaster  
UK

LEE, WILLIAM H. K.  
US Geological Survey  
Menlo Park  
USA

LE PICHON, ALEXIS  
CEA/DASE/LD  
Bruyères-le-Châtel  
France

LICHTMAN, ALLAN  
American University  
Washington D.C.  
USA

- LOHMANN, GERRIT  
Alfred Wegener Institute for Polar and Marine Research  
Bremerhaven  
Germany
- LOHMAN, ROWENA  
Cornell University  
Ithaca  
USA
- LOMAX, ANTHONY  
ALomax Scientific  
Mouans-Sartoux  
France
- LYNETT, PATRICK J.  
Texas A&M University  
College Station  
USA
- MADARIAGA, RAUL  
Laboratoire de Géologie  
Paris  
France
- MAI, P. MARTIN  
Institute of Geophysics, ETH  
Zrich  
Switzerland
- MARTINI, MARCELLO  
Istituto Nazionale di Geofisica e Vulcanologia, Sezione di  
Napoli  
Naples  
Italy
- MARTINO, CLAUDIO  
Università di Napoli "Federico II" (RISSC-Lab)  
Napoli  
Italy
- MCGUIRE, JEFFREY J.  
Woods Hole Oceanographic Institution  
Woods Hole  
USA
- MELNIK, OLEG  
Moscow State University  
Moscow  
Russia  
University of Bristol  
Bristol  
UK
- MICHELINI, ALBERTO  
Istituto Nazionale di Geofisica e Vulcanologia  
Roma  
Italy
- MOURZENKO, VALERI V.  
CNRS-LCD  
Chasseneuil du Poitou  
France
- MURRAY-MORALEDA, JESSICA  
US Geological Survey  
Menlo Park  
USA
- NAGAO, TOSHIYASU  
Tokai University  
Shizuoka  
Japan
- NISHIMURA, TAKESHI  
Tohoku University  
Sendai  
Japan
- POLET, JASCHA  
California State Polytechnic University  
Pomona  
USA
- PREJEAN, STEPHANIE G.  
Alaska Science Center  
Anchorage  
USA
- PUJOL, JOSE  
The University of Memphis  
Memphis  
USA
- ROGSCH, CHRISTIAN  
University of Wuppertal  
Wuppertal  
Germany
- ROSENZWEIG, CYNTHIA  
Columbia University  
New York  
USA
- RUNDLE, JOHN B.  
University of California  
Davis  
USA

SATAKE, KENJI  
University of Tokyo  
Tokyo  
Japan

SATO, HARUO  
Tohoku University  
Sendai-shi, Miyagi-ken  
Japan

SATRIANO, CLAUDIO  
Università di Napoli "Federico II" (RISSC-Lab)  
Napoli  
Italy

SAUL, JOACHIM  
GeoForschungsZentrum Potsdam  
Potsdam  
Germany

SCHADSCHNEIDER, ANDREAS  
Universität zu Köln  
Köln  
Germany  
Interdisziplinäres Zentrum für Komplexe Systeme  
Bonn  
Germany

SCHOLZ, CHRISTOPHER H.  
Columbia University  
New York  
USA

SEYFRIED, ARMIN  
Research Centre Jlich  
Jlich  
Germany

SOLOVIEV, ALEXANDRE  
Russian Academy of Sciences  
Moscow  
Russia  
The Abdus Salam International Center for Theoretical  
Physics  
Trieste  
Italy

SORNETTE, DIDIER  
Technology and Economics  
ETH Zurich  
Switzerland

SPARKS, R. STEPHEN J.  
University of Bristol  
Bristol  
UK

STEENEVELD, GERT-JAN  
Wageningen University  
Wageningen  
The Netherlands

STUPAZZINI, MARCO  
Politecnico di Milano  
Milano  
Italy

SUZUKI, NORIKAZU  
Nihon University  
Chiba  
Japan

TAKEI, YASUKO  
University of Tokyo  
Tokyo  
Japan

TEISSEYRE, ROMAN  
Polish Academy of Sciences  
Warsaw  
Poland

THOVERT, JEAN-FRANÇOIS  
CNRS-LCD  
Chasseneuil du Poitou  
France

TILLING, ROBERT I.  
US Geological Survey  
Menlo Park  
USA

TODOROVSKA, MARIA I.  
University of Southern California  
Los Angeles  
USA

TOL, J. RICHARD S.  
Economic and Social Research Institute  
Dublin  
Ireland  
Vrije Universiteit  
Amsterdam  
The Netherlands  
Vrije Universiteit  
Amsterdam  
The Netherlands  
Carnegie Mellon University  
Pittsburgh  
USA

TRIFUNAC, MIHAILO D.  
University of Southern California  
Los Angeles  
USA

TURCOTTE, DONALD L.  
University of California  
Davis  
USA

UYEDA, SEIYA  
Tokai University  
Shizuoka  
Japan

VERE-JONES, DAVID  
Statistical Research Associates and Victoria University  
Wellington  
New Zealand

WANG, KELIN  
Geological Survey of Canada  
Sidney  
Canada  
University of Victoria  
Victoria  
Canada

WEBER, EMANUEL  
Istituto Nazionale di Geofisica e Vulcanologia  
(RISSC-Lab)  
Napoli  
Italy

WEINSTEIN, STUART  
NOAA/NWS/Pacific Tsunami Warning Center  
Ewa Beach  
USA

WERNER, MAXIMILIAN J.  
Institute of Geophysics  
ETH Zurich  
Switzerland

WU, YIH-MIN  
National Taiwan University  
Taipei  
Taiwan

YUEN, DAVID A.  
University of Minnesota  
Minneapolis  
USA

ZÖLLER, GERT  
University of Potsdam  
Potsdam  
Germany

ZOLLO, ALDO  
Università di Napoli "Federico II" (RISSC-Lab)  
Napoli  
Italy

# Abrupt Climate Change Modeling

GERRIT LOHMANN

Alfred Wegener Institute for Polar and Marine Research,  
Bremerhaven, Germany

## Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[A Mathematical Definition](#)

[Earth System Modeling and Analysis](#)

[Example: Glacial-Interglacial Transitions](#)

[Example: Cenozoic Climate Cooling](#)

[Examples: Transient Growth](#)

[Future Directions](#)

[Bibliography](#)

## Glossary

**Atmosphere** The atmosphere is involved in many processes of abrupt climate change, providing a strong non-linearity in the climate system and propagating the influence of any climate forcing from one part of the globe to another. Atmospheric temperature, composition, humidity, cloudiness, and wind determine the Earth's energy fluxes. Wind affects the ocean's surface circulation and upwelling patterns. Atmospheric moisture transport determines the freshwater balance for the oceans, overall water circulation, and the dynamics of glaciers.

**Oceans** Because water has enormous heat capacity, oceans typically store 10–100 times more heat than equivalent land surfaces. The oceans exert a profound influence on climate through their ability to transport heat from one location to another. Changes in ocean circulation have been implicated in abrupt climate change of the past. Deglacial meltwater has freshened the North Atlantic and reduced the ability of the water to sink, inducing long-term coolings.

**Land surface** The reflective capacity of the land can change greatly, with snow or ice sheets reflecting up to 90% of the sunlight while dense forests absorb more than 90%. Changes in surface characteristics can also affect solar heating, cloud formation, rainfall, and surface-water flow to the oceans, thus feeding back strongly on climate.

**Cryosphere** The portion of the Earth covered with ice and snow, the cryosphere, greatly affects temperature. When sea ice forms, it increases the planetary re-

flective capacity, thereby enhancing cooling. Sea ice also insulates the atmosphere from the relatively warm ocean, allowing winter air temperatures to steeply decline and reduce the supply of moisture to the atmosphere. Glaciers and snow cover on land can also provide abrupt-change mechanisms. The water frozen in a glacier can melt if warmed sufficiently, leading to possibly rapid discharge, with consequent effects on sea level and ocean circulation. Meanwhile, snow-covered lands of all types maintain cold conditions because of their high reflectivity and because surface temperatures cannot rise above freezing until the snow completely melts.

**External factors** Phenomena external to the climate system can also be agents of abrupt climate change. For example, the orbital parameters of the Earth vary over time, affecting the latitudinal distribution of solar energy. Furthermore, fluctuations in solar output, prompted by sunspot activity or the effects of solar wind, as well as volcanoes may cause climate fluctuations.

**Climate time scales** The climate system is a composite system consisting of five major interactive components: the atmosphere, the hydrosphere, including the oceans, the cryosphere, the lithosphere, and the biosphere. All subsystems are open and non-isolated, as the atmosphere, hydrosphere, cryosphere and biosphere act as cascading systems linked by complex feedback processes. Climate refers to the average conditions in the Earth system that generally occur over periods of time, usually several decades or longer. This time scale is longer than the typical response time of the atmosphere. Parts of the other components of the Earth system (ice, ocean, continents) have much slower response times (decadal to millennial).

**Climate variables and forcing** State variables are temperature, rainfall, wind, ocean currents, and many other variables in the Earth system. In our notation, the variables are described by a finite set of real variables in a vector  $x(t) \in \mathbb{R}^n$ . The climate system is subject to two main external forcings  $F(x, t)$  that condition its behavior, solar radiation and the action of gravity. Since  $F(x, t)$  has usually a spatial dependence,  $F$  is also a vector  $\in \mathbb{R}^n$ . Solar radiation must be regarded as the primary forcing mechanism, as it provides almost all the energy that drives the climate system. The whole climate system can be regarded as continuously evolving, as solar radiation changes on diurnal, seasonal and longer time scales, with parts of the system leading or lagging in time. Therefore, the subsystems of the climate system are not always in equi-



librium with each other. Indeed, the climate system is a dissipative, highly non-linear system, with many instabilities.

**Climate models** are based on balances of energy, momentum, and mass, as well as radiation laws. There are several model categories, full circulation models, low-order models, and models of intermediate complexity. Climate models simulate the interactions of the atmosphere, oceans, land surface, and ice. They are used for a variety of purposes from study of the dynamics of the weather and climate system, past climate to projections of future climate.

#### **Global climate models or General circulation models**

(GCMs) The balances of energy, momentum, and mass are formulated in the framework of fluid dynamics on the rotating Earth. GCMs discretize the equations for fluid motion and energy transfer and integrate these forward in time. They also contain parametrization for processes – such as convection – that occur on scales too small to be resolved directly. The dimension of the state vector is in the order of  $n \sim 10^5 - 10^8$  depending on the resolution and complexity of the model.

**Model categories** In addition to complex numerical climate models, it can be of great utility to reduce the system to low-order, box, and conceptual models. This complementary approach has been successfully applied to a number of questions regarding feedback mechanisms and the basic dynamical behavior, e. g. [48,84]. In some cases, e. g. the stochastic climate model of Hasselmann [32], such models can provide a null hypothesis for the complex system. The transition from highly complex dynamical equations to a low-order description of climate is an important topic of research. In his book “Dynamical Paleoclimatology”, Saltzman [77] formulated a dynamical system approach in order to differentiate between fast-response and slow-response variables. As an alternative to this method, one can try to derive phenomenologically based concepts of climate variability, e. g. [21,43]. In between the comprehensive models and conceptual models, a wide class of “models of intermediate complexity” were defined [12].

#### **Earth-system models of intermediate complexity**

(EMICs) Depending on the nature of questions asked and the pertinent time scales, different types of models are used. There are, on the one extreme, conceptual models, and, on the other extreme, comprehensive models (GCMs) operating at a high spatial and temporal resolution. Models of intermediate complexity bridge the gap [12]. These models are successful in de-

scribing the Earth system dynamics including a large number of Earth system components. This approach is especially useful when considering long time scales where the complex models are computationally too expensive, e. g. [47]. Improvements in the development of coupled models of intermediate complexity have led to a situation where modeling a glacial cycle, even with prognostic atmospheric CO<sub>2</sub> is becoming possible.

**Climate simulation** A climate simulation is the output of a computer program that attempts to simulate the climate evolution under appropriate boundary conditions. Simulations have become a useful part of climate science to gain insight into the sensitivity of the system.

**Climate variability pattern** Climate variability is defined as changes in integral properties of the climate system. True understanding of climate dynamics and prediction of future changes will come only with an understanding of the Earth system as a whole, and over past and present climate. Such understanding requires identification of the patterns of climate variability and their relationships to known forcing. Examples for climate variability patterns are the North Atlantic Oscillation (NAO) or the El Niño-Southern Oscillation (ENSO).

**Abrupt climate change** One can define abrupt climate change in the time and frequency domain. (a) Time domain: Abrupt climate change refers to a large shift in climate that persists for years or longer, such as marked changes in average temperature, or altered patterns of storms, floods, or droughts, over a widespread area that takes place so rapidly that the natural system has difficulty adapting to it. In the context of past abrupt climate change, “rapidly” typically means on the order of a decade. (b) Frequency domain: An abrupt change means that the characteristic periodicity changes. Also the phase relation between certain climate variables may change in a relatively short time. For both types of changes examples will be provided.

**Regime shifts** are defined as rapid transitions from one state to another. In the marine environment, regimes may last for several decades, and shifts often appear to be associated with changes in the climate system. If the shifts occur regularly, they are often referred to as an oscillation (e. g., Atlantic Multi-decadal Oscillation, Pacific Decadal Oscillation). Similarly, one can define a regime shift in the frequency domain.

**Anthropogenic climate change** Beginning with the industrial revolution in the 1850s and accelerating ever since, the human consumption of fossil fuels has elevated CO<sub>2</sub> levels from a concentration of  $\sim 280$  ppm

to more than 380 ppm today. These increases are projected to reach more than 560 ppm before the end of the 21st century. As an example, a concomitant shift of ocean circulation would have serious consequences for both agriculture and fishing.

**Multiple equilibria** Fossil evidence and computer models demonstrate that the Earth's complex and dynamic climate system has more than one mode of operation. Each mode produces different climate patterns. The evidence of models and data analysis shows that the Earth's climate system has sensitive thresholds. Pushed past a threshold, the system can jump from one stable operating mode to a completely different one.

**Long-term climate statistics** Starting with a given initial state, the solutions  $x(t)$  of the equations that govern the dynamics of a non-linear system, such as the atmosphere, result in a set of long-term statistics. If all initial states ultimately lead to the same set of statistical properties, the system is ergodic or transitive. If, instead, there are two or more different sets of statistical properties, where some initial states lead to one set, while the other initial states lead to another, the system is called intransitive (one may call the different states regimes). If there are different sets of statistics that a system may assume in its evolution from different initial states through a long, but finite, period of time, the system is called almost intransitive [50,51,53]. In the transitive case, the equilibrium climate statistics are both stable and unique. Long-term climate statistics will give a good description of the climate. In the almost intransitive case, the system in the course of its evolution will show finite periods during which distinctly different climatic regimes prevail. The almost intransitive case arises because of internal feedbacks, or instabilities involving the different components of the climatic system. The climatic record can show rapid step-like shifts in climate variability that occur over decades or less, including climatic extremes (e. g. drought) that persist for decades.

**Feedbacks** A perturbation in a system with a negative feedback mechanism will be reduced whereas in a system with positive feedback mechanisms, the perturbation will grow. Quite often, the system dynamics can be reduced to a low-order description. Then, the growth or decay of perturbations can be classified by the systems' eigenvalues or the pseudospectrum. Consider the stochastic dynamical system

$$\frac{d}{dt}x(t) = f(x) + g(x)\xi + F(x, t), \quad (1)$$

where  $\xi$  is a stochastic process. The functions  $f, g$  de-

scribe the climate dynamics, in this case without explicit time dependence. The external forcing  $F(x, t)$  is generally time-, variable-, and space-dependent. In his theoretical approach, Hasselmann [32] formulated a linear stochastic climate model

$$\frac{d}{dt}x(t) = Ax + \sigma\xi + F(t), \quad (2)$$

with system matrix  $A \in \mathbb{R}^{n \times n}$ , constant noise term  $\sigma$ , and stochastic process  $\xi$ . Interestingly, many features of the climate system can be well described by (2), which is analogous to the Ornstein–Uhlenbeck process in statistical physics [89]. In the climate system, linear and non-linear feedbacks are essential for abrupt climate changes.

**Paleoclimate** Abrupt climate change is evident in model results and in instrumental records of the climate system. Much interest in the subject is motivated by the evidence in archives of extreme changes. Proxy records of paleoclimate are central to the subject of abrupt climate change. Available paleoclimate records provide information on many environmental variables, such as temperature, moisture, wind, currents, and isotopic compositions.

**Thermohaline circulation** stems from the Greek words “thermos” (heat) and “halos” (salt). The ocean is driven to a large extent by surface heat and freshwater fluxes. As the ocean is non-linear, it cannot be strictly separated from the wind-driven circulation. The expressions thermohaline circulation (THC) and meridional overturning circulation (MOC) in the ocean are quite often used as synonyms although the latter includes all effects (wind, thermal, haline forcing) and describes the ocean transport in meridional direction. Another related expression is the ocean conveyor belt. This metaphor is motivated by the fact that the North Atlantic is the source of the deep limb of a global ocean circulation system [10]. If North Atlantic surface waters did not sink, the global ocean circulation would cease, currents would weaken or be redirected. The resulting reorganization would reconfigure climate patterns, especially in the Atlantic Ocean. One fundamental aspect of this circulation is the balance of two processes: cooling of the deep ocean at high latitudes, and heating of deeper levels from the surface through vertical mixing.

### Definition of the Subject

The occurrence of abrupt change of climate at various time scales has attracted a great deal of interest for its theoretical and practical significance [2,3,9]. To some extent, a defini-

tion of what constitutes an abrupt climatic change depends on the sampling interval of the data being examined [28]. For the instrumental period covering approximately the last 100 years of annually or seasonally sampled data, an abrupt change in a particular climate variable will be taken to mean a statistically highly significant difference between adjacent 10-year sample means. In the paleoclimate context (i. e. on long time scales), an abrupt climate change can be in the order of decades to thousands of years. Since the climate dynamics can be often projected onto a limited number of modes or patterns of climate variability (e. g., [21,22]), the definition of abrupt climate change is also related to spatio-temporal patterns.

The concept of abrupt change of climate is therefore applied for different time scales. For example, changes in climatic regimes were described associated with surface temperature, precipitation, atmospheric circulation in North America during the 1920s and 1960s [19,75]. Sometimes, the term “climate jump” is used instead of “abrupt climate change”, e. g. [92]. Flohn [25] expanded the concept of abrupt climate change to include both singular events and catastrophes such as the extreme El Niño of 1982/1983, as well as discontinuities in paleoclimate indices taken from ice cores and other proxy data. In the instrumental record covering the last 150 years, there is a well-documented abrupt shift of sea surface temperature and atmospheric circulation features in the Northern Hemisphere in the mid-1970s, e. g. [22,67,88]. Some of the best-known and best-studied widespread abrupt climate changes started and ended during the last deglaciation, most pronounced at high latitudes.

In his classic studies of chaotic systems, Lorenz has proposed a deterministic theory of climate change with his concept of the “almost-intransitivity” of the highly non-linear climate systems. In this set of equations, there exists the possibility of multiple stable solutions to the governing equations, even in the absence of any variations in external forcing [51]. More complex models, e. g. [11,20] also demonstrated this possibility. On the other hand, variations in external forcing, such as the changes of incoming solar radiation, volcanic activity, deglacial meltwater, and increases of greenhouse gas concentration have also been proposed to account for abrupt changes in addition to climate intransitivity [9,25,38,41,49]. A particular climate change is linked to the widespread continental glaciation of Antarctica during the Cenozoic (65 Ma to present) at about 34 Ma, e. g. [93]. It should be noted that many facets of regional climate change are abrupt changes although the global means are rather smoothly changing.

Besides abrupt climate change as described in the time domain, we can find abrupt shifts in the frequency do-

main. A prominent example for an abrupt climate change in the frequency domain is the mid-Pleistocene transition or revolution (MPR), which is the last major “event” in a secular trend towards more intensive global glaciation that characterizes the last few tens of millions of years. The MPR is the term used to describe the transition between 41 ky (ky =  $10^3$  years) and 100 ky glacial-interglacial cycles which occurred about one million years ago (see a recent review in [61]). Evidence of this is provided by high-resolution oxygen isotope data from deep sea cores, e. g. [45,83].

Another example is the possibility of greenhouse gas-driven warming leading to a change in El Niño events. Modeling studies indicate that a strong enhancement of El Niño conditions in the future is not inconceivable [85]. Such a shift would have enormous consequences for both the biosphere and humans. The apparent phase shifts during the 1970s seems unique over this time period, and may thus represent a real climate shift although the available time series is probably too short to unequivocally prove that the shift is significant [90]. The inability to resolve questions of this kind from short instrumental time series provides one of the strongest arguments for extending the instrumental record of climate variability with well-dated, temporally finely resolved and rigorously calibrated proxy data.

## Introduction

One view of climate change is that the Earth’s climate system has changed gradually in response to both natural and human-induced processes. Researchers became intrigued by abrupt climate change when they discovered striking evidence of large, abrupt, and widespread changes preserved in paleoclimatic archives, the history of Earth’s climate recorded in tree rings, ice cores, sediments, and other sources. For example, tree rings show the frequency of droughts, sediments reveal the number and type of organisms present, and gas bubbles trapped in ice cores indicate past atmospheric conditions.

The Earth’s climate system is characterized by change on all time and space scales, and some of the changes are abrupt even relative to the short time scales of relevance to human societies. Paleoclimatic records show that abrupt climate changes have affected much or all of the Earth repeatedly over the last ice-age cycle as well as earlier – and these changes sometimes have occurred in periods as short as a few years, as documented in Greenland ice cores. Perturbations at northern high latitudes were spectacularly large: some had temperature increases of up to 10–20°C and a local doubling of precipitation within decades.

In the frequency domain, abrupt climate shifts are due to changes in the dominant oscillations (as in the case of the MPR), or due to a shift in the phase between different climate signals. As an example, the phase between the Indian Monsoon and ENSO exhibits significant shifts for the past 100 years [59].

The period of regular instrumental records of global climate is relatively short (100–200 years). Even so, this record shows many climatic fluctuations, some abrupt or sudden, as well as slow drifts in climate. Climatic changes become apparent on many temporal and spatial scales. Most abrupt climate changes are regional in their spatial extent. However, regional changes can have remote impacts due to atmospheric and oceanic teleconnections. Some of these shifts may be termed abrupt or sudden in that they represent relatively rapid changes in otherwise comparatively stable conditions, but they can also be found superimposed on other much slower climatic changes.

The definition of “abrupt” or “rapid” climate changes is therefore necessarily subjective, since it depends in large measure on the sample interval used in a particular study and on the pattern of longer-term variation within which the sudden shift is embedded. It is therefore useful to avoid a too general approach, but instead to focus on different types of rapid transitions as they are detected and modeled for different time periods. Although distinctions between types are somewhat arbitrary, together they cover a wide range of shifts in dominant climate mode on time scales ranging from the Cenozoic (the last 65 millions of years) to the recent and future climate.

## A Mathematical Definition

### Time Domain

Abrupt climate change is characterized by a transition of the climate system into a different state (of temperature, rainfall, and other variables) on a time scale that is faster than variations in the neighborhood (in time). Abrupt climate change could be related to a forcing or internally generated. Consider  $x(t) \in \mathbb{R}^n$  as a multi-dimensional climate state variable (temperature, rainfall, and other variables). We define an abrupt climate shift of degree  $\epsilon$  and amplitude  $B$ , if

$$\frac{d}{dt}x_i(t) \quad \text{can be approximated by a function} \quad \frac{B}{\pi} \frac{\epsilon}{x_i^2 + \epsilon^2} \quad (3)$$

for one  $i \in \{1, \dots, n\}$  in a time interval  $[t_1, t_2]$ . The case  $\epsilon \rightarrow 0$  is called instantaneous climate shift, i. e.  $x_i(t)$  can be

approximated by the Heaviside step function. The degree of approximation can be specified by a proper norm.

An alternative way of defining an abrupt climate shift is through the identification of probable breaks in a time series (e. g., the surface temperature series). The formulation of a two-phase regression (TPR) test, e. g. [55,79], describing a series  $x(t)$  is given by

$$x(t) = \mu_1 + \alpha_1 t + \epsilon(t) \quad \text{for } t \leq c \quad (4)$$

$$x(t) = \mu_2 + \alpha_2 t + \epsilon(t) \quad \text{for } t > c. \quad (5)$$

Under the null hypothesis of no changepoint, the two phases of the regression should be statistically equivalent and both the difference in means  $\mu_{1,2}$ , and the difference in slopes,  $\alpha_{1,2}$ , should be close to zero for each possible changepoint  $c$ .

In a stochastic framework one may use an appropriate stochastic differential equation (Langevin equation)

$$\frac{d}{dt}x(t) = f(x) + g(x)\xi, \quad (6)$$

where  $\xi$  is a stationary stochastic process and the functions  $f, g: \mathbb{R}^n \rightarrow \mathbb{R}^n$  describe the climate dynamics. Abrupt climate change can be defined as a transition of a short period of time  $[t_1, t_2]$ , where the probability of an event is larger than a threshold. The properties of the random force are described through its distribution and its correlation properties at different times. In the Ornstein–Uhlenbeck process  $\xi$  is assumed to have a Gaussian distribution of zero average,

$$\langle \xi(t) \rangle = 0 \quad (7)$$

and to be  $\delta$ -correlated in time,

$$\langle \xi(t)\xi(t + \tau) \rangle = \delta(\tau). \quad (8)$$

The brackets indicate an average over realizations of the random force. For a Gaussian process only the average and second moment need to be specified since all higher moments can be expressed in terms of the first two. Note that the dependence of the correlation function on the time difference  $\tau$  assumes that  $\xi$  is a stationary process.

The probability density  $p(x, t)$  for the variable  $x(t)$  in (6) obeys the Fokker–Planck equation

$$\partial_t p = -\frac{\partial}{\partial x} [f(x)p] + \frac{\partial}{\partial x} \left[ g(x) \frac{\partial}{\partial x} \{g(x)p\} \right]. \quad (9)$$

Its stationary probability density of (6) is given by

$$p_{st}(x) = \mathfrak{N} \exp \left( -2 \int_{x_0}^x \frac{f(y) - g(y)g'(y)}{g(y)^2} dy \right), \quad (10)$$

where  $\aleph$  is a normalization constant.  $g'(y)$  stands for the derivative of  $g$  with respect to its argument. The extrema  $x_m$  of the steady state density obey the equation

$$f(x_m) - g(x_m)g'(x_m) = 0 \quad (11)$$

for  $g(x_m) \neq 0$ . Here is the crux of the noise-induced transition phenomenon: one notes that this equation is not the same as the equation  $f(x_m) = 0$  that determines the steady states of the system in the absence of multiplicative noise. As a result, the most probable states of the noisy system need not coincide with the deterministic stationary states. More importantly, new solutions may appear or existing solutions may be destabilized by the noise. These are the changes in the asymptotic behavior of the system caused by the presence of the noise, e. g. [84].

### Climate Variability and Climate Change

The temporal evolution of climate can be expressed in terms of two basic modes: the forced variations which are the response of the climate system to changes in the external forcing  $F(x, t)$  (mostly called climate change), and the free variations owing to internal instabilities and feedbacks leading to non-linear interactions among the various components of the climate system [68] (mostly called climate variability). The external causes  $F(x, t)$ , operate mostly by causing variations in the amount of solar radiation received by or absorbed by the Earth, and comprise variations in both astronomical (e. g. orbital parameters) and terrestrial forcings (e. g. atmospheric composition, aerosol loading). For example, the diurnal and seasonal variations in climate are related to external astronomical forcings operating via solar radiation, while ice ages are related to changes in Earth orbital parameters. Volcanic eruptions are one example of a terrestrial forcing which may introduce abrupt modifications over a period of 2 or 3 years. The more rapid the forcing, the more likely it is that it will cause an abrupt change. The resulting evolution may be written as

$$\frac{d}{dt}x(t) = f(x) + g(x)\xi + F(x, t). \quad (12)$$

$F(x, t)$  is independent of  $x$  if the forcing does not depend on climate (external forcing).

The internal free variations within the climate system are associated with both positive and negative feedback interactions between the atmosphere, oceans, cryosphere and biosphere. These feedbacks lead to instabilities or oscillations of the system on all time scales, and can either operate independently or reinforce external forcings. Investigations of the properties of systems which are far from

equilibrium show that they have a number of unusual properties. In particular, as the distance from equilibrium increases, they can develop complex oscillations with both chaotic and periodic characteristics. They also may show bifurcation points where the system may switch between various regimes. Under non-equilibrium conditions, local events have repercussions throughout the whole system. These long-range correlations are at first small, but increase with distance from equilibrium, and may become essential at bifurcation points.

When applying (12), different concepts of climate change are in the literature. Quite often, the dynamics is governed by the following stochastic differential equation

$$\frac{d}{dt}x(t) = -\frac{d}{dx}U(x) + \sigma\xi + F(t) \quad (13)$$

with potential

$$U(x) = a_4x^4 + a_3x^3 + a_2x^2 + a_1x. \quad (14)$$

If the potential is quadratic and  $F(t) = 0$ , the Orstein-Uhlenbeck process is retained. In contrast, a bistable non-linear system with two minima in  $U(x)$  has been assumed in which shifts between the two distinctly different states are triggered randomly by stochastic forcing, e. g. [7]. In such a system, climate variability and change in the potential can interact due to stochastic resonance [1,7]. Stochastic resonance occurs when the signal-to-noise ratio of a non-linear device is maximized for a moderate value of noise intensity  $\sigma$ . It often occurs in bistable and excitable systems with sub-threshold inputs. For lower noise intensities, the signal does not cause the device to cross threshold, so little signal is passed through it. For large noise intensities, the output is dominated by the noise, also leading to a low signal-to-noise ratio. For moderate intensities, the noise allows the signal to reach threshold, but the noise intensity is not so large as to swamp it.

Strictly speaking, stochastic resonance occurs in bistable systems, when a small periodic force  $F(t)$  (which is external) is applied together with a large wide-band stochastic force  $\sigma\xi$  (which is internal). The system response is driven by the combination of the two forces that compete/cooperate to make the system switch between the two stable states. The degree of order is related to the amount of periodic function that it shows in the system response. When the periodic force is chosen small enough in order to not make the system response switch, the presence of a non-negligible noise is required for it to happen. When the noise is small very few switches occur, mainly at random with no significant periodicity in the system response. When the noise is very strong a large number of

switches occur for each period of the periodic force and the system response does not show remarkable periodicity. Quite surprisingly, between these two conditions, there exists an optimal value of the noise that cooperatively concurs with the periodic forcing in order to make almost exactly one switch per period (a maximum in the signal-to-noise ratio).

Furthermore, non-linear oscillators have been proposed where the timing of the deterministic external forcing is crucial for generating oscillations [51,77,78]. Some aspects of non-equilibrium systems can be found in the climatic system. On the climatological scale, it exhibits abrupt jumps in the long-term rate of temperature change, which are often associated with changes in circulation patterns.

### Frequency Domain

In the frequency domain, there are different ways to describe abrupt climate change. A stationary process exhibits an autocovariance function of the form

$$\text{Cov}(\tau) = \langle (x(t + \tau) - \langle x \rangle)(x(t) - \langle x \rangle) \rangle \quad (15)$$

where  $\langle \rangle$  denotes the expectation or the statistical mean. Normalized to the variance (i. e. the autocovariance function at  $\tau = 0$ ) one gets the autocorrelation function  $C(\tau)$ :

$$C(\tau) = \text{Cov}(\tau) / \text{Cov}(0). \quad (16)$$

Many stochastic processes in nature exhibit short-range correlations, which decay exponentially:

$$C(\tau) \sim \exp(-\tau/\tau_0), \quad \text{for } \tau \rightarrow \infty. \quad (17)$$

These processes exhibit a typical time scale  $\tau_0$ .

As the frequency domain counterpart of the autocovariance function of a stationary process, one can define the spectrum as

$$S(\omega) = \widehat{\text{Cov}(\tau)}, \quad (18)$$

where the hat denotes the Fourier transformation. However, geophysical processes are furthermore often non-stationary. In this regard, the optimal method is continuous wavelet analysis as it intrinsically adjusts the time resolution to the analyzed scale, e. g. [16,59].

**Wavelet Spectra** A major question concerns the significance testing of wavelet spectra. Torrence and Compo [86] formulated pointwise significance tests against reasonable background spectra. However, Maraun and Kurths [58] pointed out a serious deficiency of pointwise significance

testing: Given a realization of white noise, large patches of spurious significance are detected, making it – without further insight – impossible to judge which features of an estimated wavelet spectrum differ from background noise and which are just artifacts of multiple testing. Under these conditions, a reliable corroboration of a given hypothesis is impossible. This demonstrates the necessity to study the significance testing of continuous wavelet spectra in terms of sensitivity and specificity. Given the set of all patches with pointwise significant values, areawise significant patches are defined as the subset of additionally areawise significant wavelet spectral coefficients given as the union of all critical areas that completely lie inside the patches of pointwise significant values. Whereas the specificity of the areawise test appears to be – almost independently of the signal-to-noise ratio – close to one, that of the pointwise test decreases for high background noise, as more and more spurious patches appear [58].

**Eigenvalues and Pseudospectrum** Another spectral method characterizing the abruptness of climate change is related to the resonance of the linear system (1). As we will see later in the context of atmosphere and ocean instabilities, an eigenvalue analysis is inappropriate in describing the dynamics of the system (12). Inspection of many geophysical systems shows that most of the systems fail the normality condition

$$A A^\dagger = A^\dagger A, \quad (19)$$

where  $\dagger$  denotes the adjoint-complex operator. If a matrix is far from normal, its eigenvalues (the spectrum) have little to do with its temporal evolution [71,87]. More about the dynamics can be learned by examining the pseudospectrum of  $A$  in the complex plane. The  $\epsilon$ -pseudospectrum of operator  $A$  is defined by two equivalent formulations:

$$\begin{aligned} \Lambda_\epsilon(A) &= \{z \in \mathbb{C} : \|(zI - A)^{-1}\| \geq \epsilon^{-1}\} \\ &= \{z \in \mathbb{C} : [\text{smallest singular value of } (zI - A)] \leq \epsilon\}. \end{aligned} \quad (20)$$

This set of values  $z$  in the complex plane are defined by contour lines of the resolvent  $(zI - A)^{-1}$ . The resolvent determines the system's response to a forcing as supplied by external forcing  $F(x, t)$ , stochastic forcing  $g(x)\xi$ , or initial/boundary conditions. The pseudospectrum reflects the robustness of the spectrum and provides information about instability and resonance. One theorem is derived from Laplace transformation stating that transient growth is related to how far the  $\epsilon$ -pseudospectrum extends into

the right half plane:

$$\|\exp(A t)\| \geq \frac{1}{\epsilon} \sup_{z \in \Lambda_\epsilon(A)} \operatorname{Real}(z). \quad (21)$$

In terms of climate theory, the pseudospectrum indicates resonant amplification. Maximal amplification is at the poles of  $(zI - A)^{-1}$ , characterized by the eigenfrequencies. In a system satisfying (19), the system's response is characterized solely by the proximity to the eigenfrequencies. In the non-normal case, the pseudospectrum shows large resonant amplification for frequencies which are not eigenfrequencies. This transient growth mechanism is important for both initial value and forced problems.

## Earth System Modeling and Analysis

### Hierarchy of Models

Modeling is necessary to produce a useful understanding of abrupt climate processes. Model analyses help to focus research on possible causes of abrupt climate change, such as human activities; on key areas where climatic thresholds might be crossed; and on fundamental uncertainties in climate-system dynamics. Improved understanding of abrupt climatic changes that occurred in the past and that are possible in the future can be gained through climate models. A comprehensive modeling strategy designed to address abrupt climate change includes vigorous use of a hierarchy of models, from theory and conceptual models through models of intermediate complexity, to high-resolution models of components of the climate system, to fully coupled earth-system models. The simpler models are well-suited for use in developing new hypotheses for abrupt climate change. Model-data comparisons are needed to assess the quality of model predictions. It is important to note that the multiple long integrations of enhanced, fully coupled Earth system models required for this research are not possible with the computer resources available today, and thus, these resources are currently being enhanced.

### Feedback

One particularly convincing example showing that the feedbacks in the climate system are important is the drying of the Sahara about 5000 years before present which is triggered by variations in the Earth's orbit around the sun. Numerous modeling studies, e. g. [31], suggest that the abruptness of the onset and termination of the early to mid-Holocene humid period across much of Africa north of the equator depends on the presence of non-linear feedbacks associated with both ocean circulation and changes

in surface hydrology and vegetation, e. g. [18]. Without including these feedbacks alongside gradual insolation forcing, it is impossible for existing models to come even close to simulating the rapidity or the magnitude of climatic change associated with the extension of wetlands and plant cover in the Sahara/Sahel region prior to the onset of desiccation around 5000 years before present.

## Climate Archives and Modeling

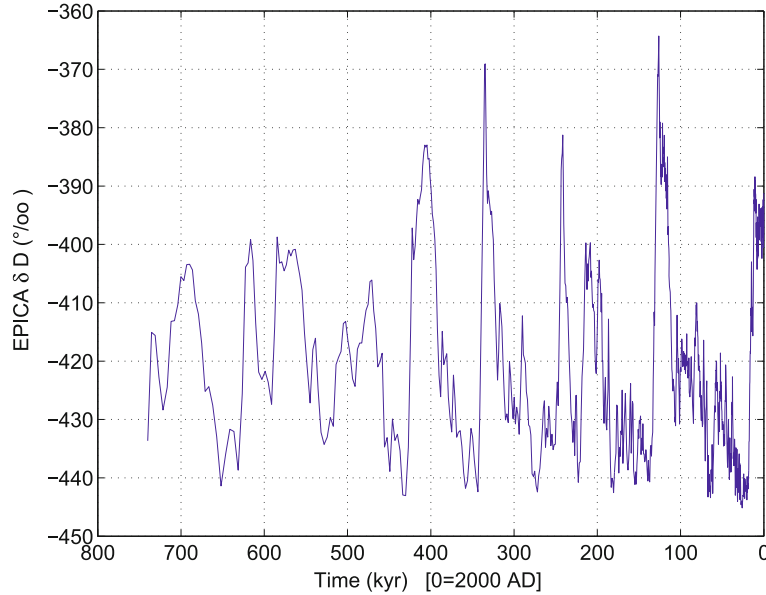
Systematic measurements of climate using modern instruments have produced records covering the last 150 years. In order to reconstruct past variations in the climate system further back in time, scientists use natural archives of climatic and environmental changes, such as ice cores, tree rings, ocean and lake sediments, corals, and historical evidence. Scientists call these records proxies because, although they are not usually direct measures of temperature or other climatic variables, they are affected by temperature, and using modern calibrations, the changes in the proxy preserved in the fossil record can be interpreted in terms of past climate.

Ice core data, coral data, ring width of a tree, or information from marine sediments are examples of a proxy for temperature, or in some cases rainfall, because the thickness of the ring can be statistically related to temperature and/or rainfall in the past. The most valuable proxies are those that can be scaled to climate variables, and those where the uncertainty in the proxy can be measured. Proxies that cannot be quantified in terms of climate or environment are less useful in studying abrupt climate change because the magnitude of change cannot be determined. Quite often, the interpretation of proxy data is already a model of climate change since it involves constraints (dating, representativeness etc.). Uncertainties in the proxies, and uncertainties in the dating, are the main reasons that abrupt climate change is one of the more difficult topics in the field of paleoclimatology.

## Example: Glacial-Interglacial Transitions

### Astronomical Theory of Ice Ages

Over the past half million years, marine, polar ice core and terrestrial records all highlight the sudden and dramatic nature of glacial terminations, the shifts in global climate that occurred as the world passed from dominantly glacial to interglacial conditions, e. g. [23,69]. These climate transitions, although probably of relatively minor relevance to the prediction of potential future rapid climate change, do provide the most compelling evidence available in the historical record for the role of greenhouse gas, oceanic



Abrupt Climate Change Modeling, Figure 1

Oxygen isotope record from a southern hemisphere ice core [23] showing the glacial-interglacial changes. Note the asymmetry: the state is longer in the cold (glacials) phases than in the warm phases (interglacials)

and biospheric feedbacks as non-linear amplifiers in the climate system. It is such evidence of the dramatic effect of non-linear feedbacks that shows relatively minor changes in climatic forcing may lead to abrupt climate response.

A salient feature of glacial-interglacial climate change is furthermore its asymmetry (Fig. 1). Warmings are rapid, usually followed by slower descent into colder climate. Given the symmetry of orbital forcings  $F(t)$ , the cause of rapid warming at glacial “terminations” must lie in a climate feedback [37,65]. Clearly, the asymmetric feedback is due to the albedo (reflectance) of ice and snow changing from high values under glacial climates to low values under warm climates. The albedo feedback helps explain the rapidity of deglaciations and their beginnings in spring and summer. Increased absorption of sunlight caused by lower albedo provides the energy for rapid ice melt. The build-up of snow and ice takes much longer than melting.

Many simplified climate models consist of only a few coupled ordinary differential equations controlled by carefully selected parameters. It is generally acknowledged that the “best” models will be those that contain a minimum of adjustable parameters [77] and are robust with respect to changes in those parameters. Rial [72] formulated a logistic-delayed and energy balance model to understand the saw-tooth shape in the paleoclimate record: A fast warm-

ing-slow cooling is described by

$$\frac{d}{dt}x(t) = R \left( 1 - \frac{x(t-\tau)}{K(t)} \right) x(t-\tau) \quad (22)$$

$$C \frac{d}{dt}T(t) = Q(1 - \alpha(x)) - (A + BT) \quad (23)$$

with  $x(t)$  for the normalized ice extent,  $\tau$  time delay,  $K(t) = 1 + e(t)T(t)$  carrying capacity,  $1/R$  response time of the ice sheet,  $T(t)$  global mean temperature,  $\alpha(x)$  planetary albedo, external parameter  $e(t)$ , and  $R\tau$  bifurcation parameter.  $A, B, C, Q$  are constants for the energy balance of the climate. The equation is calibrated so that for  $x(t) = 1$  the albedo  $\alpha(x) = 0.3$  and  $T(t) = 15^\circ\text{C}$ . With (23), saw-toothed waveforms and frequency modulation can be understood [72]. The delayed equation yields damped oscillations of  $x(t)$  about the carrying capacity for small  $\tau$ . If  $\tau$  becomes long compared to the natural response time of the system, the oscillations will become strong, and will grow in amplitude, period and duration. As in the logistic equation for growth, here the product  $R\tau$  is a bifurcation parameter, which when crossing the threshold value  $\pi/2$  makes the solutions undergo a Hopf bifurcation and settle to a stable limit cycle with fundamental period  $\sim 4\tau$  [73].

The astronomical theory of ice ages – also called Milankovitch theory [62] – gained the status of a paradigm for explaining the multi-millennial variability. A key



element of this theory is that summer insolation at high latitudes of the northern hemisphere determines glacial-interglacial transitions connected with the waxing and waning of large continental ice sheets, e.g. [33,37], the dominant signal in the climate record for the last million years. Climate conditions of glacials and interglacials are very different. During the Last Glacial Maximum, about 20,000 years before present, surface temperature in the north Atlantic realm was 10–20°C lower than today [13]. A recent study of Huybers and Wunsch [36] has shown that the most simple system for the phase of ice volume  $x(t)$  is given by

$$x(t+1) = x(t) + \sigma \xi \quad (24)$$

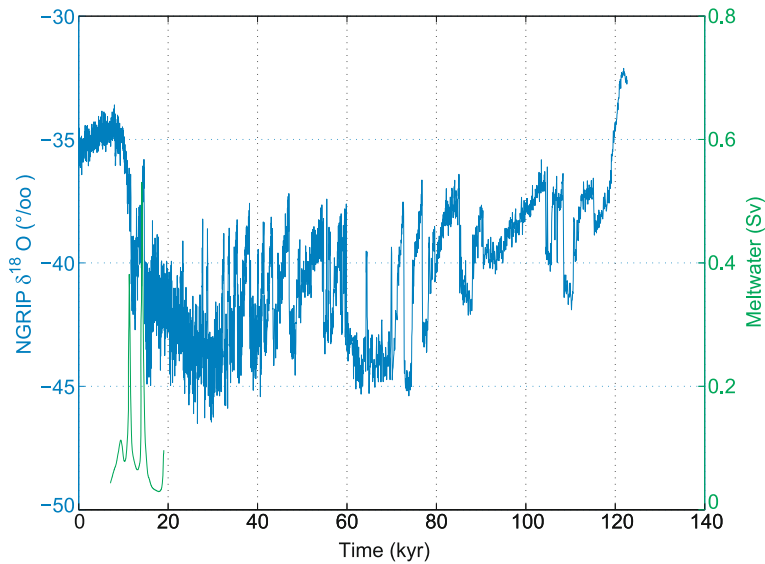
with  $\xi$  a Gaussian white noise process, but with mean  $\mu = 1$ , and  $\sigma = 2$ .  $\xi$  represents the unpredictable background weather and climate variability spanning all time scales out to the glacial/interglacial. This highly simplified model posits 1-ky steps in ice volume  $x(t)$ . The non-zero mean biases the Earth toward glaciation. Once  $x(t)$  reaches a threshold, a termination is triggered, and ice-volume is linearly reset to zero over 10 ky. The following threshold condition for a termination makes it more likely for a termination of ice volume to occur when obliquity  $\Theta(t)$  is large:

$$x(t) \geq T_0 - a\Theta(t). \quad (25)$$

$\Theta(t)$  has a frequency of about 41 ky, and is furthermore normalized to zero mean with unit variance. The other parameters are: amplitude  $a = 15$ ,  $T_0 = 105$ . Furthermore, the initial ice volume at 700 ky before present is set to  $x(t = -700) = 30$ . Equation (24) resembles an order-one autoregressive process, similar to (2), plus the threshold condition (25). Models like (24), (25) are not theories of climate change, but rather attempts at efficient kinematic descriptions of the data, and different mechanisms can be consistent with the limited observational records. In the next section, the process of deglaciation is modeled in a three-dimensional model including the spatial dimension.

### Deglaciation

The question is what causes the abrupt warming at the onset of the Boelling as seen in the Greenland ice cores (Fig. 2). There is a clear antiphasing seen in the deglaciation interval between 20 and 10 ky ago: During the first half of this period, Antarctica steadily warmed, but little change occurred in Greenland. Then, at the time when Greenland's climate underwent an abrupt warming, the warming in Antarctica stopped. Knorr and Lohmann [42], also summarizing numerous modeling studies for deglaciation, describe how global warming (which may be induced by greenhouse gases and feedbacks) can induce a rapid intensification of the ocean cir-



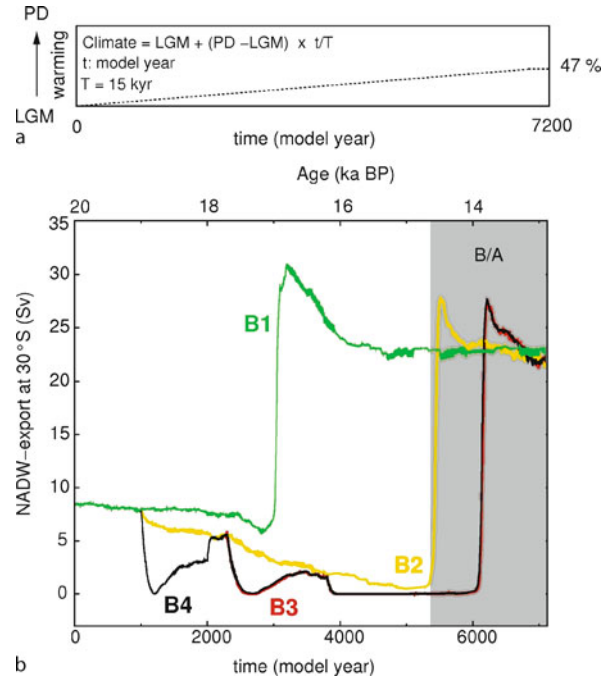
Abrupt Climate Change Modeling, Figure 2

Oxygen isotope record from a Greenland ice core record [64] using an updated time scale for this record [23]. Green: Sea-level derived rate of deglacial meltwater discharge [24] which is strong after deglacial warming

ulation (Fig. 3). During the Boelling/Alleroed, a sudden increase of the northward heat transport draws more heat from the south, and leads to a strong warming in the north. This “heat piracy” from the South Atlantic has been formulated by Crowley [15]. A logical consequence of this heat piracy is the Antarctic Cold Reversal (ACR) during the Northern Hemisphere warm Boelling/Alleroed. This particular example shows that an abrupt climate change of the ocean circulation (with large climate impacts in the North Atlantic) is related to a smooth global warming. To understand the dynamical behavior of the system, the concept of hysteresis is applied, using the global warming after the last ice ages as the control parameter [42]. The system exhibits multiple steady states (Fig. 4): a weak glacial ocean circulation and a stronger circulation (which is comparable in strength to the modern mode of operation). Deglacial warming induces a transition from a weak glacial THC state to a stronger THC state, characterizing the abrupt warming during the deglaciation.

### Millennial Climate Variability

Within glacial periods, and especially well documented during the last one, spanning from around 110 to 11.6ky ago, there are dramatic climate oscillations, including high-latitude temperature changes approaching the same magnitude as the glacial cycle itself, recorded in archives from the polar ice caps, high to middle latitude marine sediments, lake sediments and continental loess sections. These oscillations are usually referred to as the Dansgaard–Oeschger Cycle and occur mostly on 1 to 2 ky time scales, e. g. [6], although regional records of these transitions can show much more rapid change. The termination of the Younger Dryas cold event, for example, is manifested in ice core records from central Greenland as a near doubling of snow accumulation rate and a temperature shift of around 10°C occurring within a decade with world-wide teleconnections. One hypothesis for explaining these climatic transitions is that the ocean thermohaline circulation flips between different modes, with warm intervals reflecting periods of strong deep water formation in the northern North Atlantic and vice versa [29]. As an alternative approach, one can estimate the underlying dynamics (13), (14) directly from data [43]. The method is based on the unscented Kalman filter, a non-linear extension of the conventional Kalman filter. This technique allows one to consistently estimate parameters in deterministic and stochastic non-linear models. The optimization yields for the coefficients  $a_4 = 0.13 \pm 0.01$ ,  $a_3 = -0.27 \pm 0.02$ ,  $a_2 = -0.36 \pm 0.08$ , and  $a_1 = 1.09 \pm 0.23$ . The dynamical noise level of the system  $\sigma$  is estimated to



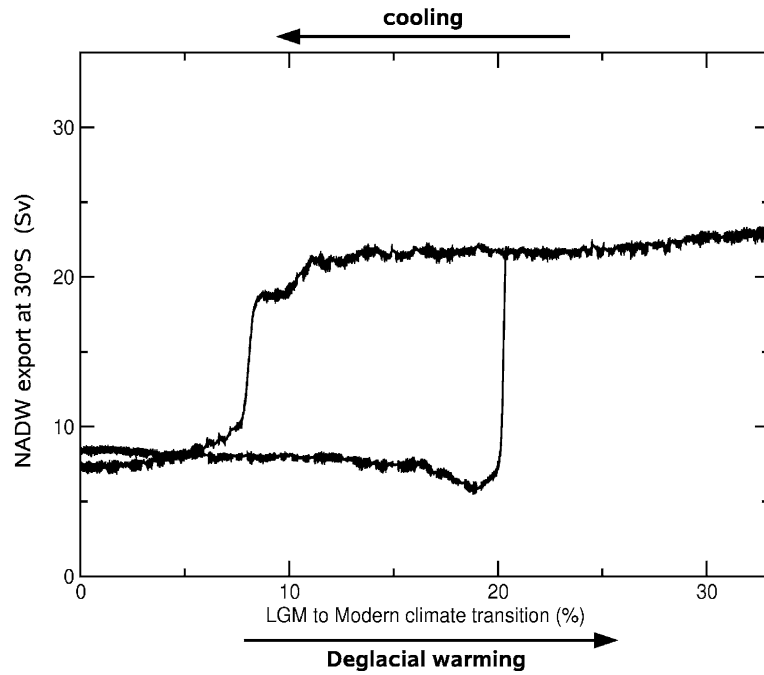
Abrupt Climate Change Modeling, Figure 3

Forcing and model response of the ocean overturning rate. **a** The background climate conditions are linearly interpolated between glacial (LGM), and modern (PD), conditions. Gradual warming is stopped after 7000 model years, which is related to  $\sim 47\%$  of the total warming. **b** Circulation strength (export at 30°S) versus time. The *green curve B1* represents the experiment without any deglacial freshwater release to the North Atlantic. Experiments *B2* (yellow curve), *B3* (red curve), and *B4* (black curve), exhibit different successions of deglacial meltwater pulse scenarios to the North Atlantic [42]

be 2.4. The potential is highly asymmetric and degenerate (that is, close to a bifurcation): there is one stable cold stadial state and one indifferently stable warm interstadial state (Fig. 5). This seems to be related to the fact that the warm intervals are relatively short-lasting.

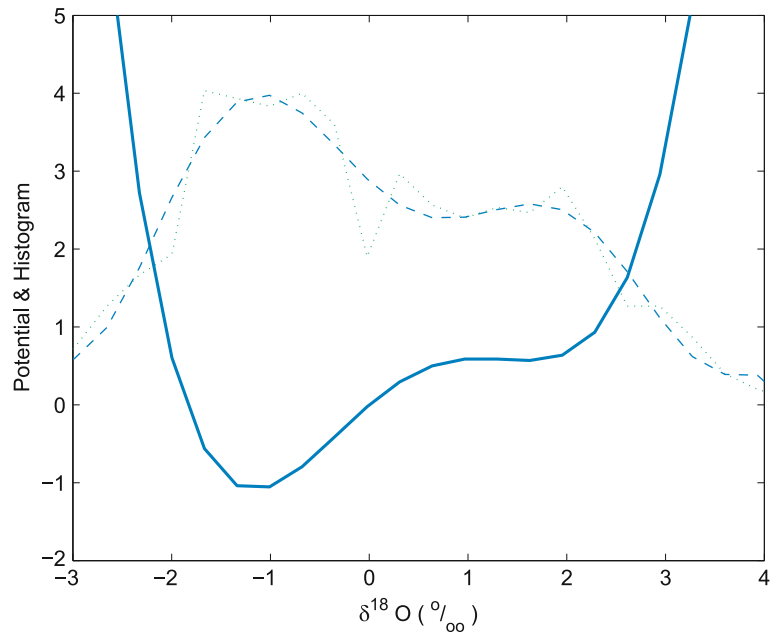
Coming back to the ice cores and a potential linkage of the hemispheres, Stocker and Johnson [81] proposed a conceptual model linking the isotopic records from Antarctica and Greenland. The basis is an energy balance with temperatures in the North and South Atlantic Ocean, as well as a “southern heat reservoir”. It is assumed that the change in heat storage of a “southern heat reservoir”  $T_S$  is given by the temperature difference between the reservoir  $T_S$  and the Southern Ocean temperature  $T$ , with a characteristic time scale  $\tau$ :

$$\frac{d}{dt} T_S(t) = \frac{1}{\tau} [T - T_S] . \quad (26)$$



Abrupt Climate Change Modeling, Figure 4

Hysteresis loop of the ocean overturning strength (*black curve*) with respect to slowly varying climate background conditions. The transition values are given in % of a full glacial-interglacial transition [42]



Abrupt Climate Change Modeling, Figure 5

Potential derived from the data (*solid*) together with probability densities of the model (*dashed*) and the data (*dotted*)

$T_N$  denotes the time-dependent temperature anomaly of the North Atlantic. The Southern Ocean temperature  $T$  is assumed to be  $-T_N$  according to the bipolar seesaw (North Atlantic cold  $\leftrightarrow$  South Atlantic warm). Using Laplace transform, one can solve for  $T_S$

$$T_S = -\frac{1}{\tau} \int_0^t T_N(t-t') \exp(-t'/\tau) dt' + T_S(0) \exp(-t/\tau). \quad (27)$$

The reservoir temperature is therefore a convolution of the northern temperature using the time scale  $\tau$  ranging from 100 to 4000 years. Equation (27) demonstrates that  $T_S$  and  $T_N$  will have entirely different time characteristics. Abrupt changes in the north appear damped and integrated in time in the southern reservoir. A sudden reduction in the thermohaline circulation causes a cooling in the North Atlantic and a warming in the South, a situation similar to the Younger Dryas period [80], see also Fig. 2.

## Example: Cenozoic Climate Cooling

### Antarctic Glaciation

During the Cenozoic (65 million years ago (Ma) to present), there was the widespread glaciation of the Antarctic continent at about 34 Ma, e.g. [93]. Antarctic glaciation is the first part of a climate change from relatively warm and certainly ice-free conditions to massive ice sheets in both, the southern and northern hemispheres [44]. Opening of circum-Antarctic seaways is one of the factors that have been ascribed as a cause for Antarctic climate change so far [40,93]. Besides gateway openings, the atmospheric carbon dioxide concentration is another important factor affecting the evolution of the Cenozoic climate [17,93]. As a third component in the long-term evolution of Antarctic glaciation, land topography is able to insert certain thresholds for abrupt ice sheet build-up. Whereas tectonics, land topography, and long-term Cenozoic  $\text{CO}_2$ -decrease act as preconditioning for Antarctic land ice formation, the cyclicities of the Earth's orbital configuration are superimposed on shorter time scales and may have served as the ultimate trigger and pacemaker for ice-sheet growth at the Eocene-Oligocene boundary around 34 Ma [14].

DeConto and Pollard [17] varied Southern Ocean heat transport to mimic gateway opening instead of an explicit simulation of ocean dynamics. They found a predominating role of  $\text{pCO}_2$  in the onset of glaciation instead of a dominating tectonic role for "thermal isolation".

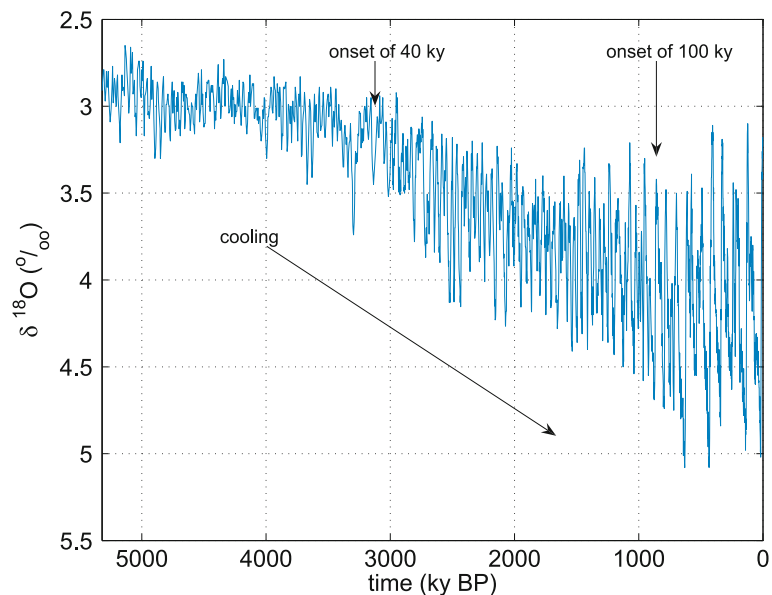
### Mid-Pleistocene Revolution

Glaciation in the Northern Hemisphere lagged behind, with the earliest recorded glaciation anywhere in the Northern Hemisphere occurring between 10 and 6 Ma and continuing through to the major increases in global ice volume around 2–3 Ma [60]. A recent compilation of 57 globally distributed records [45] is shown in Fig. 6. Let us focus now on the mid-Pleistocene transition or revolution (MPR), describing the transition from 41 ky to 100 ky glacial-interglacial cycles.

Milankovitch [62] initially suggested that the critical factor was total summer insolation at about  $65^\circ\text{N}$ , because for an ice sheet to grow some additional ice must survive each successive summer. In contrast, the Southern Hemisphere is limited in its response because the expansion of ice sheets is curtailed by the Southern Ocean around Antarctica. The conventional view of glaciation is thus that low summer insolation in the temperate North Hemisphere allows ice to survive summer and thus start to build up on the northern continents. If so, how then do we account for the MPR? Despite the pronounced change in Earth system response evidenced in paleoclimatic records, the frequency and amplitude characteristics of the orbital parameters which force long-term global climate change, e.g., eccentricity ( $\sim 100$  ky), obliquity ( $\sim 41$  ky) and precession ( $\sim 21$  and  $\sim 19$  ky), do not vary during the MPR [8]. This suggests that the cause of change in response at the MPR is internal rather than external to the global climate system.

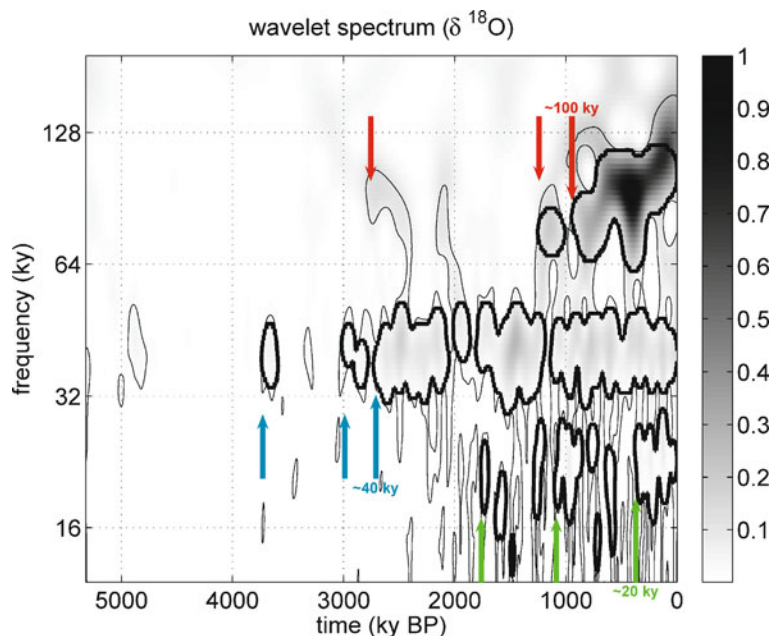
The result of a wavelet spectral analysis (Fig. 7) suggests several abrupt climate changes in the frequency domain (shown as schematic arrows in the figure). These abrupt climate shifts represent major reorganizations in the climate system. Some of them are possibly linked to the development of Northern Hemisphere ice volume. The MPR marked a prolongation to and intensification of the  $\sim 100$  ky glacial-interglacial climate. Not only does the periodicity of glacial-interglacial cycles increase going through the MPR, but there is also an increase in the amplitude of global ice volume variations.

It is likely that the MPR is a transition to a more intense and prolonged glacial state, and associated subsequent rapid deglaciation becomes possible. The first occurrence of continental-scale ice sheets, especially on Greenland, is recorded as ice-rafted detritus released from drifting icebergs into sediments of the mid- and high-latitude ocean. After a transient precursor event at 3.2 Ma, signals of large-scale glaciations suddenly started in the subtropical North Atlantic in two steps, at 2.9 and 2.7 Ma, e.g. [5].



Abrupt Climate Change Modeling, Figure 6

A compilation of 57 globally distributed records by Lisiecki and Raymo [45]: The  $\delta^{18}\text{O}$  record reflects mainly the climate variables temperature and ice volume



Abrupt Climate Change Modeling, Figure 7

Lisiecki and Raymo [45]: The corresponding wavelet sample spectrum calculated using Morlet wavelet with  $\omega_0 = 6$ . *Thin and thick lines* surround pointwise and areawise significant patches, respectively

The ice volume increase may in part be attributed to the prolonging of glacial periods and thus of ice accumulation. The amplitude of ice volume variation is also accentuated by the extreme warmth of many interglacial pe-

riods. Thus, a colder climate with larger ice sheets should have the possibility of a greater sudden warming [45]. The MPR therefore marks a dramatic sharpening of the contrast between warm and cold periods. Note however, that

the amount of energy at the 40 ka period is hardly changed in the time after 1 Ma, and notably, one sees the addition of energy at longer periods, without any significant reduction in obliquity-band energy. After about 1 Ma, large glacial-interglacial changes begin to occur on an approximately 100 ka time scale (but not periodically) superimposed upon the variability which continues largely unchanged [91]. Why did 100 ka glacial-interglacials also become possible in addition to the ice volume variability? Lowering of global CO<sub>2</sub> below some critical threshold, or changes in continental configuration, or atmospheric circulation patterns, or all together, are among the conceivable possibilities, e. g. [70].

### Examples: Transient Growth

The former examples show the power of the combination of models, data analysis, and interpretation for abrupt climate change. In the next two examples, it is shown how important the transient growth mechanism is for abrupt climate change.

### Conceptual Model of the Ocean Circulation

In this section, a category of the non-linear models following the simple thermohaline model of Stommel [82] is analyzed. The common assumption of these box models is that the oceanic overturning rate  $\Phi$  can be expressed by the meridional density difference:

$$\Phi = -c(\alpha\Delta T - \beta\Delta S), \quad (28)$$

where  $\alpha$  and  $\beta$  are the thermal and haline expansion coefficients,  $c$  is a tunable parameter, and  $\Delta$  denotes the meridional difference operator applied to the variables temperature  $T$  and salinity  $S$ , respectively. Stommel [82] considered a two-box ocean model where the boxes are connected by an overflow at the top and a capillary tube at the bottom, such that the capillary flow is directed from the high-density vessel to the low-density vessel following (28).

The equations for temperature  $T$  and salinity  $S$  are the heat and salt budgets using an upstream scheme for the advective transport and fluxes with the atmosphere:

$$\frac{d}{dt}T = -\frac{\Phi}{V}\Delta T - \frac{F_{\text{oa}}}{\rho_0 c_p h} \quad (29)$$

$$\frac{d}{dt}S = -\frac{\Phi}{V}\Delta S - \frac{S_0}{h}(P - E), \quad (30)$$

where  $V$  is the volume of the box with depth  $h$ , and  $(P - E)$  denotes the freshwater flux (precipitation minus evaporation plus runoff).  $F_{\text{oa}}$  is the heat flux at the ocean-atmo-

sphere interface,  $S_0$  is a reference salinity, and  $\rho_0 c_p$  denotes the heat capacity of the ocean.

Denoting furthermore  $x \in \mathbb{R}^2$  for the anomalies of  $(\Delta T, \Delta S)$ , Lohmann and Schneider [48] have shown the evolution equation is of the following structure:

$$\frac{d}{dt}x = Ax + \langle b|x \rangle x. \quad (31)$$

The brackets  $\langle | \rangle$  denote the Euclidean scalar product. This evolution Equation (31) can be transferred to a

$$x(t) = \frac{1}{\gamma(t)} \exp(At)x_0, \quad (32)$$

with a scaling function  $\gamma(t, x_0)$ . The models of Stommel [82], and many others are of this type, and their dynamics are therefore exactly known.

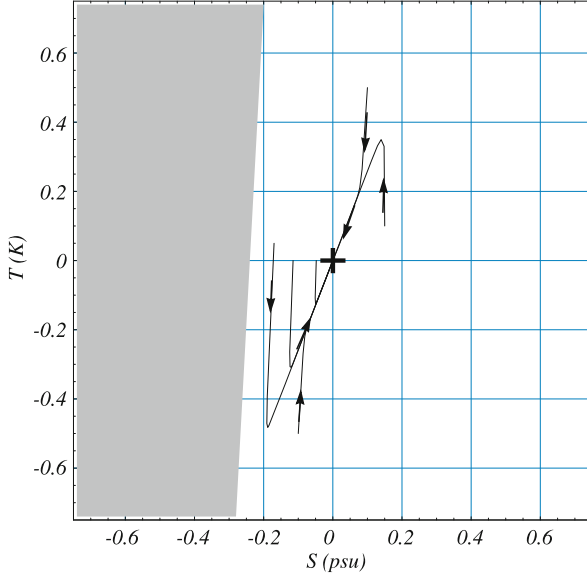
It is worth knowing that (29), (30) is equivalent to the multi-dimensional Malthus–Verhulst model (also known as the logistic equation), which was originally proposed to describe the evolution of a biological population. Let  $x$  denote the number (or density) of individuals of a certain population. This number will change due to growth, death, and competition. In the simplest version, birth and death rates are assumed proportional to  $n$ , but accounting for limited resources and competition it is modified by  $(1 - x)$ :

$$\frac{d}{dt}x(t) = a(1 - x)x. \quad (33)$$

In climate, the logistic equation is important for Lorenz's [52] error growth model: where  $x(t)$  is the algebraic forecast error at time  $t$  and  $a$  is the linear growth rate.

It is useful to analyze the dynamics in the phase space spanned by temperature and salinity anomalies and investigate the model sensitivity under anomalous high latitude forcing, induced as an initial perturbation. The lines in Fig. 8 are phase space trajectories after perturbations of different magnitude have been injected into the North Atlantic. We notice that for most trajectories, the distances from zero  $(0, 0)$  increase temporarily, where the maximal distance from zero is after a decade. After about 10 years the trajectories in Fig. 8 point into a "mixed temperature/salinity direction", denoted further as  $e_1$ .

Figure 8 implies that the adjustment of the THC involves two phases: A fast thermal response and a slower response on the  $e_1$ -direction. The vector  $e_1$  is identical with the most unstable mode in the system. Because the scaling function  $\gamma(t)$  acts upon both temperature and salinity (32), the evolution of the non-linear model can be well characterized by the eigenvectors of the matrix  $A$ , which is discussed in the following.



Abrupt Climate Change Modeling, Figure 8

The basin of attraction (*white area*) and the dynamics in the thermohaline phase space. With initial conditions outside the gray area, the trajectories converge asymptotically to the origin corresponding to the thermally driven solution of the THC. Due to the fast thermal response during the first decade of relaxation, the distance of the trajectories from zero can increase temporarily

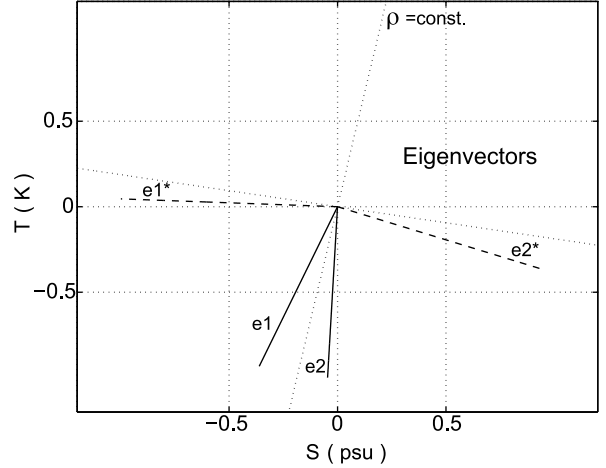
In our system, the operator  $A$  of the box model is found to be non-normal, and the eigenvectors are not orthogonal. One eigenvalue ( $e_2$ ) is closely related to temperature anomalies, whereas the other ( $e_1$ ) is a “mixed temperature/salinity eigenvector” (Fig. 9). The eigenvectors of the adjoint matrix  $A^\dagger$  are denoted by  $e_1^*$  and  $e_2^*$ , respectively. For the non-normal matrix  $A$  and  $A^\dagger$  do not coincide, but fulfill the “biorthogonality condition”:

$$e_1^* \perp e_2 \quad \text{and} \quad e_2^* \perp e_1. \quad (34)$$

Both eigenvalues  $\lambda_{1,2}$  are real and negative. Because of  $\lambda_2 < \lambda_1$ , the first term dominates for long time scales and the second for short time scales. Using the biorthogonality condition, we get furthermore the coefficients

$$c_i = \frac{\langle e_i^* | x_0 \rangle}{\langle e_i^* | e_i \rangle} \quad \text{for } i = 1, 2. \quad (35)$$

A perturbation is called “optimal”, if the initial error vector has minimal projection onto the subspace with fastest decaying perturbations, or equivalently if the coefficient  $c_1$  is maximal. This is according to (35) equivalent to  $x_0$  pointing into the direction of  $e_1^*$ . This unit vector



Abrupt Climate Change Modeling, Figure 9

Eigenvectors  $e_1, e_2$ , and adjoint eigenvectors  $e_1^*, e_2^*$  of the tangent linear operator  $A^\dagger$ . The *dotted lines* show the line of constant density and the perpendicular

$e_1^*$  is called the “biorthogonal” [66] to the most unstable eigenvector  $e_1$  which we want to excite. In order to make a geometrical picture for the mathematical considerations, assume that the tail of the vector  $x_0$  is placed on the  $e_1$ -line and its tip on the  $e_2$ -line. This vector is stretched maximally because the tail decays to zero quickly, whereas the tip is hardly unchanged due to the larger eigenvalue  $\lambda_1$ . The most unstable mode  $e_1$  and its biorthogonal  $e_1^*$  differ greatly from each other, and the perturbation that optimally excites the mode bears little resemblance to the mode itself.

It is remarkable that the optimal initial perturbation vector  $e_1^*$  does not coincide with a perturbation in sea surface density at high latitudes, which would reside on the dotted line perpendicular to  $\rho = \text{const}$  in Fig. 9. Even when using a space spanned by  $(\alpha T, \beta S)$  instead of  $(T, S)$ , to take into account the different values for the thermal and haline expansion coefficients, vector  $e_1^*$  is much more dominated by the scaled salinity anomalies than the temperature anomalies of the high latitudinal box.

Numerical simulations by Manabe and Stouffer [57] showed, for the North Atlantic, that between two and four times the preindustrial  $\text{CO}_2$  concentration, a threshold value is passed and the thermohaline circulation ceases completely. One other example of early Holocene rapid climate change is the “8200-yr BP” cooling event recorded in the North Atlantic region possibly induced by freshwater. One possible explanation for this dramatic regional cooling is a shutdown in the formation of deep water in the northern North Atlantic due to freshwater input caused by

catastrophic drainage of Laurentide lakes [4,46]. The theoretic considerations and these numerical experiments suggest that formation of deep water in the North Atlantic is highly sensitive to the freshwater forcing.

### Resonance in an Atmospheric Circulation Model

An atmospheric general circulation model PUMA [26] is applied to the problem. The model is based on the multi-level spectral model described by Hoskins and Simmons [35]. For our experiments we chose five vertical levels and a T21 horizontal resolution. PUMA belongs to the class of models of intermediate complexity [12]; it has been used to understand principle feedbacks [56], and dynamics on long time scales [76]. For simplicity, the equations are scaled here such that they are dimensionless. The model is linearized about a zonally symmetric mean state providing for a realistic storm track at mid-latitudes [27]. In a simplified version of the model and calculating the linear model  $A$  with  $n = 214$ , one can derive the pseudospectrum. Figure 10 indicates resonances besides the poles (the eigenvalues) indicated by crosses. The  $\text{Im}(z)$ -axis shows the frequencies, the  $\text{Re}(z)$ -axis the damping/amplification of the modes. Important modes for the climate system are those with  $-0.5 < \text{Im}(z) < 0.5$  representing planetary

Rosby waves. The basic feature is that transient growth of initially small perturbations can occur even if all the eigenmodes decay exponentially. Mathematically, an arbitrary matrix  $A$  can be decomposed as a sum

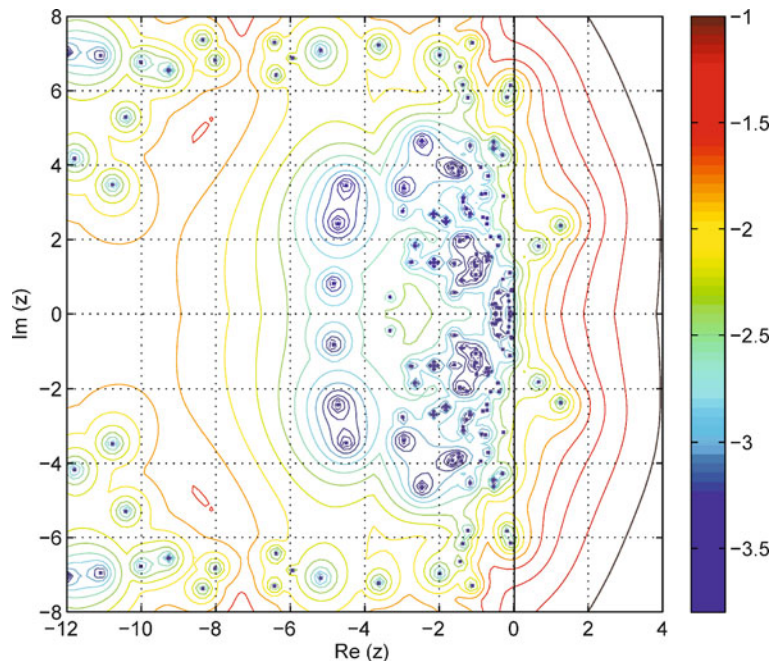
$$A = D + N \quad (36)$$

where  $A$  is diagonalizable, and  $N$  is nilpotent (there exists an integer  $q \in \mathbb{N}$  with  $N^q = 0$ ), and  $D$  commutes with  $N$  (i. e.  $DN = NA$ ). This fact follows from the Jordan–Chevalley decomposition theorem. This means we can compute the exponential of  $(A t)$  by reducing to the cases:

$$\exp(At) = \exp((D + N)t) = \exp(Dt) \exp(Nt) \quad (37)$$

where the exponential of  $Nt$  can be computed directly from the series expansion, as the series terminates after a finite number of terms. Basically, the number  $q \in \mathbb{N}$  is related to the transient growth of the system ( $q = 1$  means no transient growth).

The resonant structures are due to the mode interaction: It is not possible to change one variable without the others, because they are not orthogonal. Interestingly, one can also compute the  $A^\dagger$  model, showing the optimal



Abrupt Climate Change Modeling, Figure 10

Contours of  $\log_{10}(1/\epsilon)$ . The figure displays resonant structures of the linearized atmospheric circulation model. The modes extend to the right half plane and are connected through resonant structures, indicating the transient growth mechanism inherent in atmospheric dynamics



perturbation of a mode  $e_i$  through its biorthogonal vector (35).

The analysis indicates that non-normality of the system is a fundamental feature of the atmospheric dynamics. This has consequences for the error growth dynamics, and instability of the system, e. g. [48,66]. Similar features are obtained in shear flow systems [71,87] and other hydrodynamic applications. This transient growth mechanism is important for both initial value and forced problems of the climate system.

### Future Directions

Until now, details of abrupt climate change are not well known to accurately predict it. With better information, the society could take more confident action to reduce the potential impact of abrupt changes on agriculture, water resources, and the built environment, among other impacts. A better understanding of sea-ice and glacier stability, land-surface processes, and atmospheric and oceanic circulation patterns is needed. Moreover, to effectively use any additional knowledge of these and other physical processes behind abrupt climate change, more sophisticated ways of assessing their interactions must be developed, including:

**Better models.** At present, the models used to assess climate and its impacts cannot simulate the size, speed, and extent of past abrupt changes, let alone predict future abrupt changes. Efforts are needed to improve how the mechanisms driving abrupt climate change are represented in these models and to more rigorously test models against the climate record.

**More theory.** There are concepts to find the underlying dynamical system, to derive a theory from a high-order to low-order description similar to what is done in statistical physics (Mori–Zwanzig approach [63,94], Master equation), or in stochastic differential equations. A systematic reduction of the complex system into fewer degrees of freedom shall bring a deeper level of understanding about the underlying physics. A systematic approach was suggested by Saltzman [77]. Spectral and pseudo-spectral concepts have not been used too much in climate theory. There is a variety of phenomenological stochastic models in which non-linearity and fluctuations coexist, and in which this coexistence leads to interesting phenomena that would not arise without the complex interplay.

**Paleoclimatic data.** More climate information from the distant past would go a long way toward strengthening our understanding of abrupt climate changes and models of past climate. In particular, an enhanced effort is needed to expand the geographic coverage, temporal resolution,

and variety of paleoclimatic data. Although the present climate has no direct analogon to the past [54], the dynamical interpretation of data will improve the understanding of thresholds and non-linearities in the Earth system.

**Appropriate statistical tools.** Because most statistical calculations at present are based on the assumption that climate is not changing but is stationary, they have limited value for non-stationary (changing) climates and for climate-related variables that are often highly skewed by rapid changes over time such as for abrupt-change regimes. Available statistical tools themselves need to be adapted or replaced with new approaches altogether to better reflect the properties of abrupt climate change.

**Synthesis.** Physical, ecological, and human systems are complex, non-linear, dynamic and imperfectly understood. Present climate change is producing conditions outside the range of recent historical experience and observation, and it is unclear how the systems will interact with and react to further climate changes. Hence, it is crucial to be able to better understand and recognize abrupt climate changes quickly. This capability will involve improved monitoring of parameters that describe climatic, ecological, and economic systems. Some of the desired data are not uniquely associated with abrupt climate change and, indeed, have broad applications. Other data take on particular importance because they concern properties or regions implicated in postulated mechanisms of abrupt climate change. Research to increase our understanding of abrupt climate change should be designed specifically within the context of the various mechanisms thought to be involved. Focus is required to provide data for process studies from key regions where triggers of abrupt climate change are likely to occur, and to obtain reliable time series of climate indicators that play crucial roles in the postulated mechanisms. Observations could enable early warning of the onset of abrupt climate change. New observational techniques and data-model comparisons will also be required.

## Bibliography

### Primary Literature

1. Alley RB, Anandakrishnan S, Jung P (2001) Stochastic resonance in the North Atlantic. *Paleoceanogr* 16:190–198
2. Alley RB, Marotzke J, Nordhaus W, Overpeck J, Peteet D, Pielke R Jr, Pierrehumbert R, Rhines P, Stocker T, Talley L, Wallace JM (2002) Abrupt Climate Change: Inevitable Surprises. US National Academy of Sciences, National Research Council Committee on Abrupt Climate Change, National Academy Press, Washington
3. Alverson K, Oldfield F (2000) Abrupt Climate Change. In: Joint Newsletter of the Past Global Changes Project (PAGES) and the

- Climate Variability and Predictability Project (CLIVAR), vol 8, no 1. Bern, pp 7–10
4. Barber DC, Dyke A, Hillaire-Marcel C, Jennings AE, Andrews JT, Kerwin MW, Bilodeau G, McNeely R, Southon J, Morehead MD, Gagnonk JM (1999) Forcing of the cold event of 8,200 years ago by catastrophic drainage of Laurentide lakes. *Nature* 400:344–348
  5. Bartoli G, Sarnthein M, Weinelt M, Erlenkeuser H, Garbe-Schönberg D, Lea DW (2005) Final closure of Panama and the onset of northern hemisphere glaciation. *Earth Planet Sci Lett* 237:33–44
  6. Bender M, Malaize B, Orcharo J, Sowers T, Jouzel J (1999) Mechanisms of Global Climate Change. Clark P et al (eds) *AGU* 112:149–164
  7. Benzi R, Parisi G, Sutera A, Vulpiani A (1982) Stochastic resonance in climatic change. *Tellus* 34:10
  8. Berger A, Loutre MF (1991) Insolation values for the climate of the last 10 million years. *Quat Sci Rev* 10:297–317
  9. Berger WH, Labeyrie LD (1987) Abrupt Climatic Change, Evidence and Implications. NATO ASI Series, Series C, Mathematical and Physical Sciences, vol 216. D Reidel, Dordrecht, pp 425
  10. Broecker WS et al (1985) Does the Ocean-atmosphere System Have More than One Stable Mode of Operation? *Nature* 315:21–26
  11. Bryan F (1986) High Latitude Salinity Effects and Inter-hemispheric Thermohaline Circulations. *Nature* 323:301–304
  12. Claussen M, Mysak LA, Weaver AJ, Crucifix M, Fichefet T, Loutre M-F, Weber SL, Alcamo J, Alexeev VA, Berger A, Calov R, Ganopolski A, Goosse H, Lohmann G, Lunkeit F, Mokhov II, Petoukhov V, Stone P, Wang Z (2002) Earth System Models of Intermediate Complexity: Closing the Gap in the Spectrum of Climate System Models. *Clim Dyn* 18:579–586
  13. CLIMAP project members (1976) The surface of the ice age Earth. *Science* 191:1131–1137
  14. Coxall HK, Wilson PA, Pälike H, Lear CH, Backman J (2005) Rapid stepwise onset of Antarctic glaciation and deeper calcite compensation in the Pacific Ocean. *Nature* 433:53–57. doi:10.1038/nature03135
  15. Crowley TJ (1992) North Atlantic deep water cools the southern hemisphere. *Paleoceanogr* 7:489–497
  16. Daubechies I (1992) Ten Lectures on Wavelets. Society for Industrial and Applied Mathematics (SIAM). CBMS-NSF Regional Conference Series in Applied Mathematics, vol 61, Philadelphia
  17. DeConto RM, Pollard D (2003) Rapid Cenozoic glaciation of Antarctica induced by declining atmospheric CO<sub>2</sub>. *Nature* 421:245–249. doi:10.1038/nature01290
  18. DeMenocal et al (2000) Abrupt onset and termination of the African Humid Period: Rapid climate response to gradual insolation forcing. *Quat Sci Rev* 19:347–361
  19. Diaz HF, Quayle RG (1980) The climate of the United States since 1895: spatial and temporal changes. *Mon Wea Rev* 108:149–226
  20. Dijkstra HA, Te Raa L, Weijer W (2004) A systematic approach to determine thresholds of the ocean's thermohaline circulation. *Tellus* 56A(4):362
  21. Dima M, Lohmann G (2002) Fundamental and derived modes of climate variability. Application to biennial and interannual timescale. *Tellus* 56A:229–249
  22. Dima M, Lohmann G (2007) A hemispheric mechanism for the Atlantic Multidecadal Oscillation. *J Clim* 20:2706–2719
  23. EPICA Community Members (2006) One-to-one coupling of glacial climate variability in Greenland and Antarctica. *Nature* 444:195–198. doi:10.1038/nature05301
  24. Fairbanks RG (1989) A 17,000-year glacio-eustatic sea level record: influence of glacial melting rates on the Younger Dryas event and deep-ocean circulation. *Nature* 342:637–642
  25. Flohn H (1986) Singular events and catastrophes now and in climatic history. *Naturwissenschaften* 73:136–149
  26. Fraedrich K, Kirk E, Lunkeit F (1998) Portable University Model of the Atmosphere. DKRZ Report 16, Hamburg
  27. Frisius T, Lunkeit F, Fraedrich K, James IN (1998) Storm-track organization and variability in a simplified atmospheric global circulation model. *Q J R Meteorol Soc* 124:1019–1043
  28. Fu C, Diaz HF, Dong D, Fletcher JO (1999) Changes in atmospheric circulation over northern hemisphere oceans associated with the rapid warming of the 1920s. *Int J Climatol* 19(6):581–606
  29. Ganopolski A, Rahmstorf S (2001) Rapid changes of glacial climate simulated in a coupled climate model. *Nature* 409:153–158
  30. Ganopolski A, Rahmstorf S (2002) Abrupt glacial climate changes due to stochastic resonance. *Phys Rev Lett* 88(3):038501
  31. Ganopolski A, Kubatzki C, Claussen M, Brovkin V, Petoukhov V (1998) The influence of vegetation-atmosphere-ocean interaction on climate during the mid-Holocene. *Science* 280:1916
  32. Hasselmann K (1976) Stochastic climate models, Part 1, Theory. *Tellus* 28:289–485
  33. Hays JD, Imbrie J, Shackleton NJ (1976) Variations in the Earth's Orbit: Pacemaker of the Ice Ages. *Science* 194:1121–1132
  34. Henderson GM, Slowey NC (2000) Evidence from U-Th dating against Northern Hemisphere forcing of the penultimate deglaciation. *Nature* 404:61–66
  35. Hoskins BJ, Simmons AJ (1975) A multi-layer spectral model and the semi-implicit method. *Q J R Meteorol Soc* 101:1231–1250
  36. Huybers P, Wunsch C (2005) Obliquity pacing of the late Pleistocene glacial terminations. *Nature* 434:491–494. doi:10.1038/nature03401
  37. Imbrie J, Imbrie JZ (1980) Modeling the climatic response to orbital variations. *Science* 207:943–953
  38. IPCC (2007) Summary for Policymakers. In: *Climate change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge and New York
  39. Iwashima T, Yamamoto R (1986) Time-space spectral model of low order barotropic system with periodic forcing. *J Meteorol Soc Jpn* 64:183–196
  40. Kennett JP, Houtz RE, Andrews PB, Edwards AE, Gostin VA, Hajos M, Hampton M, Jenkins DG, Margolis SV, Ovenshine AT, Perch-Nielsen K (1974) Development of the circum-Antarctic current. *Science* 186:144–147
  41. Knorr G, Lohmann G (2003) Southern Ocean Origin for the resumption of Atlantic thermohaline circulation during deglaciation. *Nature* 424:532–536
  42. Knorr G, Lohmann G (2007) Rapid transitions in the Atlantic thermohaline circulation triggered by global warming and meltwater during the last deglaciation. *Geochem Geophys Geosyst* 8(12), Q12006:1–22. doi:10.1029/2007GC001604
  43. Kwasniok F, Lohmann G (2008) Underlying Dynamics of Glacial

- Millennial-Scale Climate Transitions Derived from Ice-Core Data. *Phys Rev E* (accepted)
44. Lawver LA, Gahagan LM (2003) Evolution of Cenozoic seaways in the circum-Antarctic region. *Palaeogeography, Palaeoclimatology, Palaeoecology* 198:11–37. doi:10.1016/S0031-0182(03)00392-4
  45. Lisiecki LE, Raymo ME (2005) A Pliocene-Pleistocene stack of 57 globally distributed benthic O-18 records. *Paleoceanography* 20:PA1003. doi:10.1029/2004PA001071
  46. Lohmann G (2003) Atmospheric and oceanic freshwater transport during weak Atlantic overturning circulation. *Tellus* 55A: 438–449
  47. Lohmann G, Gerdes R (1998) Sea ice effects on the Sensitivity of the Thermohaline Circulation in simplified atmosphere-ocean-sea ice models. *J Climate* 11:2789–2803
  48. Lohmann G, Schneider J (1999) Dynamics and predictability of Stommel's box model: A phase space perspective with implications for decadal climate variability. *Tellus* 51A:326–336
  49. Lohmann G, Schulz M (2000) Reconciling Boelling warmth with peak deglacial meltwater discharge. *Paleoceanography* 15:537–540
  50. Lorenz EN (1963) Deterministic nonperiodic flow. *J Atmos Sci* 20:130–141
  51. Lorenz EN (1976) Nondeterministic theories of climatic change. *Quat Res* 6:495–506
  52. Lorenz EN (1982) Atmospheric predictability experiments with a large numerical model. *Tellus* 34:505–513
  53. Lorenz EN (1990) Can chaos and intransitivity lead to interannual variability? *Tellus* 42A:378–389
  54. Lorenz S, Lohmann G (2004) Acceleration technique for Milankovitch type forcing in a coupled atmosphere-ocean circulation model: method and application for the Holocene. *Climate Dyn* 23(7–8):727–743. doi:10.1007/s00382-004-0469-y
  55. Lund R, Reeves J (2002) Detection of undocumented change-points: A revision of the two-phase regression model. *J Climate* 15:2547–2554
  56. Lunkeit F, Bauer SE, Fraedrich K (1998) Storm tracks in a warmer climate: Sensitivity studies with a simplified global circulation model. *Clim Dyn* 14:813–826
  57. Manabe S, Stouffer RJ (1993) Century-scale effects of increased atmospheric CO<sub>2</sub> on the ocean-atmosphere system. *Nature* 364:215–218
  58. Maraun D, Kurths J (2004) Cross wavelet analysis. Significance testing and pitfalls. *Nonlin Proc Geoph* 11:505–514
  59. Maraun D, Kurths J (2005) Epochs of phase coherence between El Niño/Southern Oscillation and Indian monsoon. *Geophys Res Lett* 32:L15709. doi:10.1029/2005GL023225
  60. Maslin MA, Li XS, Loutre MF, Berger A (1998) The contribution of orbital forcing to the progressive intensification of Northern Hemisphere Glaciation. *Quat Sci Rev* 17:411–426
  61. Maslin MA, Ridgeway A (2005) Mid-Pleistocene Revolution and the eccentricity myth. *Special Publication of the Geological Society of London* 247:19–34
  62. Milankovitch M (1941) *Kanon der Erdbestrahlung*. Royal Serb Acad Spec Publ, Belgrad, 132, Sect. Math Nat Sci 33:484
  63. Mori H (1965) A Continued-Fraction Representation of the Time-Correlation Functions *Prog Theor Phys* 33:423–455. doi:10.1143/PTP.34.399
  64. North Greenland Ice Core Project members (2004) High-resolution record of Northern Hemisphere climate extending into the last interglacial period. *Nature* 431:147–151
  65. Paillard D (1998) The timing of Pleistocene glaciations from a simple multiple-state climate model. *Nature* 391:378–381
  66. Palmer TN (1996) Predictability of the atmosphere and oceans: From days to decades. In: Anderson DTA, Willebrand J (eds) *Large-scale transport processes in oceans and atmosphere*. NATO ASI Series 44. Springer, Berlin, pp 83–155
  67. Parker DE, Jones PD, Folland CK, Bevan A (1994) Interdecadal changes of surface temperature since the late nineteenth century. *J Geophys Res* 99:14,373–14,399
  68. Peixoto JP, Oort AH (1992) *Physics of Climate*. American Institute of Physics, New York, p 520
  69. Petit JR, Jouzel J, Raynaud D, Barkov NI, Barnola JM, Basile I, Bender M, Chappellaz J, Davis M, Delaygue G, Delmotte M, Kotlyakov VM, Legrand M, Lipenkov VY, Lorius C, Pepin L, Ritz C, Saltzman E, Stievenard M (1999) Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. *Nature* 399:429–436
  70. Raymo M, Ganley K, Carter S, Oppo DW, McManus J (1998) Millennial-scale climate instability during the early Pleistocene epoch. *Nature* 392:699–701
  71. Reddy SC, Schmidt P, Henningson D (1993) Pseudospectra of the Orr-Sommerfeld operator. *SIAM J Appl Math* 53:15–47
  72. Rial JA (1999) Pacemaking the Ice Ages by Frequency Modulation of Earth's Orbital Eccentricity. *Science* 285:564–568
  73. Rial JA (2004) Abrupt Climate Change: Chaos and Order at Orbital and Millennial Scales. *Glob Plan Change* 41:95–109
  74. Ridgwell AJ, Watson AJ, Raymo ME (1999) Is the spectral signature of the 100 Kyr glacial cycle consistent with a Milankovitch origin? *Paleoceanography* 14:437–440
  75. Rogers JC (1985) Atmospheric circulation changes associated with the warming over the northern North Atlantic in the 1920s. *J Climate Appl Meteorol* 24:1303–1310
  76. Romanova V, Lohmann G, Grosfeld K, Butzin M (2006) The relative role of oceanic heat transport and orography on glacial climate. *Quat Sci Rev* 25:832–845. doi:10.1016/j.quascirev.2005.07.007
  77. Saltzman (2002) *Dynamical Paleoclimatology. Generalized Theory of Global Climate Change*. In: *International Geophysics Series*, vol 80. Harcourt-Academic Press (Elsevier Science), San Diego, p 354
  78. Schulz M, Paul A, Timmermann A (2004) Glacial-Interglacial Contrast in Climate Variability at Centennial-to-Millennial Timescales: Observations and Conceptual Model. *Quat Sci Rev* 23:2219
  79. Seidel DJ, Lanzante JR (2004) An assessment of three alternatives to linear trends for characterizing global atmospheric temperature changes. *J Geophys Res* 109:D14108. doi:10.1029/2003JD004414
  80. Stocker TF (1998) The seesaw effect. *Science* 282:61–62
  81. Stocker TF, Johnsen SJ (2003) A minimum thermodynamic model for the bipolar seesaw. *Paleoceanography* 18(4):1087
  82. Stommel H (1961) Thermohaline Convection with Two Stable Regimes of Flow. *Tellus* 13:224–230
  83. Tiedemann R, Sarnthein M, Shackleton NJ (1994) Astronomic time scale for the Pliocene Atlantic  $\delta^{18}\text{O}$  and dust flux records of Ocean Drilling Program site 659. *Paleoceanography* 9:19–638
  84. Timmermann A, Lohmann G (2000) Noise-Induced Transitions in a simplified model of the thermohaline circulation. *J Phys Oceanogr* 30(8):1891–1900
  85. Timmermann A, Oberhuber J, Bracher A, Esch M, Latif M,

- Roeckner E (1999) Increased El Niño frequency in a climate model forced by future greenhouse warming. *Nature* 398:694–696
86. Torrence C, Compo G (1998) A practical guide to wavelet analysis. *Bull Amer Meteor Soc* 79:61–78
87. Trefethen LN, Trefethen AE, Reddy SC, Driscoll TA (1993) Hydrodynamic stability without eigenvalues. *Science* 261:578–584
88. Trenberth KE (1990) Recent observed interdecadal climate changes in the Northern Hemisphere. *Bull Am Meteorol Soc* 71:988–993
89. Uhlenbeck GE, Ornstein LS (1930) On the theory of Brownian Motion. *Phys Rev* 36:823–841
90. Wunsch C (1999) The interpretation of short climate records, with comments on the North Atlantic and Southern Oscillation. *Bull Amer Meteor Soc* 80:245–255
91. Wunsch C (2004) Quantitative estimate of the Milankovitch-forced contribution to observed Quaternary climate change. *Quat Sci Rev* 23(9–10):1001–1012
92. Yamamoto R, Iwashima T, Sanga NK (1985) Climatic jump: a hypothesis in climate diagnosis. *J Meteorol Soc Jpn* 63:1157–1160
93. Zachos J, Pagani M, Sloan L, Thomas E, Billups K (2001) Trends, Rhythms, and Aberrations in Global Climate 65 Ma to Present. *Science* 292(5517):686–693
94. Zwanzig R (1980) Thermodynamic modeling of systems far from equilibrium. In: Garrido L (ed) *Lecture Notes in Physics* 132, in *Systems Far From Equilibrium*. Springer, Berlin

### Books and Reviews

- Dijkstra HA (2005) *Nonlinear Physical Oceanography*, 2nd revised and extended edition. Springer, New York, pp 537
- Hansen J, Sato M, Kharecha P (2007) Climate change and trace gases. *Phil Trans R Soc A* 365:1925–1954. doi:10.1098/rsta.2007.2052
- Lockwood JG (2001) Abrupt and sudden climate transitions and fluctuations: a review. *Int J Climat* 21:1153–1179
- Rial JA, Pielke RA Sr, Beniston M, Claussen M, Canadell J, Cox P, Held H, N deNoblet-Ducudre, Prinn R, Reynolds J, Salas JD (2004) *Nonlinearities, Feedbacks and Critical Thresholds Within the Earth's Climate System*. *Clim Chang* 65:11–38
- Ruddiman WF (2001) *Earth's Climate. Past and Future*. WH Freeman, New York, p 465
- Stocker TF (1999) Abrupt climate changes from the past to the future—a review. *Int J Earth Sci* 88:365–374

## Brittle Tectonics: A Non-linear Dynamical System

CHRISTOPHER H. SCHOLZ

Lamont-Doherty Earth Observatory,  
Columbia University, New York, USA

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Scaling Relations: Self-similarity](#)

[Earthquake and Fault Populations: Self-organization](#)

[Models of the System](#)

[Future Directions](#)

[Bibliography](#)

### Glossary

**Ductile shear zone** A quasi-planar tabular zone of localized shear deformation in the semi-brittle to fully plastic regimes.

**Earthquake** Dynamically running shear instability on a fault.

**Fault** A shear crack with friction between its interfaces.

**Mylonite** A metamorphic rock with a fabric produced by shear deformation.

**Suprafault** The shear relaxation structure that includes a fault and its associated ductile shear zone.

### Definition of the Subject

Brittle deformation is the primary mode of deformation of Earth's crust. At the long timescale it is manifested by faulting, and on the short timescale by earthquakes. It is one of the best-known examples of a system exhibiting self-organized criticality. A full understanding of this system is essential to the evaluation of earthquake hazard.

### Introduction

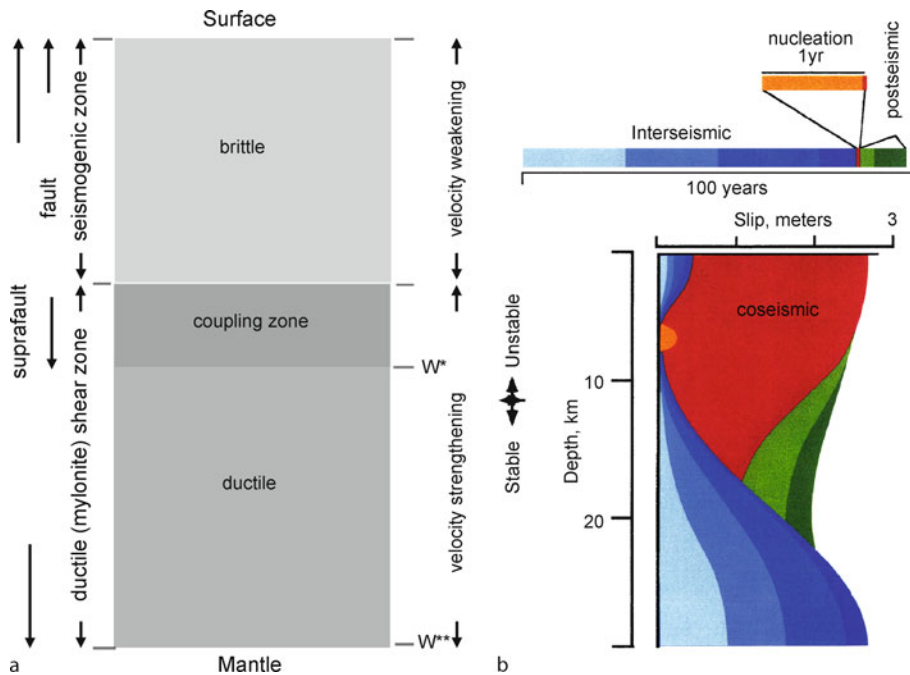
The upper part of Earth's crust is brittle and under a state of all-round compression. It responds to deformation by faulting: the formation and propagation of shear cracks. The crack walls support normal stresses and hence fault propagation must overcome not only the rupture resistance of the fault tips but friction between its interior interfaces. This friction is usually velocity weakening, such that any slippage results in stick-slip instability. The resulting

dynamically running crack-like shear instability radiates elastic waves, producing the shaking known as an earthquake. Thus brittle tectonics separates into two greatly different timescales: the long timescale of the faulting and the short timescale of the earthquakes. Faults propagate quasi-statically at rates set by tectonic driving rates, whereas earthquakes propagate at a rate set by the shear velocity of the rock. Almost all displacement on continental faults occurs by earthquakes, so that there is a coupling between the two timescales.

This system is illustrated in Fig. 1a. The term *fault* is usually restricted to the rupture of the brittle part, but large faults continue downwards as ductile shear zones, which, in the case of major faults, may extend as deep as the crust-mantle boundary. We will refer to the entire shear release structure, including both the fault and its corresponding ductile shear zone as the *suprafault*.

The division between brittle and ductile behavior occurs at what is usually called the brittle-ductile transition. This transition is gradual, with a change in deformation mechanism from brittle crack propagation and friction to fully plastic deformation through a region of mixed behavior known as semi-brittle. For our purposes, however, it is more useful to divide the suprafault into two regimes, of velocity weakening and velocity strengthening rheologies, at a depth  $S$ . These regions are predicted by a rate/state variable friction law that has been established in the laboratory (for a review, see [25]). Here, we also refer to the ductile regime as velocity strengthening because the creep law for velocity strengthening friction is the same form as for ductile creep.  $S$  corresponds to the upper onset of the brittle-ductile transition and it varies regionally because it depends on temperature (heat flow) and strain rate. For faults in continental crust it varies between about 10–20 kms. Earthquakes can only nucleate above  $S$ , so this constitutes the *seismogenic zone*. When earthquakes become large, they can, owing to the very high strain rates at their rupture tips, propagate below  $S$  to  $W^*$ , which defines the maximum width of earthquakes. Similarly, the maximum width of a suprafault is  $W^{**}$ .

Although the ductile region is velocity (strain-rate) strengthening, it is also, seemingly paradoxically, strain softening, which leads to strain localization into narrow shear zones. As the rock shears, it gradually reorganizes internally to form a mylonitic fabric, which facilitates further shearing. With finite strain, the mylonitic fabric develops further and the mylonitic foliation rotates closer to the plane of shear, progressively reducing the rock's resistance to shear in that particular direction and plane. It is this mechanism of *fabric weakening* that produces the shear localization.



Brittle Tectonics: A Non-linear Dynamical System, Figure 1

**a** A schematic longitudinal section of the rheological components of a crustal suprafault. **b** A simulation of a model with the properties of (a), shown throughout a seismic cycle (after Tse and Rice [37])

The system has two characteristic length scales,  $W^*$  and  $W^{**}$ . An earthquake nucleates within the seismogenic zone and initially propagates in all directions along its perimeter, acting as a 3D crack. When its dimension exceeds  $W^*$ , it has breached the free surface and penetrated to  $W^*$  and then is prohibited from growing deeper. It is then restricted to propagating in the horizontal direction, acting as a 2D crack. Thus there is a symmetry breakage at the dimension  $W^*$ . *Small* earthquakes, with dimensions smaller than  $W^*$ , are not self-similar with *large* earthquakes, those with lengths larger than  $W^*$ . The same occurs for suprafaults at the dimension  $W^{**}$ .

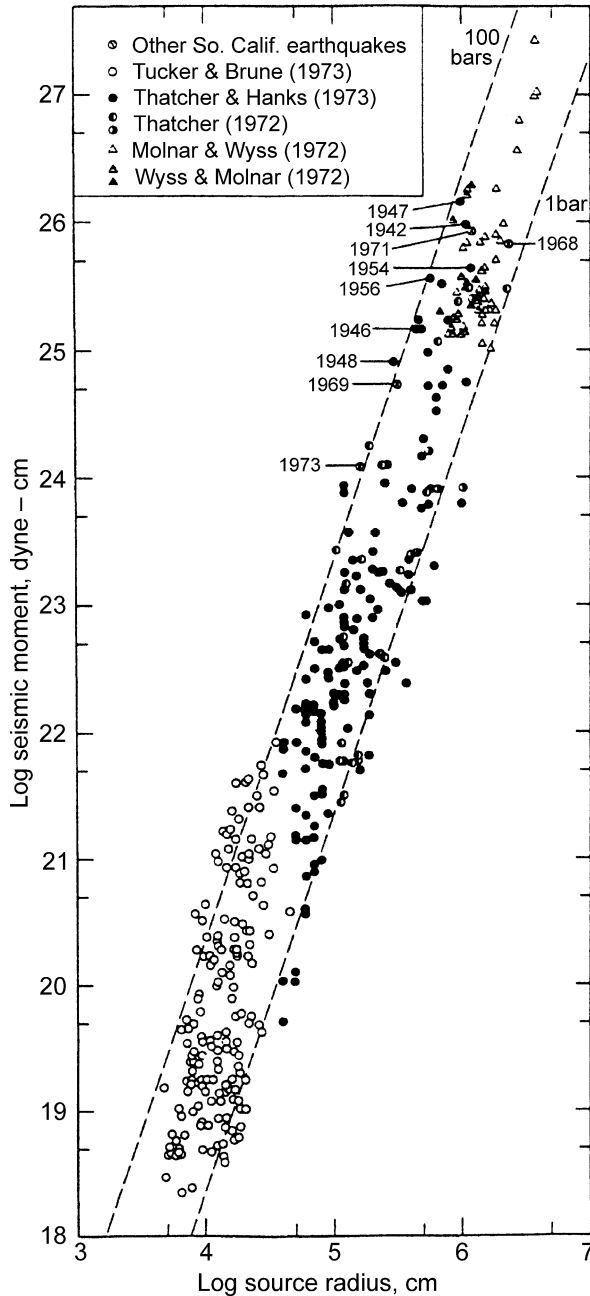
The coupling of the system is illustrated in Fig. 1b. This simulation shows the deformation of a strike slip suprafault (one in which the slip is horizontal), over the seismic cycle, i.e. the recurrence time of a large earthquake. The recurrence time depends on the long-term (geologic) slip rate of the suprafault. During the *interseismic* period, the ductile region shears at a constant rate and gradually propagates upwards. This loads the upper region until nucleation occurs above  $S$  as a quasi-static crack that eventually goes dynamic, resulting in an earthquake that propagates over  $S$  and down to  $W^*$ . The ductile region is instantaneously reloaded by the downward propagation of the earthquake. This results in *postseismic* relaxation that exponentially decays over a period of years to decades.

## Scaling Relations: Self-similarity

### Earthquakes

The primary scaling relation for cracks is that between the shear displacement (slip),  $D$ , for faults or  $\Delta D$  for earthquakes, and the dimension  $L$  of the rupture. The ratio of these parameters is a stress-drop  $\Delta\sigma$  normalized by an elastic constant. Figure 2 shows data for small earthquakes, in which seismic moment  $M_0 = \mu\Delta DA$  is plotted vs.  $L$  ( $\mu$  is the shear modulus and  $A$  is rupture area). Because these are 3D cracks the  $L^3$  lines are loci of equal stress-drop. Although there is considerable scatter, the data follow this trend. Thus, while  $\Delta\sigma$  varies by about 2 orders of magnitude among earthquakes, it is scale-invariant. These earthquakes are thus self-similar.

A plot of slip vs. length for large earthquakes is shown in Fig. 3. What is seen is that initially  $\Delta D$  increases linearly with  $L$  and, following a long crossover, becomes constant. From the static analysis of a 2D crack, we would expect, for constant stress-drop, that  $\Delta D$  be proportional to  $W^*$ , and hence constant on a plot like this, but this is seen only for the rare earthquakes with aspect ratios greater than 10. A static scaling analysis suggests that the crossover be complete by an aspect ratio of 4. A numerical model of dynamic shear cracking produced the same delayed crossover [30]. Further study of the same model



Brittle Tectonics: A Non-linear Dynamical System, Figure 2  
 A collection of data for small earthquakes showing the relationship between seismic moment  $M_0 = \mu \Delta D A$  and source radius. Dashed lines are of constant stress drop. From Hanks [15]

suggests that the long rollover results from significant slip below the seismogenic depth [31]. The scaling of large earthquakes thus exhibit two self-similar regimes, neither of which is self-similar with small earthquakes.

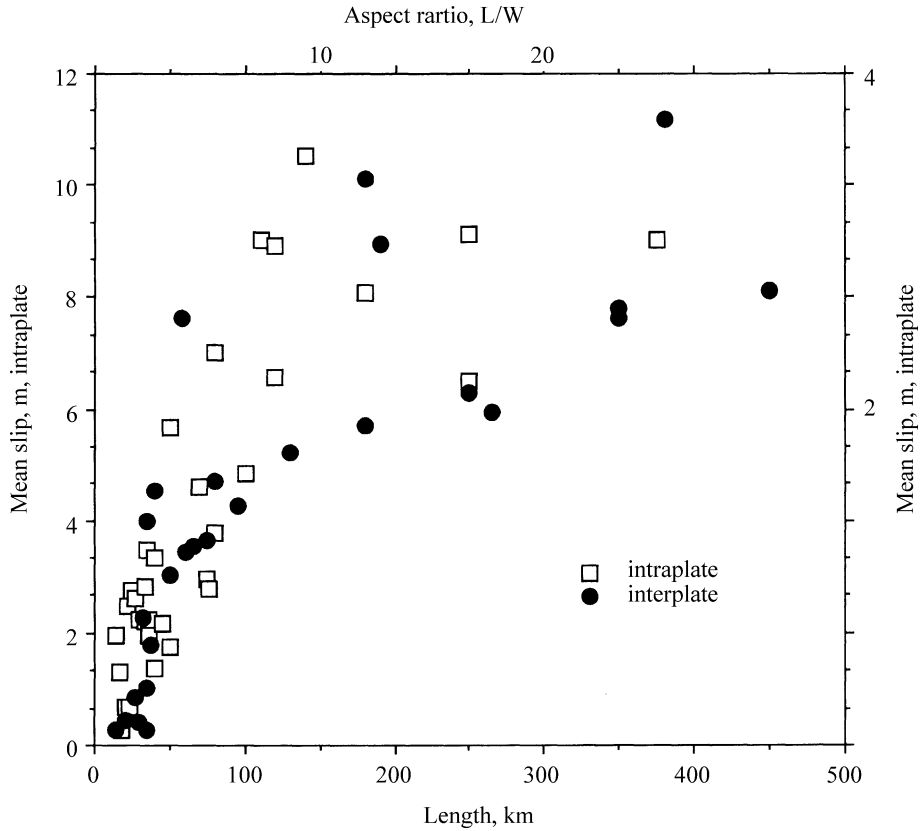
The two  $\Delta D$  scales in Fig. 3, for interplate and intraplate earthquakes, illustrate the rate effect on stress drop. Earthquakes on intraplate faults, for which geological sliding rates are several orders of magnitude less than for interplate faults, have, on average, stress drops about three times larger than interplate earthquakes. Such fault healing is expected from the rate/state friction constitutive law [4].

## Faults

The displacement profiles for some faults in the same rock type, ranging in length from 690 to 2200 m, are shown, normalized by length, in Fig. 4. The data collapse shows the self-similarity of these profiles. Displacement amplitude scales linearly with length, whereas the linear displacement tapers as the tips are approached are scale invariant. A global plot of  $D$  vs.  $L$  is shown in Fig. 5. There is, similar to earthquakes, considerable scatter, but the overall trend shows a linear relationship between  $D$  and  $L$ . Here more can be said about the origin of the scatter than in the case of earthquakes. Those faults with  $D/L$  ratios less than  $10^{-2}$  are in soft sedimentary rock at shallow depth; those with greater  $D/L$  are in strong metamorphic or igneous rock. Faults of length greater than 5 km extend increasingly deeper into the crust and exhibit an increase in  $D/L$  (stress-drop) owing to the increase of strength with pressure. Scatter is also produced by the interaction of faults through their stress fields, which distorts their displacement profiles (see pp. 126–129 in [27]). These effects of lithology and fault interaction also occur for the displacement tapers near the fault tips, which are another measure of strength [28]. In a study in which the faults are in a single lithology and were selected to be isolated [13], the scatter in both  $D/L$  and fault tip taper was correspondingly reduced (as in Fig. 4).

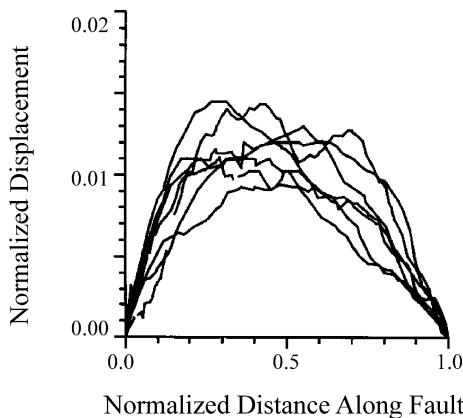
We do not have data for long enough faults to observe a crossover from 3D to 2D scaling, as we did for earthquakes. Because  $W^{**} \sim 40$  km, we would need data from faults of length exceeding 400 km to clearly see this effect. Although faults of that length and longer exist, they are predominately plate boundaries, which have no tips because their ends connect to other plate boundaries and are stress relieved by them. Therefore the scaling of Fig. 5 does not apply to them. The San Andreas fault in California, for example, terminates at its southern end at a spreading ridge. Its net slip is more than an order of magnitude greater than would be expected from simple application of the scaling shown in Fig. 5.

Faults, like earthquakes, are self-similar. The scaling parameters relating  $D$  and  $L$  differ greatly, however. For



Brittle Tectonics: A Non-linear Dynamical System, Figure 3

Mean slip vs. rupture length for large earthquakes. Two scaling regimes are observed. Slip increases approximately linearly with  $L$  up to an aspect ratio of about 10, and is independent of length thereafter. After Scholz [22]



Brittle Tectonics: A Non-linear Dynamical System, Figure 4  
 Displacement profiles along faults of different length in the Volcanic Tablelands of eastern California, normalized to length. Data from Dawers et al. [13]

faults it is  $\sim 10^{-2}$  and for earthquakes,  $10^{-4}$ – $10^{-5}$ . Both represent stress-drops. For faults, it is the stress-drop from the fracture strength of intact rock to the residual frictional strength. For earthquakes, it is the stress-drop from static to dynamic friction.

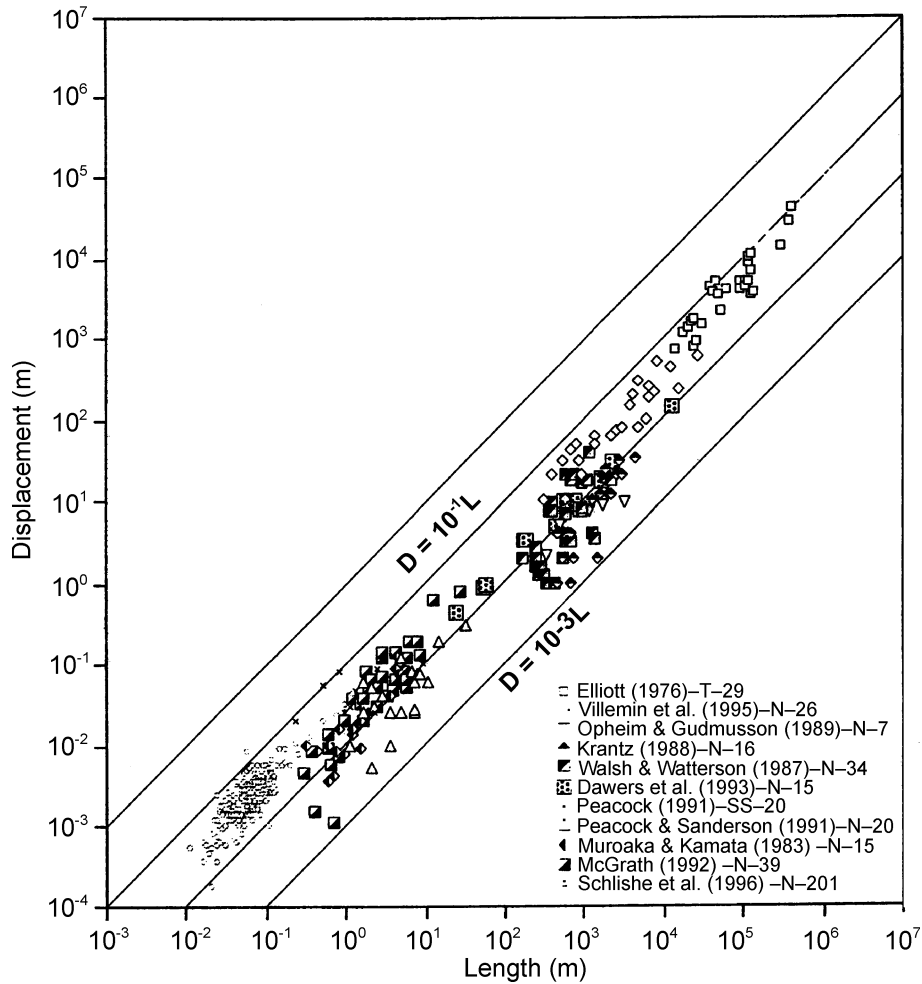
**Earthquake and Fault Populations:  
 Self-organization**

Earthquakes and faults both self-organize into populations with well defined statistical properties, thus exhibiting complexity in the physics of this system.

**Earthquakes**

The frequency  $N(M)$  of earthquakes greater than magnitude  $M$  has long been known to obey the Gutenberg-Richter law,  $\log N(M) = a - bM$  in which  $b \approx 1$  is found universally for small earthquakes. In laboratory experiments  $b$  is found to decrease with stress [21]. The fact that  $b$  is narrowly limited in the Earth’s crust may be be-





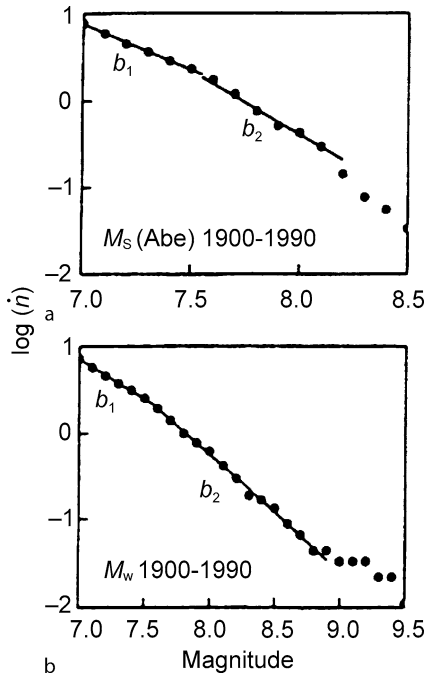
Brittle Tectonics: A Non-linear Dynamical System, Figure 5

Maximum slip vs. length for fault data sets from various tectonic settings. Modified from Schlishe et al. [20]

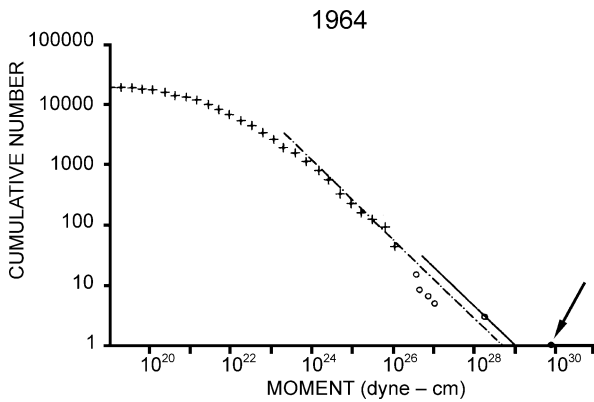
cause stress in the crust is limited to a narrow range by the frictional strength of faults [36].  $M$  is a logarithmic measure of energy  $E$  or seismic moment,  $M_0$ . In terms of  $M_0$  this relation is  $N(M_0) = aM_0^{-B}$  where  $B = 2/3b$ . However, as shown in Fig. 6, in the global catalog of earthquakes there is a crossover from  $b = 1$  to  $b = 1.5$  at about  $M7.5$ . This corresponds to a rupture dimension of about 50 km, which is an average value of  $W^*$  for subduction zones, which dominate this data set. For a subset of continental earthquakes, the same crossover is observed at  $M6.5$ , corresponding to a rupture length of 10–20 km, similar to  $W^*$  for crustal faults. Thus in terms of a linear measure of earthquake size like  $E$  or  $M_0$ , earthquakes have a power law size distribution. Where there is a breakdown of self-similarity between small and large

earthquakes there is a corresponding change in the exponent  $B$  in this distribution.

The size distribution discussed above applies to regions large enough to include many active faults or plate boundary segments. If we study instead a single fault or plate boundary segment that ruptures in a single large earthquake and consider earthquakes that occur within a single seismic cycle of that earthquake, a different picture emerges. An example is shown in Fig. 7. There we see that the small earthquakes during that interval obey a power law with  $B = 2/3$ , as usual, but that the single large earthquake is more than an order of magnitude greater than expected from extrapolation of the small events. This is another manifestation of large and small earthquakes belonging to different fractal sets.



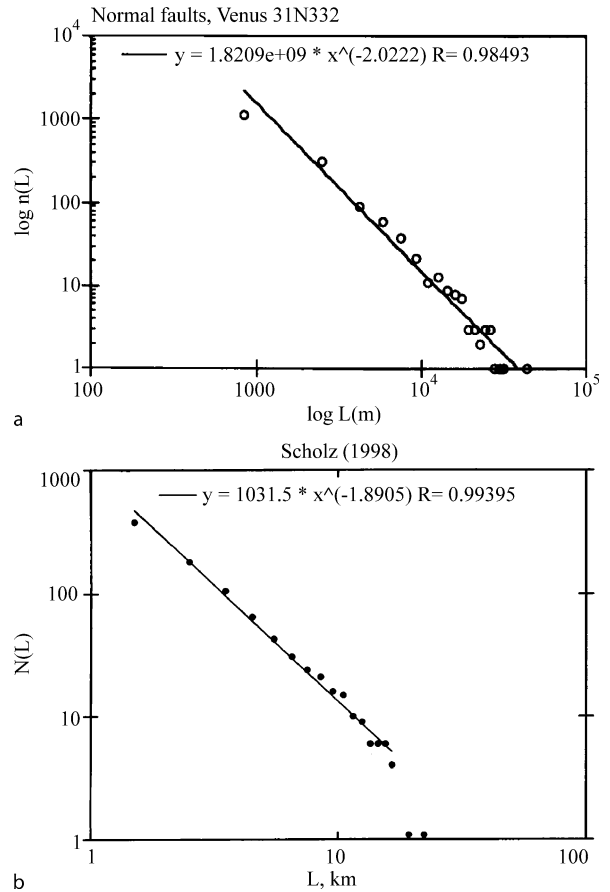
Brittle Tectonics: A Non-linear Dynamical System, Figure 6  
 The frequency size distribution from a global cataloger of Pacheco and Sykes [18] for two magnitude scales. From Pacheco et al. [19]



Brittle Tectonics: A Non-linear Dynamical System, Figure 7  
 Distribution of small earthquakes in the rupture zone of the 1964 Alaskan  $M_{9.4}$  earthquake, normalized to the recurrence time of that earthquake. An arrow indicates the 1964 earthquake. It is about  $1\frac{1}{2}$  times larger than an extrapolation of the small earthquakes would indicate. The roll off at  $M_0 < 3 \times 10^{23}$  dyne cm is cause by a lack of perceptibility of smaller events. From Davison and Scholz [11]

**Faults**

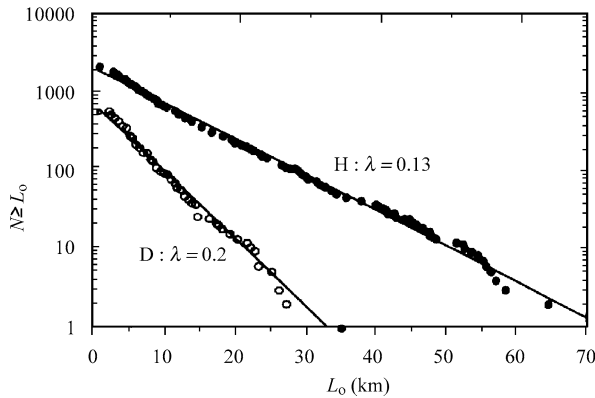
Fault data are usually obtained by geological mapping. Because mapping has a dynamic range limited to about one



Brittle Tectonics: A Non-linear Dynamical System, Figure 8  
 Power law size distributions of faults. a Frequency-length distribution for faults on the plains of Venus from a Magellan SAR image [23]. b Cumulative length distribution of subfaults of the San Andreas fault [26]

order of magnitude, it is difficult to convincingly demonstrate a power law size distribution with this method. When the results of mapping at several scales are combined [29], the case for a power law is strengthened, but the exponent is poorly determined. This problem was overcome with the data in Fig. 8a, from a Magellan SAR image of normal faults on the plains of Venus. This image has a far greater dynamic range, and unequivocally demonstrates a power law distribution, with an exponent  $-2$ . This is a frequency plot, rather than a cumulative plot such as shown in Fig. 7.

On close examination, faults are found to be not a continuous surface, but a complex of myriad strands or subfaults. A cumulative size distribution of the subfaults of the San Andreas fault is shown in Fig. 8b, where they are also shown to obey a power law. The upper fractal limit of the subfaults appears to be  $W^*$ .



Brittle Tectonics: A Non-linear Dynamical System, Figure 9  
 The distribution of faults at two places on the flanks of mid-ocean ridges. These distributions are exponential, rather than power law. From Cowie et al. [9]

The size distributions of earthquakes and faults are not unrelated. If one equates faults with large earthquakes and subfaults with small earthquakes, then by convolving over the fault distributions of Fig. 8 one can obtain the earthquake distributions of Figs. 6 and 7 [24,26].

Not all fault distributions are power law, however. Figure 9 shows the size distribution of faults on the flanks of the East Pacific Rise, obtained by sonar swath mapping. These clearly obey an exponential rather than a power law distribution. One can also find examples of periodic distributions, such as the regularly spaced faults of the Basin and Range province of the Western US. We show later that three regimes, characterized by power law, exponential, and periodic distributions, evolve with increasing strain.

### Models of the System

The first model of the earthquake system was a bench-top apparatus of a string of weights connected by springs and pulled along a frictional surface [6]. Their main finding was that this model obeyed the Gutenberg–Richter law. Interest in this result was revived by the paper of Bak, Tang, and Wiesenfeld [3], which introduced the concept of *self-organized criticality*.

Bak et al. [3] illustrated this concept with a cellular automata model of a sandpile, in which the cells are governed by a nearest neighbor rule for avalanches and the system is driven by the slow random addition of grains to the pile. The result was a power law distribution of avalanches and a  $1/f$  energy flux. From this they hypothesized that systems containing many interacting elements may exhibit general statistical properties that are described by power laws. They called this behavior self-organized

criticality (SOC). Although this concept is intuitively appealing, there is no clear-cut definition of what SOC is.

Because the Gutenberg–Richter law is the most famous of such power laws, adaptations of this cellular automata model to earthquakes soon followed (e. g. [2,5]). Carlson and Langer [7] explored a numerical spring-slider model that contained the full dynamics of the slider blocks. All these models reproduced the Gutenberg–Richter law and also explored other aspects of the system, for example, in the case of Brown et al. [5], the distribution of recurrence times of large earthquakes. Olami et al. [17] explored a non-conservative spring block model, and Christensen and Olami [8] showed how Omori’s law of aftershocks might be explained without having to appeal to an added viscosity as did Burridge and Knopoff. Sornette and Virieux [32], using a mean field theory approach, showed that the scaling difference between small and large earthquakes results in the change in  $b$  value shown in Fig. 6. Extended discussion of these models and their results can be found in Turcotte [38] and Jensen [16].

The study of the self-organization of faults was done first in an analog model of continental deformation by Davy et al. [12], who found a geometrically fractal fault distribution. Cowie et al. [10] studied a model simulating a brittle sheet undergoing stretching. For large strains this model converges on a single system size crack, but at small strains power law size crack distributions were observed.

Spyropoulos et al. [34] expanded on this with a spring-block model of the stretching of a brittle sheet supported by a ductile substrate. They found an evolution of the crack population with strain, from power law to exponential distributions and finally to a periodic distribution of system size cracks. In this model, the brittle layer contains only nearest neighbor interactions but the ductile substrate provides longer-range force interactions. It is the longer-range interactions that result in the evolution of the system and is what differs this model from the sandpile family of models. The evolution from a power law distribution to an exponential distribution of cracks has been demonstrated with physical models [1,33]. It has also been demonstrated in the field, for faults in the Asal rift of Ethiopia [14].

### Future Directions

Our understanding of brittle tectonics as a dynamical system is still very sketchy. Approaches to it tend to be piecemeal, which leads to self-limiting results. For example, although it is now well known that earthquakes interact through their stress fields (e. g. [35]), this phenomenon is modeled by assuming that the background ‘tectonic’ stress

field is uniform, thus ignoring the stress fields of faults, which also interact. The interaction of faults has so far been studied in only the simplest case, that of sub-parallel normal faults. For strike-slip faults, in which the fault tip stresses are asymmetric, more interesting possibilities exist, as yet unexplored. The San Andreas fault system in California is a complex of faults several hundred km wide that somehow act together to accommodate most of the Pacific-North America plate motion. How does this work? How do the faults interact in such a way to preserve continuity of plate motion along strike, as geodetic data indicate? How does the interaction between faults lead to the geometric evolution of fault patterns? Are fault slip rates stationary, as often assumed, or do they vary at intermediate time scales between that of earthquakes and that of the faults themselves? If so, can this explain the temporal clustering of earthquakes that seem to occur over spatial distances long compared to the stress fields of earthquakes? These are but a few of the problems one can imagine studying in the future, once the idea is accepted that the system must be considered in its entirety before being broken into digestible chunks.

## Bibliography

### Primary Literature

- Ackermann RV, Schlische RW, Withjack MO (2001) The geometric and statistical evolution of normal fault systems: an experimental study of the effects of mechanical layer thickness on scaling laws. *J Struct Geol* 23:1803–1819
- Bak P, Tang C (1989) Earthquakes as a self-organized critical phenomenon. *J Geophys Res* 94:15635–15637
- Bak P, Tang C, Wiesenfeld K (1987) Self-organized criticality: An explanation of  $1/f$  noise. *Phys Rev Lett* 59:381–384
- Beeler NM, Hickman SH, Wong TF (2001) Earthquake stress drop and laboratory-inferred interseismic strength recovery. *J Geophys Res-Solid Earth* 106:30701–30713
- Brown SR, Scholz CH, Rundle JB (1991) A simplified spring-block model of earthquakes. *Geophys Res Lett* 18:215–218
- Burridge R, Knopoff L (1967) Model and theoretical seismicity. *Bull Seism Soc Am* 57:341–362
- Carlson JM, Langer JS (1989) Properties of earthquakes generated by fault dynamics. *Phys Rev Lett* 62:2632–2635
- Christensen K, Olami Z (1992) Variation of the Gutenberg-Richter B Values and Nontrivial Temporal Correlations in a Spring-Block Model for Earthquakes. *J Geophys Res-Solid Earth* 97:8729–8735
- Cowie PA, Scholz CH, Edwards M, Malinverno A (1993) Fault strain and seismic coupling on midocean ridges. *J Geophys Res-Solid Earth* 98:17911–17920
- Cowie PA, Sornette D, Vanneste C (1995) Multifractal scaling properties of a growing fault population. *Geophys J Int* 122:457–469
- Davison F, Scholz C (1985) Frequency-moment distribution of earthquakes in the Aleutian Arc: A test of the characteristic earthquake model. *Bull Seismol Soc Am* 75:1349–1362
- Davy P, Sornette A, Sornette D (1990) Some consequences of a proposed fractal nature of continental faulting. *Nature* 348:56–58
- Dawers NH, Anders MH, Scholz CH (1993) Growth of normal faults – displacement-length scaling. *Geology* 21:1107–1110
- Gupta A, Scholz CH (2000) Brittle strain regime transition in the Afar depression: implications for fault growth and seafloor spreading. *Geology* 28:1087–1090
- Hanks TC (1977) Earthquake Stress Drops, Ambient Tectonic Stresses and Stresses That Drive Plate Motions. *Pure Appl Geophys* 115:441–458
- Jensen HJ (1998) Self-organized criticality: Emergent complex behavior in physical and biological systems. Cambridge Univ. Press, Cambridge
- Olami Z, Feder HJS, Christensen K (1992) Self-organized criticality in a continuous, nonconservative cellular automaton modeling earthquakes. *Phys Rev Lett* 68:1244–1247
- Pacheco JF, Sykes LR (1992) Seismic moment catalog of large shallow earthquakes, 1900 to 1989. *Bull Seismol Soc Am* 82:1306–1349
- Pacheco JF, Scholz CH, Sykes LR (1992) Changes in frequency-size relationship from small to large earthquakes. *Nature* 355:71–73
- Schlichte RW, Young SS, Ackermann RV, Gupta A (1996) Geometry and scaling relations of a population of very small rift-related normal faults. *Geology* 24:683–686
- Scholz CH (1968) The frequency-magnitude relation of microfracturing in rock and its relation to earthquakes. *Bull Seismol Soc Am* 58:399–415
- Scholz CH (1994) A reappraisal of large earthquake scaling. *Bull Seismol Soc Am* 84:215–218
- Scholz CH (1997) Earthquake and fault populations and the calculation of brittle strain. *Geowissenschaften* 3–4:124–130
- Scholz CH (1997) Size distributions for large and small earthquakes. *Bull Seismol Soc Am* 87:1074–1077
- Scholz CH (1998) Earthquakes and friction laws. *Nature* 391:37–42
- Scholz CH (1998) A further note on earthquake size distributions. *Bull Seismol Soc Am* 88:1325–1326
- Scholz CH (2002) The mechanics of earthquakes and faulting, 2nd edn. Cambridge University Press, Cambridge
- Scholz CH, Lawler TM (2004) Slip tapers at the tips of faults and earthquake ruptures. *Geophys Res Lett* 31:L21609, doi:10.1029/2004GL021030
- Scholz CH, Dawers NH, Yu JZ, Anders MH (1993) Fault growth and fault scaling laws – preliminary-results. *J Geophys Res-Solid Earth* 98:21951–21961
- Shaw BE, Scholz CH (2001) Slip-length scaling in large earthquakes: observations and theory and implications for earthquake physics. *Geophys Res Lett* 28:2995–2998
- Shaw BE, Wesnouski SG (2008) Slip-length Scaling in large earthquakes: The role of deep penetrating slip below the seismogenic layer. *Bull Seismol Soc Am* 98:1633–1641
- Sornette D, Virieux J (1992) Linking short-timescale deformation to long-timescale tectonics. *Nature* 357:401–403
- Spyropoulos C, Griffith WJ, Scholz CH, Shaw BE (1999) Experimental evidence for different strain regimes of crack populations in a clay model. *Geophys Res Lett* 26:1081–1084

34. Spyropoulos C, Scholz CH, Shaw BE (2002) Transition regimes for growing crack populations. *Phys Rev E* 65:056105, doi:10.1103/PhysRevE.65.056105
35. Stein RS (1999) The role of stress transfer in earthquake occurrence. *Nature* 402:605–609
36. Townend J, Zoback MD (2000) How faulting keeps the crust strong. *Geology* 28:399–402
37. Tse S, Rice J (1986) Crustal earthquake instability in relation to the depth variation of frictional slip properties. *J Geophys Res* 91:9452–9472
38. Turcotte DL (1999) Seismicity and self-organized criticality. *Phys Earth Planet Inter* 111:275–293

### Books and Reviews

- Sornette D (2003) *Critical phenomena in natural systems: Chaos, fractals, self-organization, and disorder*. Springer, Berlin
- Turcotte DL (1997) *Fractals and chaos in geology and geophysics*. Cambridge, New York

## Climate Change and Agriculture

CYNTHIA ROSENZWEIG  
NASA/Goddard Institute for Space Studies,  
Columbia University, New York, USA

### Article Outline

Glossary  
Introduction  
Conclusions  
Future Directions  
Bibliography

### Glossary

**Anthropogenic emissions** Greenhouse gas emissions that are produced as a result of humans through such developments as industry or agriculture.

**Greenhouse gases** The gases of the atmosphere that create the greenhouse effect, which keeps much of the heat from the sun from radiating back into outer space. Greenhouse gases include, in order of relative abundance: Water vapor, carbon dioxide, methane, nitrous oxide, ozone, and CFCs. Greenhouse gases come from natural sources and human activity; present CO<sub>2</sub> levels are ~380 ppmv, approximately 100 ppmv higher than they were in pre-industrial times.

**Soybean cyst nematode** *Heterodera glycines*, a plant-parasite that infects the roots of soybean, with the female becoming a cyst. Infection causes various symptoms, including a serious loss of yield.

**El Niño-southern oscillation (ENSO)** A phenomenon in the equatorial Pacific Ocean characterized by a positive sea-surface temperature departure from normal (for the 1971–2000 base period) in the Niño 3.4 region greater than or equal in magnitude to 0.5°C, averaged over three consecutive months.

**North atlantic Oscillation (NAO)** A hemispheric, meridional oscillation in atmospheric mass with centers of action near Iceland and over the subtropical Atlantic.

**Vegetative index** A simple numerical indicator used to analyze remote sensing measurements, often from space satellites, to determine how much photosynthesis is occurring in an area.

**Soil organic carbon** All the organic compounds within the soil without living roots and animals.

### Introduction

The term climate change refers to an overall shift of mean climate conditions in a given region. The warming

trend associated with anthropogenic emissions of greenhouse gases and the enhanced greenhouse effect of the atmosphere can and should be regarded as a “climate change” when viewed on the time scale of decades or a few centuries.

Climate change exacerbates concerns about agricultural production and food security worldwide. At global and regional scales, food security is prominent among the human concerns and ecosystem services under threat from dangerous anthropogenic interference in the earth’s climate [17,29,50]. At the national scale, decision-makers are concerned about potential damages that may arise in coming decades from climate change impacts, since these are likely to affect domestic and international policies, trading patterns, resource use, regional planning, and human welfare.

While agro-climatic conditions, land resources and their management are key components of food production, both supply and demand are also critically affected by distinct socio-economic pressures, including current and projected trends in population and income growth and distribution, as well as availability and access to technology and development. In the last three decades, for instance, average daily per capita intake has risen globally from 2,400 to 2,800 calories, spurred by economic growth, improved production systems, international trade, and globalization of food markets [12]. Feedbacks of such growth patterns on cultures, personal tastes, and lifestyles have in turn led to major dietary changes – mainly in developing countries – where shares of meat, fat, and sugar in total food intake have increased significantly [13]. Thus, the consequences of climate change on world food demand and supply will depend on many interactive dynamic processes.

Agriculture plays two fundamental roles in human-driven climate change. It is one of the key human sectors that will be affected by climate change over the coming decades, thus requiring adaptation measures. Agriculture is also a major source of greenhouse gases to the atmosphere, including carbon dioxide (CO<sub>2</sub>) due to land-use change and farm operations; methane (CH<sub>4</sub>) from rice production and livestock husbandry, and nitrous oxide (N<sub>2</sub>O) from application of nitrogen fertilizer. As climate changes as well as socio-economic pressures shape future demands for food, fiber and energy, synergies can be identified between adaptation and mitigation strategies, so that robust options that meet both climate and societal challenges can be developed. Ultimately, farmers and others in the agricultural sector will be faced with the dual task of contributing to global reductions of carbon dioxide and other greenhouse gas

emissions, while coping with an already-changing climate.

A changing climate due to increasing anthropogenic emissions of greenhouse gases will affect both the productivity and geographic distribution of crop and pasture species. The major climate factors contributing to these responses include increasing atmospheric carbon dioxide, rising temperature, and increasing extreme events, especially droughts and floods. These factors in turn will affect water resources for agriculture, grazing lands, livestock, and associated agricultural pests. Effects will vary, depending on the degree of change in temperature and precipitation and on the particular management system and its location. Several studies have suggested that recent warming trends in some regions may have already had discernible effects on some agricultural systems [17].

Climate change projections are uncertain in regard to both the rate and magnitude of temperature and precipitation variation in the coming decades. This uncertainty arises from a lack of precise knowledge of how climate system processes will change and of how population growth, economic and technological developments, and land-use patterns will evolve in the coming century [16,17]. Despite these uncertainties, the ultimate significance of the climate change issue is related to its global reach, affecting agricultural regions throughout the world in complex ways. After approximately two decades of research, ten major conclusions may be drawn in regard to climate change and agriculture.

### **Effects on Agricultural Systems Will Be Heterogeneous**

Global studies on projected climate change effects on agriculture show that negative and positive effects will occur both within countries and across the world. In large countries such as the United States, Russia, Brazil, and Australia, agricultural regions will likely be affected quite differently. Some regions will experience increases in production and some declines (see, e.g., [34]). At the international level, this implies possible shifts in comparative advantage for production of export crops. This also implies that adaptive responses to climate change will necessarily be complex and varied. Due to differences in global climate model projections and decadal variability, it is impossible to project exact effects in any one location for any given decade.

### **Developing Countries Are More Vulnerable**

Despite general uncertainties about the rate and magnitude of climate change and especially about consequent hydrological changes, regional and global studies have

consistently shown that agricultural production systems in the mid and high latitudes are more likely to benefit in the near term (to mid-century), while production systems in the low-latitudes are more likely to decline (Fig. 1) [17]. In biophysical terms, rising temperatures will likely push many crops beyond their limits of optimal growth and yield. Higher temperatures will intensify the evaporative demand of the atmosphere, leading to greater water stress, especially in semi-arid regions. Since most developing countries are located in lower-latitude regions (some which are indeed semi-arid) while most developed countries are located in the more humid mid- to high latitudes, this finding suggests a divergence in vulnerability between these groups of nations, with far-reaching implications for future world food security [31,37].

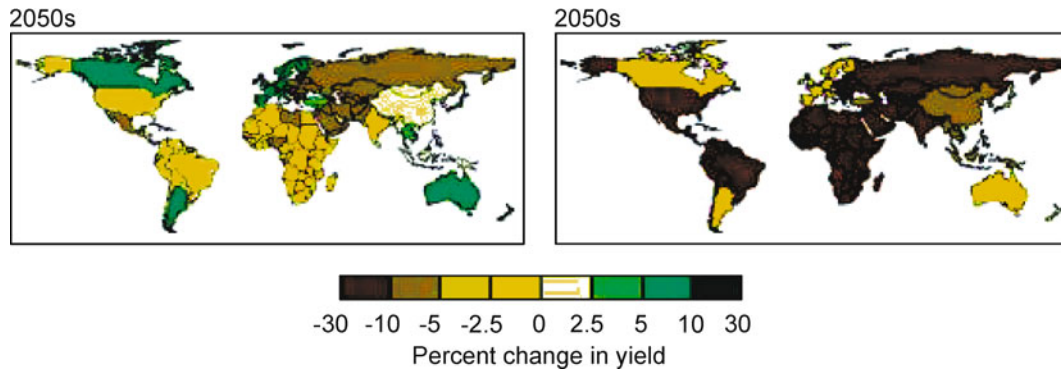
Furthermore, developing countries often have fewer resources with which to devise appropriate adaptation measures to meet changing agricultural conditions. The combination of potentially greater climate stresses and lower adaptive capacity in developing countries creates different degrees of vulnerability between rich and poor nations as they confront global warming. This difference is due in part to the potentially greater detrimental impacts of a changing climate in areas that are already warm (particularly if such areas are also dry), and in part to the generally lower levels of adaptive capacity in developing countries.

### **Development Path Matters**

Since climate is not the only driving force on agriculture, researchers now conduct scenario analysis that include linked sets of population projections, economic growth rates, energy technology improvements, land-use changes, and associated emissions of greenhouse gases [31]. Regional patterns related to economic development and adaptive capacity contribute to differing levels of climate change impacts [17]. Scenarios with higher economic growth rates and less attention to environmental issues lead to high temperatures and reduced adaptive capacity, which in turn lead to pronounced decreases in yields both regionally and globally. Scenarios with lower greenhouse gas emissions and greater attention to environmental issues lead to lower amounts of temperature rise and crop production declines.

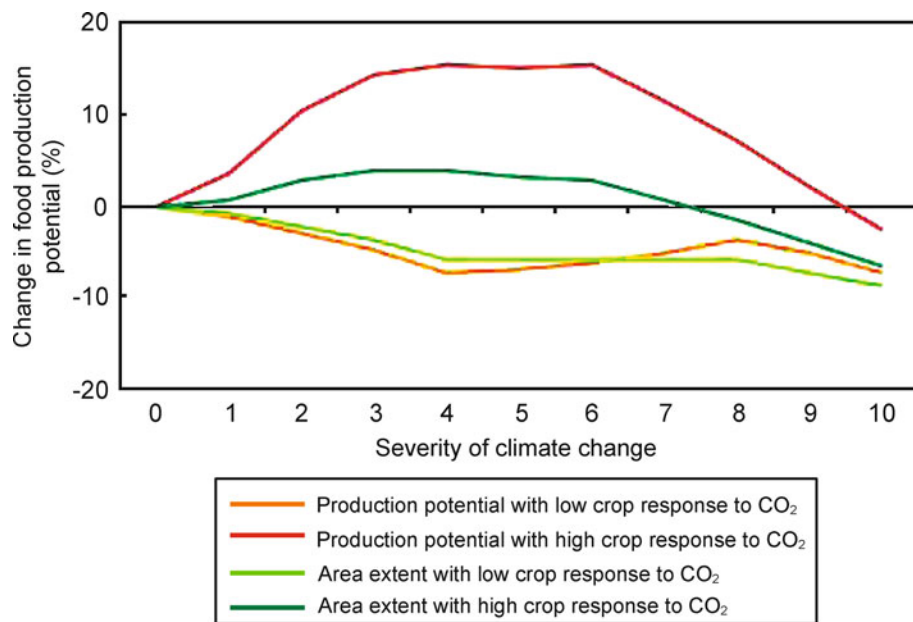
### **Long-Term Effects Are Negative for Both Developed and Developing Countries**

If the effects of climate change are not abated, production in the mid- and high-latitudes is likely to decline in the longer term (i.e., above  $\sim 3^{\circ}\text{C}$  warming) (Fig. 2) [16].



Climate Change and Agriculture, Figure 1

Potential changes (%) in national cereal yields for the 2050s (compared with 1990) under the HadCM3 SRES A1FI with (*left*) and without (*right*) CO<sub>2</sub> effects [31]



Climate Change and Agriculture, Figure 2

Change in food production potential in relation to severity of climate change [13]

These results are consistent over a range of temperature, precipitation, and direct CO<sub>2</sub> effects tested, and are due primarily to the detrimental effects of heat and water stress as temperatures rise. While the beneficial effects of CO<sub>2</sub> may eventually level out, the detrimental effects of warmer temperatures and greater water stress are more likely to be progressive in all regions. Although the precise levels of CO<sub>2</sub> effects on crops and their contribution to global crop production are still active areas of research [47], global im-

pacts are likely to turn negative in all regions sometime during the second half of the century, and perhaps before then for some major crops. For instance, by 2050 climate change is projected to have a downward pressure on yields of almonds, walnuts, avocados, and table grapes in California. Opportunities for expansion into cooler regions have been identified, but this adaptation would require substantial investments and may be limited by non-climatic constraints [24].



### Water Resources Are Key

Recent flooding and heavy precipitation events in the US and worldwide have caused great damage to crop production. If the frequency of these weather extremes were to increase in the near future, as recent trends for the US indicate and as projected by global climate models [17,48], the cost of crop losses in the coming decades could rise dramatically. US corn production losses due to excess soil moisture, already significant under current climate, may double during the next thirty years, causing additional damages totaling an estimated \$3 billion per year (Fig. 3) [41]. These costs may either be borne directly by those farmers affected or may need to be transferred to private or governmental insurance and disaster relief programs. There is also concern for tractability in the spring and water-logging in the summer in mid and high latitudes where precipitation is projected to increase.

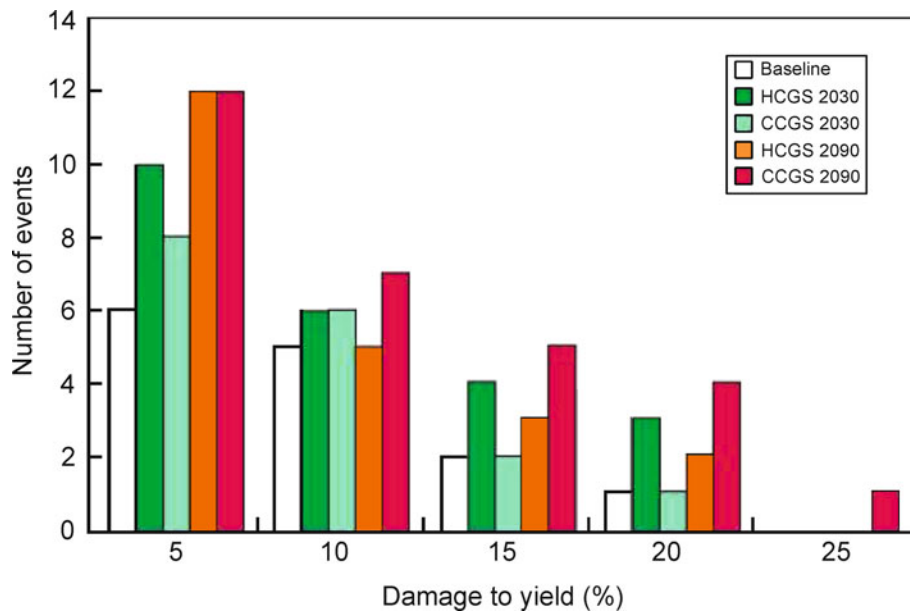
Changes in crop water demand and water availability will affect the reliability of irrigation, which competes for growing municipal and industrial demands [42]. Studies link climate change scenarios with hydrologic, agricultural, and planning models to estimate water availability for agriculture under changing climate conditions, to explore changes in ecosystem services, and to evaluate adaptation strategies for the water resources and agriculture

sectors. Major irrigated agricultural regions are very likely to be affected by changing supplies of and demands for water due a changing climate especially under conditions of expansion of irrigated lands [42].

Cultivars are available for agricultural adaptation to the projected changes, but their demand for water may be higher than currently adapted varieties. Thus, even in relatively water-rich areas, changes in water demand due to climate change effects on agriculture and increased demand from urban growth will require timely improvements in crop cultivars, irrigation and drainage technology, and water management. In tropical regions, the use of agroforestry may be an economically feasible way to protect crop plants from extremes in microclimates and soil moisture [22].

### Agricultural Pests and Diseases May Spread

Increased pest damage arises from changes in production systems, enhanced resistance of some pests to pesticides, and the production of crops in warmer and more humid climatic regions where crops are more susceptible to pests. Changes in crop management techniques, particularly the intensification of cropping, reduction in crop rotations, and increase in monocultures, have increased the activity of pests. The expansion of worldwide trade in food and



Climate Change and Agriculture, Figure 3

Number of events causing damage to maize yields due to excess soil moisture conditions, averaged over all study sites, under current baseline (1951–1998) and climate change conditions. The Hadley Center (HC) and Canadian Center (CC) scenarios with greenhouse gas and sulfate aerosols (GS) were used. Events causing a 20% simulated yield damage are comparable to the 1993 US Midwest floods [41]

plant products has also increased the impact of weeds, insects, and diseases on crops. The geographical ranges of several important insects, weeds, and pathogens in the US have recently expanded, including soybean cyst nematode (*Heterodera glycines*) and sudden death syndrome (*Fusarium solani* f. sp. *glycines*) (Fig. 4) [15,39,43].

Current climate trends and extreme weather events may be directly and indirectly contributing to the increased pest damage [17,39,52]. Downy mildew (*Plasmopara viticola*) epidemics on grape, the most serious grapevine disease in northern Italy, may increase under climate change, even though reduced precipitation may have a counterbalancing effect on disease pressure [44].

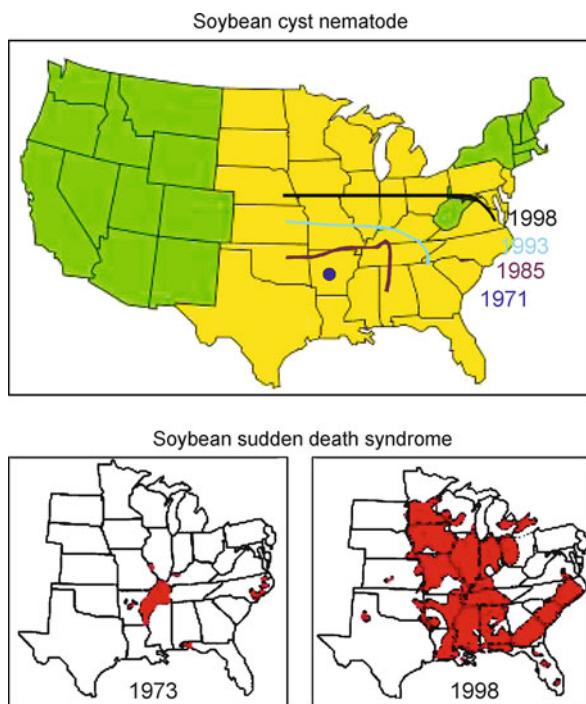
Such changes need to be put in the context of the global increases in pest-induced losses of crops in all regions since the 1940s [30,33] and the more than 33-fold increase in both the amount and toxicity of pesticide used over the same period [33]. Climate change thus may exacerbate environmental and public health issues related to agricultural chemicals [39], since increased applications of agricultural chemicals are likely to be needed in response to increasing disease pressure. Improved knowledge of the

effects of climate on host–pathogen interactions will contribute to the adaptive capacity of agro-ecosystems.

### Current Climate Stress Is a Key Entry Point for Climate Change

There is an important interplay between current and future climate stresses. Since farmers have dealt with climatic fluctuations since the advent of agriculture, improving strategies for dealing with present climate extremes – such as droughts, floods, and heatwaves – is an important way to prepare for climate change. Many agricultural regions are affected by the major climate variability systems, including the processes known as the El Niño–Southern Oscillation (ENSO) and the North Atlantic Oscillation (NAO) [36]. The El Niño phase of the ENSO cycle tends to bring rainfall to Uruguay, while La Niña brings drought, as shown in Fig. 5 for 1998, an El Niño year, and 2000, the following La Niña.

In terms of prediction tools, ENSO models provide the opportunity for testing and validation of climate prediction and assessment on shorter seasonal-to-interannual time-scales. Skill in predicting climate changes on shorter time-scales, particularly the ENSO periods of the last twenty years when good observations exist, may lend credence to projections of global warming over the longer-term. As global climate models are further developed with improved parametrizations and higher spatial resolution, they are likely to improve simulations of ENSO and other large-scale variability processes. The interaction of these systems with underlying anthropogenic trends caused by increasing greenhouse gas concentrations in the atmosphere is an active area of contemporary climate science. For regions directly affected by ENSO and other systems, such changes, if they do indeed occur, may become important manifestations of global warming.

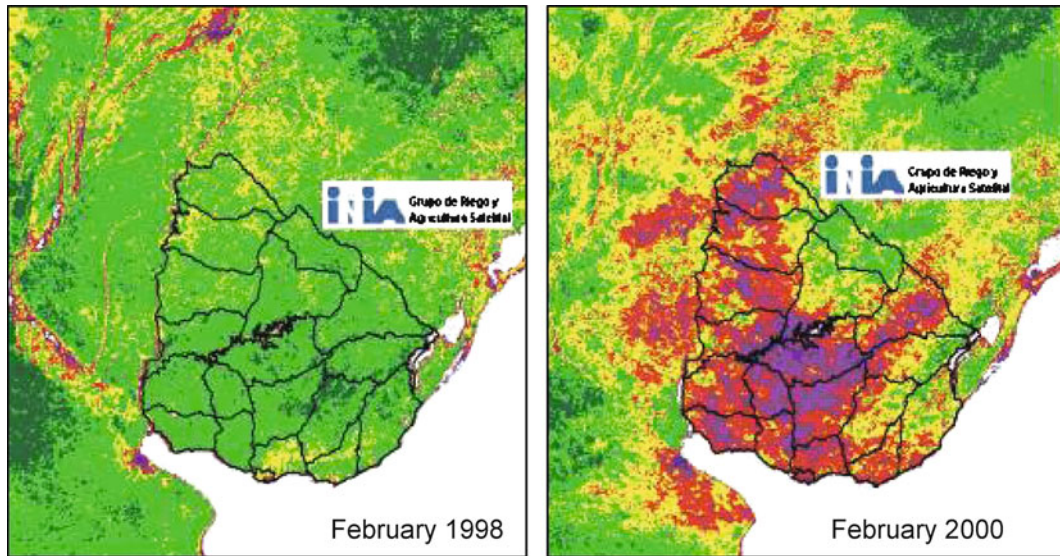


Climate Change and Agriculture, Figure 4  
Range expansion of soybean cyst nematode (*Heterodera glycines*) from 1971 to 1989 (top) and soybean sudden death syndrome (*Fusarium solani* f. sp. *Glycines*) from 1973 to 1998 (bottom) in North America [40]

### Adaptation Is Necessary

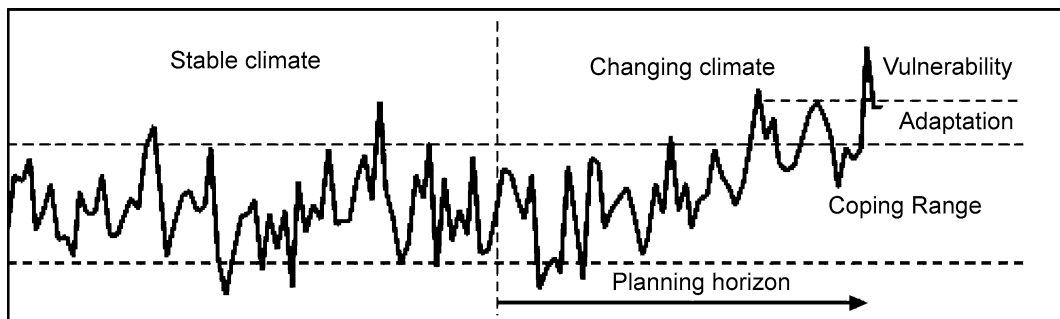
‘Coping range’ is a useful paradigm for improving responses to climate stresses of today and preparing for the climate changes of tomorrow [20]. An agricultural system may currently exist within a ‘coping range’ of climate variability that may be exceeded as incidence of extreme events increases under changing climate conditions (Fig. 6). The goal is to increase the coping range over which an agricultural system may thrive under such changes through the process of adaptation.

Adaptation can help farmers to minimize negative impacts of climate on human activities and ecosystems and to take advantage of potential beneficial changes. *Adaptation* to climate change can be defined as the range of ac-



Climate Change and Agriculture, Figure 5

Vegetative Index (NDVI) for El Niño (1998) and La Niña (2000) years in Uruguay. *Green* = adequate water conditions; *red/purple* = drought conditions [2]



Climate Change and Agriculture, Figure 6

Coping range of climate variability (adapted from [20])

tions taken in response to changes in local and regional climatic conditions [45]. Adaptation responses include both *autonomous adaptation* actions (i. e., those taken spontaneously and independently by individual farmers), and *planned adaptation* actions (i. e., those facilitated by climate-specific regulations and incentives put in place by regional, national and international policies) (see Table 1) [17]. In terms of the multiple factors impinging on agriculture, however, system responses to socio-economic, institutional, political or cultural pressures may outweigh response to climate change alone in driving the evolution of agricultural systems. The *adaptive capacity* of a system, in the context of climate change, can be viewed as the full set of system skills – i. e., technical solutions available to farmers in order to respond to climate stresses – as deter-

mined by the socio-economic and cultural settings, plus institutional and policy contexts, prevalent in the region of interest [17].

While current agronomic research confirms that at the field level crops would respond positively to elevated CO<sub>2</sub> in the absence of climate change (e. g., [1,17,19,21]), the associated impacts of high temperatures, altered patterns of precipitation, and possibly increased frequency of extreme events (such as drought and floods) are likely to require a range of adaptation responses, some of which are listed in Table 2.

#### Mitigation Reduces Long-Term Risk

Agriculture has an important role to play in mitigation of climate change. *Mitigation* is defined as intervention

Climate Change and Agriculture, Table 1  
Adaptation approaches to climate impacts on agriculture

Approach	Definition	Operation
Autonomous	Adaptation that does not constitute a conscious response to climatic stimuli but is triggered by ecological changes in natural systems and by market or welfare changes in human systems. Also referred to as spontaneous adaptation.	Crop calendar shifts (planting, input schedules, harvesting) Cultivar changes Crop-mix changes
Planned	Adaptation that is the result of a deliberate policy decision, based on an awareness that conditions have changed or are about to change and that action is required to return to, maintain, or achieve a desired state.	Land-use incentives Irrigation infrastructure Water pricing Germplasm development programs

Climate Change and Agriculture, Table 2  
Key agronomic impacts and adaptation responses

Agricultural impacts	Adaptation response
Biomass increase under elevated CO <sub>2</sub>	Cultivar selection and breeding to maximize yield
Acceleration of maturity due to higher temperature	Cultivar selection and breeding of slower maturing types
Heat stress during flowering and reproduction	Early planting of spring crops
Crop losses due to increased droughts and floods	Changes in crop mixtures and rotations; warning systems; insurance
Increased pest damage	Improved management; increased pesticide use; biotechnology

aimed at reducing the severity of climate change by reducing the atmospheric concentration of greenhouse gases, either by reducing emissions or by enhancing sinks. There are several major ways that the agricultural sector can contribute to climate change mitigation.

**Soil Carbon Sequestration** Of the approximately 150 GT of carbon that were lost in the last century due to land conversion to agriculture and subsequent production, about two thirds were lost due to deforestation and one-third, roughly 50 GT, due to cultivation of current agricultural soils and exports as food products [50]. The latter figure represents the maximum theoretical amount of carbon that could be restored in agricultural soils. In practice, as long as 40–50% of total above-ground grain or fruit production is exported as food to non-agricultural areas, the actual carbon amount that can be restored in agricultural soils is much lower.

Efforts to improve soil quality and raise soil organic carbon (SOC) levels include crop management and conservation tillage techniques. These practices have evolved as means to enhance sustainability and resilience of agricultural systems, rather than with soil carbon sequestration as primary motivation. They include so-called “best practice” agricultural techniques, such as use of cover crops and/or nitrogen fixers in rotation cycles, judicious use of fertilizers and organic amendments, soil water management improvements to irrigation and drainage, and improved varieties with high biomass production.

Conventional tillage is defined [27] as the mechanical manipulation of the topsoil that leaves no more than 15% of the ground surface covered with crop residues. In contrast, no-till management is defined as the avoidance of mechanical manipulation of the topsoil so as to leave it undisturbed and covered with surface residues from harvesting the prior crop to planting the new crop.

Best agricultural practices can result in a net augmentation of soil carbon and in enhanced productivity due to better soil structure and soil moisture management. The relevant practices include precise and timely applications and spatial allocation of fertilizers, use of slow-release fertilizers, prevention of erosion, shortening or elimination of fallow periods, use of high-residue cover crops and green-manure crops, and minimized mechanical disturbance of soil (e. g., zero tillage). Altogether, such practices may lead to partial or even complete restoration of the soil’s organic carbon content where it had been depleted. In some cases, it might even be possible to store more carbon than had originally been present in the “virgin” soil. Where the soils had been severely degraded and their agricultural productivity greatly impaired, they may be converted to grassland or afforested so as to serve as carbon sinks.

The overall potential for carbon storage depends in each case on such factors as climate, type of vegetation, topography, depth and texture of the soil, past use (or abuse), and current management.

Along with sequestering carbon, these practices have the potential to improve soils in developing countries. In

areas such as West Africa, soil fertility depletion has been described as the single most important constraint to food security [4]. Studies in smallholder agricultural farms in Africa have already illustrated significant increases in system carbon and productivity through organic-inorganic resources management (Roose and Barthes [35])

**Biofuels** Agriculture may help to mitigate anthropogenic greenhouse emissions through the production of biofuels. As has been demonstrated by ethanol based on corn production, issues involved with biofuel production include potential competition with food production, increased pollution from fertilizers and pesticides, and loss of biodiversity. Biofuels derived from low-input high-diversity mixtures of native grassland perennials can provide more usable energy, greater greenhouse gas reduction, and less agrichemical pollution per hectare than corn grain ethanol or soybean biodiesel [46]. The higher net energy results arise because perennial grasses require lower energy inputs and produce higher bioenergy yield. Furthermore, all aboveground biomass of the grasses can be converted to energy, rather than just the seed of either corn or soybean. These perennial grasses also sequester carbon at significant rates [46].

**Other Greenhouse Gases** Because of the greater global warming potential (GWP) of methane (21) and nitrous oxide (310) compared to carbon dioxide (1), reductions of non-CO<sub>2</sub> greenhouse gas emissions from agriculture can be quite significant and achieved via the development of more efficient rice (for methane) and livestock production systems (for both methane and nitrous oxide). In intensive agricultural systems with crops and livestock production, direct CO<sub>2</sub> emissions are predominantly connected to field crop production and are typically in the range of 150–200 kg C ha<sup>-1</sup> yr<sup>-1</sup> [14,51]. Recent full greenhouse gas analysis of different farm systems in Europe showed that such CO<sub>2</sub> emissions represent only 10–15% of the farm total, with emissions of CH<sub>4</sub> contributing 25–30% and emissions of N<sub>2</sub>O accounting for as much as 60% of total CO<sub>2</sub>-equivalent greenhouse gas emissions from farm activities. The N<sub>2</sub>O contribution arises from substantial nitrogen volatilization from fertilized fields and animal waste, but it is also a consequence of its very high GWP.

In Europe, methane emissions are mostly linked to cattle digestive pathways; its contribution also dominates that of CO<sub>2</sub>, due in part to methane's high GWP. Mitigation measures for methane production in livestock include improved feed and nutrition regimes, as well as recovery of bio-gas for on-farm energy production. Effective reduction of N<sub>2</sub>O emissions is more difficult, given the largely

heterogeneous nature of emissions in space and time and thus the difficulty of timing fertilizer applications and/or manure management. Large uncertainties in emission factors also complicate the assessment of efficient N<sub>2</sub>O-reduction strategies. Current techniques focus on reduction of absolute amounts of fertilizer nitrogen applied to fields, as well as on livestock feeding regimes that reduce animal excreta.

### Climate Change Effects on Agriculture Are Occurring Already

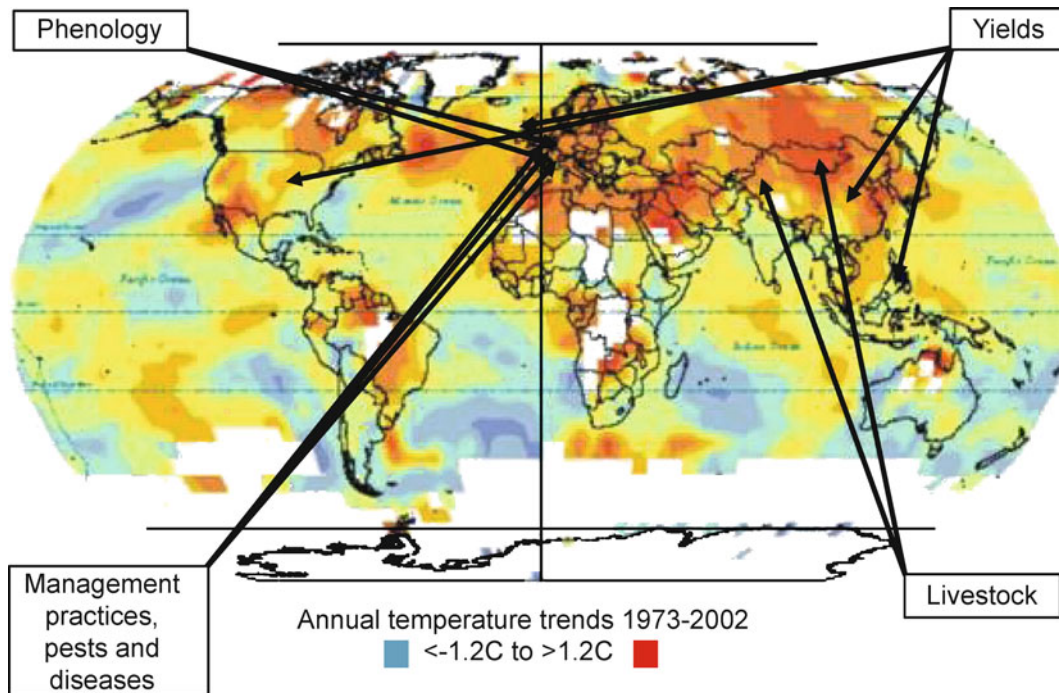
Agricultural effects of observed climate changes are being documented in many regions of the world (Fig. 7). Changes in crop phenology provide important evidence of responses to recent regional climate change. Such changes are apparent in perennial crops, such as fruit trees and wine-making varieties of grapes, which are less dependent on yearly management decisions by farmers than annual crops and are also often easier to observe.

Phenological changes are often observed in tandem with changes in management practices by farmers. Between 1951 and 2004 in Europe, agricultural crops have advanced 2.1 days/decade in response to recent warming in spring [28]. In Sahelian countries, increasing temperature in combination with rainfall reduction has led to a reduced length of vegetative period, no longer allowing present varieties to complete their cycle [5].

A negative effect of warming for local rice production has been observed by the International Rice Research Institute (IRRI) in the Philippines (yield loss of 15% for 1°C increase of growing-season minimum temperature in the dry season) [32]; a similar effect has been noted on hay yield in the UK (1°C increase in July–August led to a 0.33 t/ha loss) [6]. At the county level, US maize and soybean yields are demonstrating a positive effect of cooler and wetter years in the Midwest and hotter and drier years in the North-west plains [23]. In the case of the Sahel region of Africa, warmer and drier conditions have served as a catalyst for a number of other factors that have accelerated a decline in groundnut production [49]. For livestock, one study in Tibet reports a significant relationship of improved performance with warming in high mountainous conditions [9], while pasture biomass in Mongolia has been negatively affected by the warmer and drier climate, as observed at a local station [3] or at the regional scale by remote sensing [10].

### Conclusions

Climate change brings both challenges and opportunities to agriculture. Farmers and researchers are being



Climate Change and Agriculture, Figure 7

Locations of observed changes in agriculture in response to climate changes

called on to simultaneously adapt to and mitigate climate change through a myriad of activities involving management practices, crop breeding, and new production systems. Some of these can be mutually re-enforcing, especially in view of the projected increased climate variability under climate change. This is because, most mitigation techniques currently considered in agriculture, including reduced tillage, were originally designed as “best practice” management strategies, aimed at enhancing the long-term stability and resilience of cropping systems in the face of climate variability or of increased cultivation intensity. By increasing the ability of soils to hold soil moisture and to better withstand erosion, and by enriching ecosystem biodiversity through the establishment of more diversified cropping systems, mitigation techniques implemented locally for soil carbon sequestration may also help cropping systems to better withstand droughts and/or floods, both of which are projected to increase in frequency and severity in future warmer climates. As climate change progresses, agriculture will continue to play a leading role in responding to a dynamic environment.

### Future Directions

The time has come to incorporate climate as an essential factor in development planning and implementation. In the past, responses to climate variability were often too

narrowly focused and lacking in institutional fit. Development programs are now beginning to include recommendations to mainstream responses to climate variability and change.

Magalhães [25] gave as an example of the need for a broad focus when considering climate in planning the early drought policies in Northern Brazil. Responses had focused on improving the water-supply infrastructure (e.g. building dams and digging wells) rather than on redressing the social and economic vulnerabilities and the need to build human capital by means of education, institutions, and market incentives for sustainability. Adaptive policies should be broadly conceived so as to increase and secure household entitlements, to change land-use patterns that lead to degradation, and to develop means of support for inhabitants that are less sensitive to the vagaries of climate.

At what levels or scales of organization (national, regional, household, and individual) can the variability of climate and the sustainability of agricultural production be addressed effectively? Because many nations encompass several and often most numerous climatic zones, the challenge faced by national agricultural managers and policy make is to foster sustainability at the regional level while building a foundation from the bottom-up at the individual as well as at the household and community levels. Dil-

ley [8] believes that greater benefits of food security could be realized if knowledge were made more readily available at the household level, thereby improving the ability of more people to make even small adjustments based on anticipated climatic conditions. Multiple-scale efforts are clearly needed, with pathways of communication among the various levels and sectors of society.

Regions do not exist in isolation, as evidenced by the effects that extreme climate events occurring along the Atlantic coast of South America have on the sustainability of the inland Amazon rainforest. At least part of the pressure to deforest the Amazon region arises from the westward migration of farmers from Northern Brazil who suffer from ENSO-related droughts there (Magalhães [25]). Thus, policies related to sustainability and climate issues need to take regional interactions and their direct and indirect linkages into account.

Beyond interconnections among regions within a nation, there are the larger national concerns of economics, finance, and international relations that may affect the range of climate-adaptation policy choices. Engaging with the international community on issues of climate variability and change can lead to capacity building in both developed and developing countries, as climate and societal processes are studied and as improved understanding is incorporated into policies. This can be accomplished through interactions with international bodies dealing with climate variability and change, such as the World Meteorological Organization and the United Nations Framework Convention on Climate Change. There is a growing realization that climate plays an important role in sustainable development: It is a component of natural capital, an occasional trigger to socio-economic crises caused by extreme events, and a long-term component of global environmental change.

## Bibliography

- Ainsworth EA, Long SP (2005) What have we learned from 15 years of free-air CO<sub>2</sub> enrichment (FACE)? A meta-analysis of the response of photosynthesis, canopy properties and plant production to rising CO<sub>2</sub>. *New Phytologist* 165:351–275
- Baethgen WE, Giménez A (2002) Seasonal climate forecasts and the agricultural sector of Uruguay. In: *Examples of ENSO-Society Interactions*. The International Research Institute for Climate and Society, New York. <http://iri.columbia.edu/climate/ENSO/societal/resource/example/Baethgen.html>
- Batimaa P (2005) The potential impact of climate change and vulnerability and adaptation assessment for the livestock sector of Mongolia. *Assessments of Impacts and Adaptations to Climate Change*. AIACC, Washington DC, 20 pp
- Bationo A, Kihara J, Vanlauwe B, Waswa B, Kimetu J (2007) Soil organic carbon dynamics, functions and management in West African agro-ecosystems. *Agric Syst* 94:13–25
- Ben Mohamed A, Duivenbooden NV, Abdoussallam S (2002) Impact of climate change on agricultural production in the Sahel, Part 1. Methodological approach and case study for millet in Niger. *Clim Chang* 54:327–348
- Cannell MGR, Palutikof JP, Sparks TH (1999) *Indicators of climate change in the UK*. DETR, London, 87 pp
- Chmielewski FM, Muller A, Bruns E (2004) Climate changes and trends in phenology of fruit trees and field crops in Germany, 1961–2000. *Agric Forest Meteorol* 121(1-2):69–78
- Dilley M (2003) Regional responses to climate variability in Southern Africa. In: O'Brian K, Vogel C (eds) *Coping with climate variability: The use of seasonal climate forecasts in Southern Africa*. Ashgate, Hampshire, pp 35–47
- Du MY, Kawashima S, Yonemura S, Zhang XY, Chen SB (2004) Mutual influence between human activities and climate change in the Tibetan Plateau during recent years. *Global Planet. Change* 41:241–249
- Erdenetuya M (2004) Application of remote sensing in climate change study: Vulnerability and adaptation assessment for grassland ecosystem and livestock sector in Mongolia project. *AIACC Annual Report*, Washington DC
- Fischer G, Shah M, Velthuizen H, Nachtergaeal FO (2001) *Global agro-ecological assessment for agriculture in the 21st century*. International Institute for Applied Systems Analysis. IIASA, Laxenburg
- Fischer G, Shah M, van Velthuizen H (2002) *Climate change and agricultural vulnerability, special report to the UN World Summit on Sustainable Development, Johannesburg 2002*. IIASA, Laxenburg
- Fischer, Shah GM, Tubiello FN, van Velhuizen H (2005) Socio-economic and climate change impacts on agriculture: An integrated assessment, 1990–2080. *Philos Trans R Soc B-Biol Sci* 360(1463):2067–2083
- Flessa H, Ruser R, Dörsch P, Kamp T, Jimenez MA, Munch JC, Beese F (2002) Integrated evaluation of greenhouse gas emissions (CO<sub>2</sub>, CH<sub>4</sub>, N<sub>2</sub>O) from two farming systems in southern Germany. *Agric Ecosyst Environ* 91:175–189
- Hartman GL, Noel GR, Gray LE (1995) Occurrence of soybean sudden death syndrome in east-central Illinois and associated yield losses. *Plant Disease* 79:314–318
- IPCC (2007) *Climate change 2001: The scientific basis*. Contributions of Working Group I to the fourth assessment report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge
- IPCC (2007) *Climate change 2007: Impacts, adaptation, and vulnerability*. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge
- IPCC (2007) *Climate Change 2001: Mitigation*. Contribution of Working Group III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, 1076 pp
- Jablonski LM, Wang X, Curtis PS (2002) Plant reproduction under elevated CO<sub>2</sub> conditions: A meta-analysis of reports on 79 crop and wild species. *New Phytologist* 156(1):9–26
- Jones PD, Mann ME (2004) Climate over past millennium. *Rev Geophys* 42:RG2002
- Kimball BA, Kobayashi K, Bindi M (2002) Response of agriculture crops to free-air CO<sub>2</sub> enrichment. *Adv Agron* 77:293–368

22. Lin BB (2007) Agroforestry management as an adaptive strategy against potential microclimate extremes in coffee agriculture. *Agric Forest Meteorol* 144(1-2):85–94
23. Lobell DB, Asner GP (2003) Climate and management contributions to recent trends in US agricultural yields. *Science* 299:1032
24. Lobell DB, Field CB, Cahill KN, Bonfils C (2006) Impacts of future climate change on California perennial crop yields: Model projections with climate and crop uncertainties. *Agricult Forest Meteorol* 141:208–218
25. Magalhães AR (2000) Sustainable development: Climate and policy linkages. In: *Proceedings of the International Forum on Climate Prediction, Agriculture and Development*, 26–28 April. International Research Institute for Climate Prediction, Palisades, New York, pp 3–10
26. Magalhães AR, Glantz MH (eds) (1992) Socioeconomic impacts of climate variations and policy response in Brazil. *Esquel Brazil Foundation, Brasília*
27. Marland G, West TO, Schlamadinger B, Canella L (2003) Managing soil organic carbon in agriculture: The net effect on greenhouse gas emissions. *Tellus* 55B:613–621
28. Menzel A, von Vopelius J, Estrella N, Schleip C, Dose V (2006) Farmers' annual activities are not tracking speed of climate change. *Climate Res* 32:201–207
29. Millennium Ecosystem Assessment (2005) *Ecosystem and human well-being: Synthesis*. Island, Washington DC
30. Oerke EC, Dehne HW, Schohnbeck F, Weber A (1995) Crop production and crop protection: Estimated losses in major food and cash crops. Elsevier, Amsterdam, 830 pp
31. Parry ML, Rosenzweig C, Iglesias A, Livermore M, Fischer G (2004) Effects of climate change on global food production under SRES emissions and socio-economic scenarios. *Glob Environ Chang* 14:53–67
32. Peng SB, Huang JL, Sheehy JE, Laza RC, Visperas RM, Zhong XH, Centeno GS, Khush GS, Cassman KG (2004) Rice yields decline with higher night temperature from global warming. *Proc Natl Acad Sci USA* 101(27):9971–9975
33. Pimentel D (1997) Pest management in agriculture, In: Pimentel D (ed) *Techniques for reducing pesticide use: Environmental and economic benefits*. Wiley, Chichester, pp 1–12
34. Reilly J, Tubiello F, McCarl B, Abler D, Darwin R, Fuglie K, Hollinger S, Izaurrealde C, Jagtap S, Jones J, Mearns L, Ojima D, Paul E, Paustian K, Riha S, Rosenberg N, Rosenzweig C (2003) US agriculture and climate change: New results. *Clim Chang* 57:43–69
35. Roose E, Barthes B (2001) Organic matter management for soil conservation and productivity restoration in Africa: a contribution from Francophone research. *Nutrient Cycling in Agroecosystems* 61(1–2):159–170
36. Rosenzweig C, Hillel D (2008) *Climate variability and the global harvest*. Oxford University Press, Oxford
37. Rosenzweig C, Parry ML (1994) Potential impacts of climate change on world food supply. *Nature* 367:133–138
38. Rosenzweig C, Tubiello F (2007) The interactions of adaptations and mitigation strategies in agriculture. *Mit Adapt Strategies Glob Change* 12(5):855–873
39. Rosenzweig C, Iglesias A, Yang XB, Epstein PR, Chivian E (2000) Implications of climate change for US agriculture: Extreme weather events, plant diseases, and pests. Center for Health and the Global Environment, Harvard Medical School. Cambridge, 56 pp
40. Rosenzweig C, Iglesias A, Yang XB, Epstein PR, Chivian E (2000) Climate change and extreme weather events: Implications for food production, plant diseases, and pests. *Global Change and Human Health* 2(2)
41. Rosenzweig C, Tubiello FN, Goldberg R, Mills E, Bloomfield J (2002) Increased crop damage in the US from excess precipitation under climate change. *Glob Environ Chang* 12:197–202
42. Rosenzweig C, Strzepek KM, Major DC, Iglesias A, Yates DN, McClusky A, Hillel D (2004) Water resources for agriculture in a changing climate: International case studies. *Glob Environ Chang* 14:345–360
43. Roy KW, Rupe JC, Hershman DE, Abney TS (1997) Sudden death syndrome. *Plant Dis* 81:1100–1111
44. Salinari F, Giosue S, Tubiello FN, Rettori A, Rossi V, Spanna F, Rosenzweig C, Gullino ML (2006) Downy mildew (*Plasmopara viticola*) epidemics on grapevine under climate change. *Glob Chang Biol* 12(7):1299–1307
45. Smit B, Burton I, Klein RJT et al (2000) An anatomy of adaptation to climate change and variability. *Climate Chang* 45(1):223–251
46. Tillman D, Hill J, Lehman C (2006) Carbon-negative biofuels from low-input high-diversity grassland biomass. *Science* 314:1598–1600
47. Tubiello FN, Amthor JS, Boote KJ, Donatelli M, Easterling W, Fischer G, Gifford RM, Howden M, Reilly J, Rosenzweig C (2006) Crop response to elevated CO<sub>2</sub> and world food supply – A comment on “Food for Thought ...” by Long et al. *Science* 312:1918–1921:2006. *Eur J Agron* 26(3):215–223 APR 2007
48. National Assessment Synthesis Team (2001) *Climatic Change Impacts on the US: The potential consequences of climate variability and change*. US Global Change Research Program, Washington DC
49. Van Duivenbooden N, Abdoussalam S, Mohamed AB (2002) Impact of climate change on agricultural production in the Sahel. Part 2. Case study for groundnut and cowpea in Niger. *Clim Chang* 54:349–368
50. Watson RT, Noble IR, Bolin B, Ravindranath NH, Verardo DJ, Dokken DJ (2000) IPCC special reports. Land use, land-use change, and forestry. Cambridge Univ Press, Cambridge, 324
51. West TO, Marland G (2002) A synthesis of carbon sequestration, carbon emissions, and net carbon flux in agriculture: Comparing tillage practices in the United States. *Agr Ecosyst Environ* 91:217–232
52. Yang XB, Scherm H (1997) El Niño and infectious disease. *Science* 275:739



## Climate Change and Human Health

HARTMUT GRASSL

Max Planck Institute for Meteorology,  
Hamburg, Germany

### Article Outline

Glossary

Introduction

Climate Change Impact on Human Health

Climate Change Impacts on Plants

with Consequences for Human Health

Concluding Remarks

Bibliography

### Glossary

**Health** As defined by the World Health Organization (WHO) health is the state of complete physical, mental and social well-being and not merely the absence of disease or infirmity.

**Human bioclimate** The fundamental issue in human biometeorology is the assessment of the direct health effects of the atmospheric environment from heat exchange, to solar radiation and air pollution.

**Climate change related direct health effects** Climate change always impacts on human bioclimate, presently it leads to increased heat stress and heat stress frequency, higher ultraviolet radiation doses especially in summer, longer allergic pollen seasons and new allergens as well as intensified photo-smog.

**Thermal stress and mortality** Summer heat waves in mid-latitudes and elsewhere increase without doubt mortality; hence also highlight lack of correct adaptive measures, i. e. heat waves impact most strongly in societies with lack of social cohesion.

**Global expansion of tropical diseases** The observed recent global warming has increased the incidence and enlarged the distribution of some tropical diseases due to the expansion of suitable conditions for both vectors and pathogens. A northward spread has been observed for West Nile fever, Leishmaniasis and Chikungunya fever and a climate-driven spread has in parts also been recorded for malaria, dengue fever and other vector-borne infectious diseases.

**Vector-borne diseases** In epidemiology a vector is an organism transmitting a pathogen from one of its reservoirs (e. g. ruminants, birds) to another one (e. g. human) without falling ill. Such vectors for tropical diseases are: mosquitoes, biting flies, bugs, lice, flea's and

mites. Typical vector-borne diseases are malaria, yellow fever, dengue fever, West Nile fever, Leishmaniasis, Chikungunya fever. For some of these diseases global warming is the cause of the observed expansion or intensification. The complex web of reservoir organism, pathogens, vectors and infected organisms with different dependence on climate parameters often hinders a full understanding. Hence, surprises are and will be common.

**Arbo viruses** Arbo viruses are transmitted by arthropods (arthropod-borne) to vertebrates and hence in parts also to humans. Besides yellow fever, tick borne encephalitis and dengue fever about 150 other diseases are due to virus infections by insects and spiders (arthropods). In very complex transmission cycles climatic conditions play a central role. The occurrence of unusual arbo virus infections is often related to changes in climatic conditions. Therefore, the partly dramatic global increase of some arbo virus infections is also driven at least in part by the ongoing global anthropogenic climate change.

**Arbo viruses transmitted by Aedes mosquitoes** Aedes mosquitoes and Aedes-transmitted arbo viruses such as the dengue and yellow fever viruses are a growing global threat. The primary vector of these diseases, *Aedes aegypti*, has re-emerged throughout the tropics, but also *Aedes albopictus* has emerged as one of the worst invasive species taking the role of *Aedes aegypti*. Direct human activities like global trade are mainly responsible for the spread of these vectors and global warming – indirectly anthropogenic as well – cannot be ruled out as a contributor. With further warming temperate regions like Central Europe could also become areas for *Aedes albopictus*.

**Malaria and global warming** Although the Anopheles mosquitoes transmitting the protozoae (e. g. *Plasmodium falciparum*) causing malaria are strongly dependent on temperature and suitable small water reservoirs for the larvae, the spread of malaria in recent years is more a consequence of deficiencies in public health systems of many countries rather than due to the observed global warming and concomitant precipitation changes.

**Blue-tongue disease in Europe** Since August 2006 the blue-tongue-disease of cattle and sheep (serotype 8 from South Africa) spread within months from the Netherlands to Belgium, Luxembourg, France and Germany alone at the end of 2006. The vector carrying the virus is the ceratopogonid, biting midge, *Culicoides obsoletus*. The new disease for ruminants in Western and Central Europe is primarily a conse-

quence of globalization but the extremely warm winter 2006/2007 in Western and Central Europe supported further spreading.

**Carbon dioxide fertilization and quality of food** From field experiments at elevated carbon dioxide concentrations (close to a doubling of preindustrial values) it is known that agricultural yield increases for C3 plants by 10 to 30 percent; however, frequently also reduced nitrogen content in plant tissue including seeds is observed. Hence, food quality may be lowered.

**Changes in the pollen season and new pollen** The onset of flowering of plants in mid and high latitudes is mainly triggered by temperature. Therefore, warming in recent decades has caused an earlier start of pollen in the air, often also leading to a longer pollen season and for some pollen also to higher abundance stimulated by higher carbon dioxide concentration. Hence, susceptible individuals will suffer from pollen allergy longer and even perennial allergic symptoms may become possible.

## Introduction

The basis of our life is energy from the sun, water from the skies and biomass production by plants on land and in the oceans. If we ask for the key climate parameters – given the size of the planet and its mean distance to the sun – we get a very similar answer: energy flux density of the sun, precipitation and land surface parameters, mostly determined by vegetation. Hence, climate is the key natural resource. If this resource changes rapidly, as it does now, life in all its forms is affected as well. Therefore, it is a must for decision makers to deal with climate change. Here only one facet of climate change is discussed: the health of humans. It is rather astonishing that it has not already been studied intensively since the beginning of the climate change debate, as other environmental policy decisions were nearly totally driven by health consequences for humans [12].

The consequences of climate change for the health of humans, animals and plants are complex with direct and indirect relations between causes and impacts. As any organism on land is in permanent struggle with the local weather a rather strong capacity for adaptation exists; however, the organisms are trained only with existing weather and climate variability during the life-time of an organism. For new weather extremes accompanying any climate change this adaptation capability is no longer given. Hence, a changing climate will often become a threat to health. As several health related factors may change simultaneously, “multiplying” impacts, which – if alone – would not have gone beyond existing adaptive ca-

pability, may do so. An example is heat stress during heat waves accompanied not only by high ultraviolet radiation levels but also by high near surface concentrations of the strong oxidant ozone, exacerbating the heat stress.

According to the World Health Organization (WHO) “Health” is defined as “the state of complete physical, mental and social well being and not merely the absence of disease or infirmity”. This comprehensive and also ambitious understanding of the term health can also be applied to animals and plants as well as ecosystems (the terms ecosystem health and environmental health have been used very often in recent literature). Therefore, consequences of climate change on health are not confined to diseases but also include reduced well-being or reduced strength of an organism as well as weakened functions in an ecosystem or a socio-economic system.

This contribution to the Encyclopedia of Complexity does not focus exclusively on climate change and human health (see Sect. “Climate Change Impact on Human Health”) because climate change impacts on animals and plants (see Sect. “Climate Change Impacts on Plants with Consequences for Human Health”) also have consequences for human well-being and human health. An example of the complexity is the changed composition of nutrients in grains as a consequence of higher carbon dioxide levels in the atmosphere, in addition modulated by changed climate parameters that may in turn enhance infestations of plant diseases with consequences for the composition of our food. The section on Climate Change Impacts on Plants with Consequences for Human Health tries to collect known knowledge, reports on potential threats and chances for political reactions as well as pointing out major open research questions.

## Climate Change Impact on Human Health

Whenever climate changes all ecosystems have to react, i. e. they try to adapt to changed climate that must also include new weather extremes. Typical reactions are shift of biomes, altered ecosystem composition, new geographical patterns of animal and plant diseases, new relations between predator and prey. As major so-called natural climate changes have occurred also in the recent few million years the first major question in the present anthropogenic climate change era is: Can earlier so-called abrupt climate changes be used as an example for projections into the future? The answer is no, because the mean global temperature change rate to occur in the 21st century exceeds by far even the most rapid natural ones. The largest and most rapid *global* mean temperature change rates in the recent few million years have been the collapses of

the large northern hemisphere ice sheets that have led to a mean global warming of up to 5°C and a sea level rise of about 120 m in roughly 10,000 years. Projections for the 21st century without stringent climate policy [4] range between 1.5°C and about 4°C, i. e. an acceleration of at least a factor 30. Hence, the past cannot be used as an analogue. In other words: The key goal of the United Nations Framework Convention on Climate Change (UNFCCC) that speaks of avoidance of a dangerous interference with the climate system, cannot be met without globally coordinated climate policy. It requests, in addition, a climate policy helping to stabilize greenhouse gas concentrations within a time frame that firstly natural ecosystems are able to adapt to climate change, secondly food production is not generally threatened and a sustainable economic development remains possible.

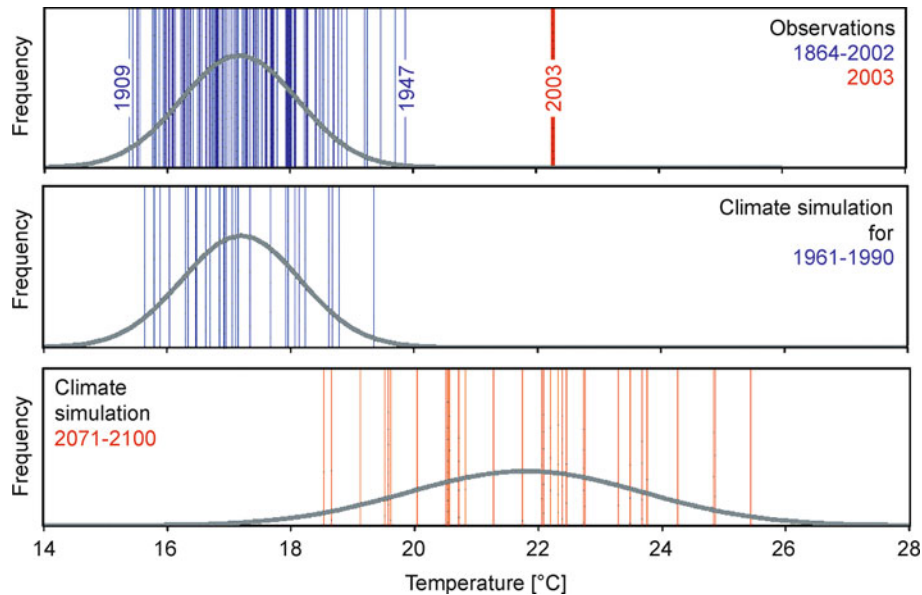
In this situation with ill-adapted forests, rapidly changed patterns of infectious diseases for plants, animals and humans looking back through climatic history does also not help directly. Besides intensified research on changed disease patterns in the very recent past a stringent globally coordinated climate policy under UNFCCC is the best insurance against massively altered disease patterns.

This paper will concentrate on climate change impacts on human health but will not exclude totally impacts on food production. Major points will be “thermal stress” be-

fore “vector-borne diseases” and also prolonged allergen seasons and new allergens are discussed.

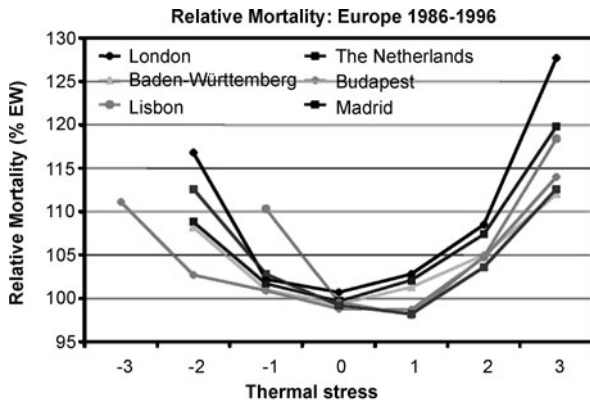
### Changed Thermal Stress

Heat or cold stress forces our body to adapt to keep the core body temperature within a narrow temperature interval of about five degrees centigrade (35 to a maximum of slightly above 40°C). While extreme cold stress events have diminished and will further diminish as a consequence of ongoing global warming, extreme heat stress will increase dramatically (see Fig. 1), when the frequency distributions of temperature at a certain location are shifted by only a few degrees centigrade and may be broadened. Up to now only very few places on the Earth’s surface exist where survival of a human being is nearly impossible. This would certainly happen if the wet bulb temperature (roughly equivalent to a ventilated sweating naked body) surmounts about 35°C. Under present climate conditions such areas do not exist, but coastal areas of the Red Sea come closest to it during on-shore winds after a sunny day that was heating surface waters to about 35°C. Hence, heat strokes can in principle be avoided by adequate behavior, if buildings are well insulated and properly ventilated and if an individual behaves. Therefore, the huge death toll caused by major heat waves, namely about 55,000 people



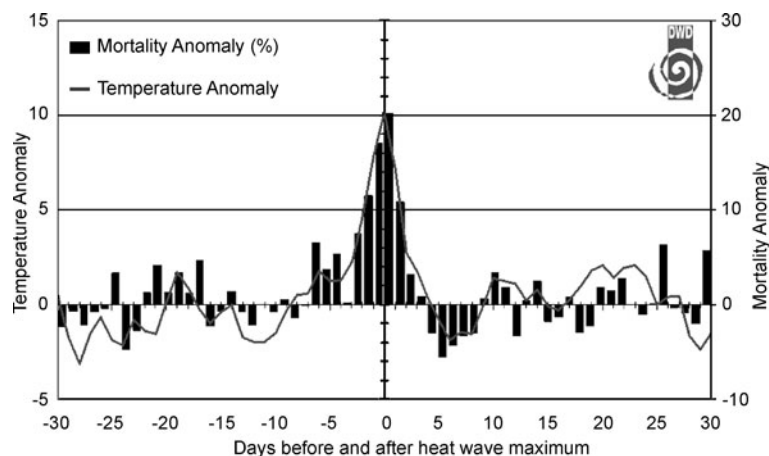
Climate Change and Human Health, Figure 1

Observed (*upper panel*) and modeled summer mean temperatures for the Swiss Plateau both for present climate (1961 to 1990) (*central panel*) and the last three decades of the 21st century for scenario A2. Please note that the exceptional summer 2003 would occur every second year, if globally coordinated climate policy would not exist. From [13]



Climate Change and Human Health, Figure 2  
 Mean relative mortality in percent for different thermal stress categories observed in Europe during 1986 to 1996. Please note that the mortality increases by more than a factor 2 when heat stress category 2 is replaced by category 3. From [5]

that died during the heat wave in summer 2003 in Europe [5] is rather an indication for “social freeze” and ill-adapted buildings in our industrialized or developing societies than for really intolerable thermal conditions. Many have died because of lack of care. New weather extremes always demask weaknesses in our security-related infrastructure. A famous example is the large difference in the number of people dying during a summer heat wave in 1995 in two US cities (Chicago and Philadelphia) just because of intensified public care in Philadelphia where the weak and poor citizens were brought by the city administration to the cooled malls during daytime.



Climate Change and Human Health, Figure 3  
 Mean mortality anomaly during a heat wave (30 days before and after its maximum) resulting from nine observed heat waves in the state of Baden Württemberg (Germany) for the period 1968 to 1997. Source: [5]

As Fig. 2 clearly demonstrates the mortality anomaly caused by heat waves is a fact observed over several decades (here in a developed country). Lowering the anomaly means not only investment in better warning systems but also enhanced social care in general.

Both heat and cold stress increase mortality as underlined by Fig. 3 for European countries. In the future heat stress category 3 will occur much more often and the lowering of the increased mortality will also be a sign of an improved public health system.

A further point to be made with respect to enhanced thermal stress as a consequence of climate change is the rapidly mounting heat stress in the inner tropics, where dew points of about 25°C will more often occur, if global warming also continues there. The ability to work with high efficiency is shrinking there rapidly with rising temperatures and dew points. As is well known economic development of developing countries needs a cooled or well-ventilated work place.

### Impact on Photochemical Smog

Photochemical smog is formed if solar radiation stimulates chemical reactions in a polluted atmosphere. Emissions of non-methane hydrocarbons and nitrogen oxides ( $\text{NO} + \text{NO}_2$ ) lead to the formation of ozone and other oxidizing toxic trace gases as well as aerosol particles. In mid-latitudes photochemical smog is typically strongest in late spring and summer and it has been the reason for some environmental policy making. Heat waves with intense solar radiation lead to major photochemical smog episodes. The higher frequency and longer duration of

Climate Change and Human Health, Table 1

Days with 8-hour mean ozone concentration above  $120 \mu\text{g m}^{-3}$  since 1990 for all stations in the German network

1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
22	22	28	23	32	29	20	22	19	21	19	21	19	51	19

Climate Change and Human Health, Table 2

Days with  $\text{PM}_{10}$  values above  $50 \mu\text{g m}^{-3}$  for the years 2001 to 2004

2001	2002	2003	2004
Stations with highest values			
117	103	132	73
Stations in cities (average)			
65	75	83	55
Remote stations in cities (average)			
22	30	38	16
Rural stations (average)			
7	12	17	5

heat waves during recent and foreseen global warming will intensify health problems already existing during heat waves. As Tables 1 and 2 demonstrate the hot summer in Germany in 2003 has increased the number of days for both ozone and aerosol load ( $\text{PM}_{10}$ ) where limit values have been surmounted in Germany dramatically [14,15], e. g. by a factor of 2 or more both for  $\text{PM}_{10}$  in rural areas and ozone in all areas.

Fighting against photochemical smog will become even more demanding in a warming world.

### Changed Ultraviolet Radiation

Ozone protects us to a large extent from the dangerous part of ultraviolet solar radiation through absorption in the ultraviolet (UV-B) range from 0.28 to  $0.32 \mu\text{m}$  wavelength. The observed latitude-dependent reduction of stratospheric ozone (from 8 to 50 km height in high latitudes), which constitutes about 90% of the total ozone column content and which is caused by chlorine compounds stemming from decay products of chlorofluorocarbons (CFCs) and other halocarbons, has increased UV-B radiation especially in spring of both hemispheres at higher latitudes. This stratospheric ozone decrease is strongest in Antarctic spring in the so-called ozone hole. This ozone depletion has certainly increased and still will increase the skin cancer and cataract incidence that has dramatically grown in recent decades. However, the ozone depletion contribution is largely buried in the variability and change of exposure of our body to UV radiation, which is due to changed behavior, especially in developed countries. The

key question for the forthcoming decades is: How fast will ozone column content recovery be after the banning of CFCs and other chlorine and bromine containing compounds? Present knowledge says: Full recovery after several decades with a chance of a super-recovery caused by a further increase of the greenhouse effect of the atmosphere, which lowers stratospheric temperatures.

### Health Effects

#### Caused by Other Changed Climate Parameters

Climate change shifts and reshapes frequency distributions of meteorological and hydrological parameters, thereby multiplying the occurrence of known extremes and leading to new ones (see Fig. 1, where this is demonstrated for temperature). Therefore, the health of many more millions is affected by intensified flooding, higher storm surges, and many other weather-related disasters.

While the highest death toll of weather-related disasters was – for thousands of years – due to droughts, with up to 10 million people dying per decade around 1930, the highest death toll is now caused by flooding and wind damage due to tropical cyclones and hurricanes (Red Cross, ...). The main reason for this change is – besides more dwellings in flood-prone areas, often already lying slightly below mean sea level – the international aid bringing food and seeds into the drought-affected areas; as long as civil war does not prevent this help.

Hence, climate change also calls – irrespective of globally coordinated climate change mitigation policies – for a coordinated climate change adaptation policy in the coming decades, because we have to adapt to the already unavoidable climate change. Mitigation measures meant to avoid the un-tolerable climate change will only become effective in decades due to the inertia of the climate system caused by the slow reaction of oceans and ice sheets. In other words: If flooding is prevented by strengthened dikes, anticipating shifted frequency distributions of precipitation, cholera epidemics will not occur. Diking has to become an international activity, as the emitters in industrialized countries are causing more flooding and sea level rise on a global scale, co-financed by the already existing but strongly to be increased adaptation fund under the UNFCCC, its Kyoto Protocol and the follow-up protocol envisaged to be signed in 2009.

As an aside I will here report on reactions of our body to high carbon dioxide levels in the atmosphere. Very often if many people are gathered in the same closed room and breathe the same air some will ask for fresh air (i. e. they require more oxygen). The need for fresh air is because of too high carbon dioxide levels. Regulations in some German states concerning ventilation in classrooms provide ventilation rules to avoid carbon dioxide concentrations above 1,500 ppm, a level after which concentration diminishes and some students may even develop signs of a beginning headache. The oxygen content of air has fallen by a bit more than a tenth of a percent only to about 20.84 percent at which point ventilation by opening windows becomes a must. Hence, ventilation of our living rooms means pushing out carbon dioxide.

### **Transmission of Infectious Diseases from Birds to Humans**

In recent years migratory birds have been named as a cause for the transmission of bird flu (avian influenza) to humans because they can in principle transmit the influenza virus to chickens, geese, turkeys and ducks in our farms from where the virus infects humans that come into close contact with the fowl or their products. However, very often the cause for the long-range transmission is global trade and tourism on the one hand plus industrialized animal husbandry in developed and emerging countries on the other. The latter has been found as the principal cause for the comparably rapid mutation of slightly pathogenic bird flu viruses to highly pathogenic ones (Bairlein and Metzger, 2008). These in turn can reach wild bird species which then can rapidly transmit them to new areas, if the viruses are only slightly pathogenic for them. But also pathogens, vectors or reservoir species can inadvertently be introduced by globalization to regions where no longer-term co-evolution could have taken place. Further global warming will intensify the shift of pathogens with the shift of (migratory) birds, but will only add to the further rapid distribution of pathogens caused by globalization (Smith et al., 2007).

### **Allergies Caused by Pollen**

In most countries allergies have recently increased dramatically. In Germany, for example, 20 to 30% of the population suffers from allergies. Most abundant is the allergic rhino conjunctivitis (hay fever), often turning into asthma bronchiale, caused by allergic pollen in air. Climate change has led to longer pollen seasons, in parts to more pollen, changed pollen spectrum and also new pollen [9].

From studies in Europe, North America and Japan an earlier flowering of 1 to 3 days per decade has been reported during the recent decades [11]. For some species, especially the late flowering ones a prolonged pollen season has been found [2,3], in parts caused also by long-range transport of the pollen. Consequently, for many people in mid-latitudes suffering from several pollen the pollen season became longer, sometimes already starting in December and ending only in October after the flowering of the neophyte ragweed (*Ambrosia artemisiifolia*) for Europe.

From differences between cities and rural areas as well as from laboratory studies it became known that pollen abundance increases with carbon dioxide concentration for some species.

In combination with air pollution allergic reactions have been shown to intensify [3].

**Invasion of Allergenic Neophytes** With ongoing global warming two processes combine for the spread of neophytes: Firstly, global trade and tourism transmitting plants and their seeds within days and weeks around the globe to all inhabited places and into ocean basins and secondly, increased temperatures allowing more and more often establishment of exotic species in new areas. A famous example where both mentioned processes work in combination is the invasion of the strongly allergenic ragweed (*Ambrosia artemisiifolia*) from North America to Europe in the 19th century with a large spread after the Second World War. Ragweed flowers from late August to September and its pollen also undergo long-range transport [1]. In recent decades it spread strongly in Central Europe facilitated by agricultural practice and by inadvertent transport with bird food as well as higher temperatures.

### **Indirect Health Effects Caused by Climate Change**

Many infectious diseases are transmitted by vectors, i. e. by animals (very often insects) transferring a pathogen from another animal and/or human without suffering from the pathogen themselves. The complicated web of hosts, pathogens, reservoirs and vectors is dependent on many factors, among them climate variables, foremost temperature and precipitation. Therefore, distribution patterns and incidence of infectious diseases will also be modified by climate change; however, disentangling the cause and effect relationships is often extremely difficult. Since tropical infectious diseases are the most common of these diseases and especially temperature sensitive they will be a focus in this section. The World Health Organization (WHO) points to a highly probable increase of morbidity

and mortality due to higher prevalence and a pole-ward shift of tropical infectious diseases due to further global warming.

### Increase of Vector-Borne Tropical Infectious Diseases

As always in scientific investigations it is especially difficult to derive long-term trends of certain variables influenced by many parameters like changes in land use, socio-economic conditions and climate or weather patterns. This is also true for tropical infectious diseases. Hence, only few trend analyses exist, e. g. for malaria reaching higher elevations in Africa.

The main vectors for infectious diseases are mosquitoes, biting flies, bugs, lice, fleas, and mites. Of about more than a million insect species roughly 17,000 have adapted to a blood-sucking mode of life. A small minority of these are vectors of pathogens. The pathogens are viruses, bacteria, protozoa or filarioses. Pathogens are either multiplied within the vectors without change of form (Arboviruses, Rickettsiae, Bacteria), multiplied with change of form (Protozoa) or changed without multiplication (Filaria). From uptake of a pathogen during a blood meal until the infective stage temperature is *the* factor determining duration, hence, higher temperatures can lead to enhanced spread of vector-borne diseases as for example observed for dengue fever.

According to number of people infected by blood-sucking insects *malaria* comes first with 300 to 500 million cases per year of which 1 to 2 million die, especially children. About 70 mosquito species of the genera *Anopheles* transmit four different protozoa (*Plasmodium falciparum*, *Plasmodium vivax*, *Plasmodium ovale*, *Plasmodium malariae*). The mosquito *Anopheles gambiae* in Africa is the most important vector. Because malaria has been eradicated in many developed countries in Europe an enhanced potential for malaria must not lead to a re-introduction there. However, in countries with a weak public health system higher prevalence and new malaria infected areas are highly probable.

About 120 million people in Asia, Africa, South and Central America are infected by *lymphatic filariasis* (elephantiasis) transmitted by different genera of mosquitoes. The larvae of the worm develop (from 0.4 to about 1.5 mm) in the vector and can be transmitted during the next blood meal to humans. The development period is temperature-dependent. Hence, a potential for a further spread exists, but the Global Program to Eliminate Lymphatic Filariasis (WHO, 2006) may reduce incidence strongly.

*Dengue fever* affects about 50 million persons per year and it is a clearly growing threat to human health in about

100 tropical and subtropical countries. The virus belonging to the flaviviruses is transmitted mainly by the *Aedes aegypti* mosquito. Vector and virus show strong temperature dependence in their development. But the disappearance of *Aedes aegypti* from Europe around 1950 points to the probably strong application of the insecticide DDT. As often, measures taken against infectious diseases can easily off-set climate influences.

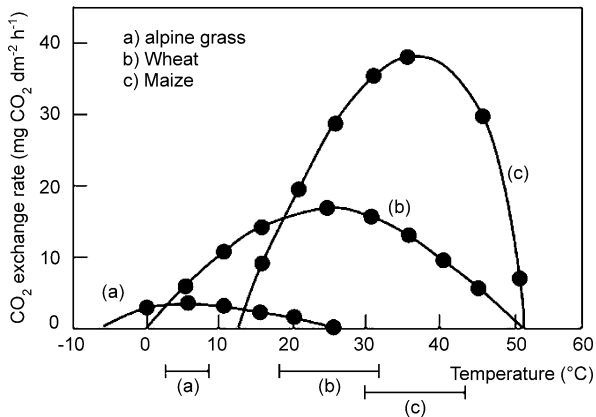
*Onchocerciasis* (in severe cases leading to river blindness) affects about 37 million persons, mainly in Africa, the microfilaria of the worm "*Onchocerca volvulus*" are taken up from the human skin by the vector, a simuliidae species, during a first blood meal, they develop through two skinings within the vector and are transmitted back to a human being during a further blood meal after about 7 to 10 days. Whether the vector, developing in running water and sucking blood during daylight, is already reacting to climate change is not known although a temperature and precipitation dependence clearly exists.

Many other tropical vector-borne infectious diseases exist, like loiasis affecting about 13 million people in tropical Africa, "Schlafkrankheit" with about 60,000 new cases per year or West Nile fever that has reached the USA. In all cases it is a complex mix of influencing factors with positive and negative feedbacks, which inhibits a clear separation of a climate contribution to changed patterns and severity. Hence, there will be many surprises often based both on transport of vectors by growing trade and tourism and better survival conditions in higher latitudes due to higher temperatures and/or changed moisture conditions.

The best means to cope with changed disease distribution patterns is a strong public health system and links between the systems in different countries.

### Climate Change Impacts on Plants with Consequences for Human Health

Our food is produced by plants on land and in the ocean, even if we eat meat or fish because animals also feed ultimately on plants. Hence, a climate change impact on plants may have strong indirect effects on human health. However, even if the climate change impact were small on certain plants, their way of reacting to enhanced CO<sub>2</sub> concentration could still have consequences for food production. In reality climate change and elevated CO<sub>2</sub> concentration act together and both have largely different impacts on plant types and species. The temperature dependence of photosynthesis rate as presented in Fig. 4 for very different plants, shows the rather steep decline of this rate for maize, a C4-plant, at high temperatures above about 40°C. Adding to this finding that there is nearly no pos-



Climate Change and Human Health, Figure 4  
**Temperature dependence of photosynthesis rate for an alpine grass, wheat and maize. Please note the strongly differing optimal temperature ranges. From [10]**

itive feedback to higher CO<sub>2</sub> levels for C<sub>4</sub>-plants, yield in tropical areas would be reduced if plants have to assimilate at leaf temperatures of about 40°C. Therefore, Working Group II of IPCC concluded [4]: “Crop productivity is projected to increase slightly at mid- to high latitudes for local mean temperature increases of up to 1–3°C depending on the crop, and then decrease beyond that in some regions. At lower latitudes, especially seasonally dry and tropical latitudes, crop productivity is projected to decrease for even small local temperature increases (1–2°C), which would increase the risk of hunger.”

If C<sub>3</sub>-plants (e. g. wheat, rice, potato, sugar beet) live at higher CO<sub>2</sub>-concentrations their photosynthesis rate increases rather linearly with CO<sub>2</sub>-concentration, if water stress and nutrient scarcity are not limiting their photosynthesis rate. Consequently, the delay of decades between CO<sub>2</sub>-concentration rise and full expression of the warming, to which we are already committed, is a “window of opportunity” for high crop yields of C<sub>3</sub>-plants. This may in the long-term also have global consequences, because the CO<sub>2</sub>-concentration could rise additionally if the (high latitude) forests (C<sub>3</sub>-plants), acting as a sink for anthropogenic CO<sub>2</sub> presently, would lose this capacity under higher climate change stress. Whether and when this will occur is not yet known.

### Changes in Food Composition for Main Crops at Elevated CO<sub>2</sub>-Concentration

An important consequence of elevated CO<sub>2</sub>-concentration would be changed composition of plant tissue, and especially of seeds, as it could have immediate health conse-

quences for animals and humans eating them. The sparse body of published studies is nearly unanimously pointing to a loss of nitrogen content in leaves and stems as well as in seeds (see Fig. 5 and [6] for an overview). The results of so-called Free Air Carbon Dioxide Enrichment (FACE) studies at roughly doubled CO<sub>2</sub>-concentrations (~ 550 ppm) are all reporting higher nitrogen content for plant tissue and a bit less for seeds. This negative consequence for our food is not yet fully acknowledged in the public, because of the very different impacts for different parts of society. For a grain-producing farmer the CO<sub>2</sub> fertilization effect leads to higher yields of C<sub>3</sub>-plants, a dampening of photo-smog yield reductions, higher water use efficiency for both C<sub>3</sub>- and C<sub>4</sub>-plants and thus less drought impact while for the baker the wheat quality for baking bread declines, and the cattle as well as the consumer get less healthy food.

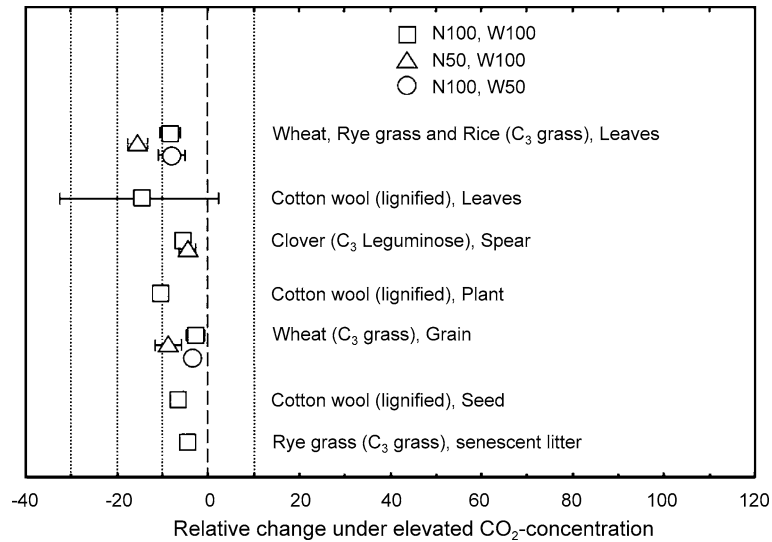
### Shift of Biomes and Migration

As already observed, precipitation is redistributed due to rather different warming patterns. And projections of precipitation changes in the 21st century, as published in the Fourth Assessment Report of [4], can be summarized in the following sentence: Rather humid areas will on average get more precipitation, especially in high latitudes, while semi-arid areas will get on average less (with strongly lowered soil moisture at higher temperatures). This is bad news for many millions of people in semi-arid zones of the tropics and subtropics. The main impact on plants is water stress, hence lower crop yields. But also less food for cattle in pasture lands, continued or enhanced malnutrition for the poor, aggravation of existing deficiencies in public health care systems in these areas will be the negative consequences of climate change impact on the biomes of these areas. In other words: Undermined public health systems, loss of livelihoods and finally migration will be the dire consequences of anthropogenic climate change. The socio-economic and political reactions to this threat cannot be foreseen in detail. However, the international economic cooperation between developing and industrialized countries has to take as a priority adaptation to unavoidable climate change as the best means to lower impacts on public health.

### Concluding Remarks

Although threat to human health was often the cause for environmental policy making, e. g. in the case of desulfurization of power plant exhaust, the manifold threats to human health as a consequence of global anthropogenic climate change have rarely been named as a key reason for





Climate Change and Human Health, Figure 5

Changes in nitrogen content in plant tissue under elevated CO<sub>2</sub>-concentration (~ 550 ppm) depending on nitrogen fertilizer and water availability (50 or 100%). From [16]

climate change policy measures. What else is the partly still growing mal-nutrition of subsistence farming communities in the desertification-prone semi-arid tropics and subtropics than a threat to human health? On the other hand most new threats to our health caused by the spread of vector-borne infectious diseases due to higher temperatures or those caused by new weather extremes can be strongly reduced by proper health system up-grading and pre-cautionary measures that strengthen security-related infrastructure. However, this will probably not be the case in developing countries already suffering from (very) weak public health systems; unless the preliminary decisions of the 13th Conference of the Parties to the UNFCCC lead to a new international and binding protocol in 2009 as a follow-on to the Kyoto Protocol that then stipulates that a fixed portion of the revenues of international greenhouse gas emission trading should be used for adaptation measures in developing countries. A large share has to be invested in public health systems in poor developing countries in order to help the poorer parts of societies typically suffering most from epidemics and weather-related catastrophes.

If earlier tropical vector-borne infectious diseases, like the West Nile fever, reach developed countries, research to get proper vaccines will be stimulated within large pharmaceutical companies. If the threat remains confined to the poor South, this research effort will often not exist, because developing countries' normal population cannot afford the expensive new drugs or vaccines that remain

property right protected for years. It is high time that political summits deal with this problem and WHO helps to circumvent this barrier as pointed out recently by the Nobel Prize laureate for economics (Stiglitz, 2006).

## Bibliography

1. Bohren C, Memillod G, Delabays N (2006) Common ragweed in Switzerland: development of a nation-wide concerted action. *J Plant Dis Prot, Special Issue XX, Ulmer, Stuttgart*, pp 497–503
2. Confaloni U, Menne B, Akhtar R, Ebi KL, Hanengue M, Kovats RS, Revich B, Woodward A (2007) Human health. In: Parry MO, Lanziani OF, Palutikof JP, van der Linden PJ, Hanson CE (eds) *Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of IPCC*. Cambridge University Press, Cambridge, pp 391–431
3. D'Amato G, Cecci L, Bonini S, Nunes C, Annesi-Maesano I, Behrendt H, Liccardi G, Popov T, van Canwenberge P (2007) Allergenic pollen and pollen allergy in Europe. *Allergy* 62:976–990
4. Intergovernmental Panel on Climate Change (IPCC) (2007) *Science of Climate Change, Contribution of Working Group I to the Fourth Assessment Report*. Cambridge University Press, Cambridge
5. Jendritzky G, Koppe C (2008) Die Auswirkungen von thermischen Belastungen auf die Mortalität. In: Lozán JL, Grassl H, Jendritzky G, Karbe L, Reise K (eds) *Warnsignale Klima: Gesundheitsrisiken. Wissenschaftliche Auswertungen*, Hamburg, ISBN 978-39809668-4-9
6. Kimball BA (2004): Global environmental change: implications for agricultural productivity. *Crop Env Bioinform* 1:251–263
7. Laschewski G, Jendritzky G (2002) Effects of the thermal environment on human health: an investigation on 30 years

- of daily mortality data from SW Germany. *Clim Res* 21:91–103
8. McMichael A, Campbell-Lendrum DH, Corvalán CF, Ebi KL, Githeko AK, Schraga ID, Woodward A (2003) *Climate Change and Human Health: Risks and Responses*. WHO, Geneva
  9. Menzel A et al. (2006) European phenological response to climate change matches the warming pattern. *Glob Chang Biol* 12:1969–1976
  10. Rosenzweig C, Hillel D (1998) Carbon Dioxide, Climate Change and Crop Yields. In: Rosenzweig D, Hillel D (eds) *Climate Change and the Global Harvest. Potential Impacts of the Greenhouse Effect on Agriculture*. Oxford University Press, Oxford, pp 70–100
  11. Rosenzweig C, Cassassa G, Karoly DJ, Imeson A, Liu C, Menzel A, Rawlins S, Toot TL, Seguin B, Tryjanowski P (2007) Assessment of observed changes and responses in natural and managed systems. In: Parry MO, Lanziani OF, Palutikof JP, van der Linden PJ, Hanson CE (eds) *Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of IPCC*. Cambridge University Press, Cambridge, pp 79–131
  12. Sauerborn R (2007) Climate change: an agenda for research and teaching in public health. *Scand J Public Health* 1–3
  13. Schär C, Vidale PL, Lüthi D, Frei C, Häberli C, Liniger MA, Appenzeller C (2004) The role of increasing temperature variability for European summer heat waves. *Nature* 427:332–336
  14. UBA (2007) [www.env-it.de/luftdaten/download/public/docs/pollutant/03/Jahr/Ozberi06.pdf](http://www.env-it.de/luftdaten/download/public/docs/pollutant/03/Jahr/Ozberi06.pdf)
  15. UBA (2007) [www.env-it.de/luftdaten/download/public/docs/pollutant/PM10\\_gesamt\\_2001-2006.pdf](http://www.env-it.de/luftdaten/download/public/docs/pollutant/PM10_gesamt_2001-2006.pdf)
  16. Weigel H-J, Manderscheid R, Fangmeier A, Högy P (2008) Mehr Kohlendioxid in der Atmosphäre: Fluch oder Segen für die Landwirtschaft. In: Lozán JL, Grassl H, Jendritzky G, Karbe L und Reise K (eds) *Warnsignale Klima: Gesundheitsrisiken. Wissenschaftliche Auswertungen*, Hamburg, ISBN 978-39809668-4-9

## Climate Change, Economic Costs of

RICHARD S. J. TOL<sup>1,2,3,4</sup>

<sup>1</sup> Economic and Social Research Institute, Dublin, Ireland

<sup>2</sup> Institute for Environmental Studies, Vrije Universiteit, Amsterdam, The Netherlands

<sup>3</sup> Department of Spatial Economics, Vrije Universiteit, Amsterdam, The Netherlands

<sup>4</sup> Department of Engineering and Public Policy, Carnegie Mellon University, Pittsburgh, USA

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Issues](#)

[Total Costs](#)

[Marginal Costs](#)

[Policy Implications](#)

[Future Directions](#)

[Acknowledgments](#)

[Bibliography](#)

### Glossary

**Direct costs** The direct cost equals quantity times price.

**Discount factor** The discount factor  $t$  years into the future equals one over one plus the discount rate, raised to the power  $t$ .

**Discount rate** The discount rate is the annual rate of decline of the value of consumption. It is roughly equal to the rate of interest, or the opportunity cost of capital. The discount rate consists of two components: the rate at which consumers get better off, and the rate of pure time preference (or impatience). The discount rate is not related to the rate of inflation, which is the annual rate of decline of the value of money.

**Economics** Economics is the social science which studies human behavior with regard to the relationship between ends and scarce means which have alternative uses.

**Equity weights** Equity weights are applied to aggregate national impacts to a global total. Equity weights are often set to unity, but sometimes equity weights equal the ratio of nationally average per capita income over global average per capita income.

**Indirect costs** The indirect costs equal all costs that are not direct costs. This includes the price change induced by the change in quantity (partial equilibrium effects), the changes in other markets (general equilib-

rium effects), and the changes at later times (dynamic effects).

**Marginal costs** The marginal cost of greenhouse gas emissions equals the first partial derivative of the net present value of the total costs of climate change to emissions.

**Monetary valuation** Monetary valuation is a set of techniques and their application that attempts to express in monetary terms the value to humans of changes in environmental goods and services. Negative impacts are typically expressed as an income loss that would give an equivalent loss in welfare.

**Neo-classical economics** Although historians would refer to neo-classical economics as the dominant form of economic research between 1860 and 1910, common usage has neo-classical economics as a synonym for mainstream or orthodox economics. In that sense, neo-classical economics is a style of research, characterized by empirical rigour, mathematical rigour, and micro-founded macro-relationships.

**Net present value** The net present value is the sum of all future costs and benefits, weighted by the discount factor.

**Total costs** The total cost of climate change equals the direct and the indirect costs of climate change, that is, the difference in welfare between a scenario with climate and a scenario without.

### Definition of the Subject

The economic costs of climate change include all positive and negative impacts of the enhanced greenhouse effect and the resulting changes in the atmosphere and ocean on all human consumers and producers. Total costs refer to the difference in human welfare between a scenario with climate change and a scenario without climate change. Marginal costs refer to the difference in human welfare between two scenarios with a slightly different climate, normalized by the amount of greenhouse gas emissions that would induce that difference. Estimates of the economic costs of climate change are important to assess the size of the climate problem relative to other problems, and to compare the costs of climate change to the costs of greenhouse gas emission reduction.

### Introduction

Calls for greenhouse gas emission reduction are often phrased as a moral imperative. While tempting, this is wrong. Firstly, there is no moral agreement. Emission reduction could save polar bears but it would cost coal miner's jobs and raise the price of food for the malnour-

ished. Moral imperatives are easy if a policy has only benefits. As soon as a policy has both costs and benefits, one has to make trade-offs and choose the lesser evil. Secondly, there is no avoiding dangerous interference with the climate system. Emission reduction would slow down the melting of the Greenland ice cap, and reduce the probability of a collapse of the West-Antarctic Ice Sheet – but it would not stop the melting or bring the change to zero. Thirdly, we have no obligations to future generations or poor people. Such duties are self-imposed. And even if we choose to help others, there are many ways to do this. Would our grandchildren prefer a richer but warmer world, or a poorer but colder one? Would the grandchildren of the Bangladeshis like us to reduce greenhouse gas emissions, help them to adapt to climate change, or help their grandparents grow rich?

This chapter looks into such questions. It is written from the thoroughly relativistic perspective of a neo-classical economist. The basic principles of the economic theory of climate change are quite simple. Greenhouse gas emissions are an externality. Externalities are unintended – we burn coal to make electricity, not to emit carbon dioxide – and uncompensated – carbon dioxide is freely dumped in the atmosphere – consequences of economic activity. Externalities should be internalized, that is, emitters should pay for their emissions. The price of emissions should equal the damage done by the emissions. That is all.

There is now a vast literature on the economics of climate change – started by Nordhaus [73,74]. A large part of that literature is about the deviation between the simple policy prescription of economic theory and the complexities of actual policy. Another large part of the literature is about the costs of greenhouse gas emission reduction. The literature on the economic costs of climate change is only a small one. It is reviewed here.

Section “**Issues**” discusses the methodological, conceptual, and moral issues one has to confront when estimating the economic impacts of climate change. Section “**Total Costs**” reviews estimates of the total economic impact. Section “**Marginal Costs**” surveys estimates of the marginal impacts. Section “**Policy Implications**” concludes by assessing the policy implications.

## Issues

### Scenarios

A scenario is a set of assumptions on future conditions that is coherent, internally consistent, and not implausible [70]. Climate scenarios are usually derived from modeling experiments with General Circulation Models (GCM). Climate scenarios include simple statistics such as the global

mean surface air temperature, and complex results such as spatial patterns of rainfall extremes. Climate scenarios may also include low-probability events, such as a disruption of the thermohaline circulation in the Atlantic Ocean, or the collapse of the West Antarctic Ice Sheet.

Scenarios also include population, economic activity, greenhouse gas emissions, and land use. Besides driving the climate models, these components are also important as they determine the vulnerability of social and economic systems to climate change over time. Although poorer societies are generally believed to be more vulnerable to climate change, this is by no means a simple relationship. Some impacts tend to fall with economic growth. The impacts of climate change on infectious diseases are a prime example. Malaria does not kill middle income people, because they can afford prevention and (if necessary) cure. Some impacts tend to rise with economic growth. The impact of climate change on biodiversity and species loss is one example. People tend to care more about these matters as their income grows, and further developed economies put more pressure on nature. Other impacts may rise first and then fall with economic growth. Urban air quality is one example. The very poor have nothing to foul the air with, and the very rich do not like foul air and have the wherewithal to prevent it. Climate change is likely to increase malaria, reduce biodiversity and worsen air quality. In some case, climate change and economic growth work together to increase the impacts, while in other cases they pull in opposite directions.

## Valuation Approach

There are various techniques for the monetary valuation of climate change impacts. Some values of impacts are directly based on observed prices. Agriculture and dike building are examples. Other values can be indirectly measured on the basis of observed market prices for surrogate products or services. One example is human health, which is not traded directly but indirectly through safety measures, labor markets, and health care. The challenge in these instances is to model future market prices that are consistent with the underlying socioeconomic scenario. For yet other impacts, no market values exist, and hypothetical prices are needed. Notable impacts are on non-commercial ecosystems and biodiversity.

Because it is practically impossible to estimate each exposure-response relationship or value at the respective geographical location of a climate change impact, data from previous studies focusing on different locations and different policy contexts are inevitable. Furthermore, most climate change impacts will take place in the future, for

which by definition no data are available. Therefore it is important to know when data from other studies can be used and under what conditions, and how to extrapolate values from today to tomorrow.

The majority of recent studies still adopt benefit transfer methods for the evaluation of climate impacts. However, benefit transfer is not very reliable [91]. For this reason, more attention should be given to original valuation research in the context of climate change.

An example of such a study is Li et al. [54] who analyze the willingness-to-pay (WTP) of American citizens for climate policy by means of the contingent valuation method. They find that the median American citizen is willing to pay about \$ 15/tC. Berrens et al. [7] find a willingness-to-pay between \$ 200 and \$ 1760 per US household per year (0.2–2.3% of income) for US ratification of the Kyoto Protocol. (Manne and Richels [60] estimate that the costs of US ratification would be 0.75% of GDP in 2010.) Hersch and Viscusi [46] find that Europeans are willing to pay up to 3.7% more for petrol if that helps combat climate change. Viscusi and Zeckhauser [117] find that Harvard students are willing to pay \$ 0.50/gallon (a 25% price increase) or 3% of their expected annual income for greenhouse gas emission reduction. (This study also showed that these students underestimate projected warming in Boston by about 50%, while the authors made them believe that carbon dioxide emission reduction would be effective for slowing climate change in the next 30 years.) There is scope for similar applications of WTP techniques, mainly to account for spatial and socio-economic differences in individuals' preferences.

### Direct and Higher Order Impacts

Most studies to date have estimated the direct costs of climate change. Direct costs equal the physical change (e. g., the dikes to be reinforced) times their price (in this example, the costs per dike length and dike height). Direct costs are easy to compute (but see Subsect. “**Valuation Approach**”), but probably underestimate the real economic costs.

The higher order impacts come in three kinds. Firstly, climate change may impact the market under consideration. For example, if dikes are being reinforced everywhere, then the costs of dike building is likely to go up as materials, machinery, and skilled labour is difficult to get. Secondly, the impact of climate change on one market may spill over into other markets. For example, dike building may increase the costs of construction, as the same materials and skills are used. Dike building is capital intensive and may drive up the interest rate. Thirdly, the impact of

climate change may affect economic growth. For example, money invested in dike building is not invested elsewhere.

A number of recent studies have examined the economy-wide implications of sea level rise [10], tourism [8], and health [9]. While it is perhaps too early to draw firm conclusions from this body of research, the studies suggest that the indirect effects of climate change impacts can both enlarge and diminish the direct economic impacts of climate change. The distribution of gains and losses is another difference between direct costs and general equilibrium effects. Whereas direct costs are limited to those directly affected, markets would spread the impact to their suppliers, clients, and competitors as well as to financial markets.

Fankhauser and Tol [35] show that the economic growth impact of climate change is as large as the direct impact of climate change. Acemoglu et al. [1] and Masters and McMillan [62] show that differences in climate explain part of observed differences in economic development. Easterly and Levine [28] show that the link is at most weak and indirect, and it is not clear whether the mechanisms that may have been active in the past, still hold for present and future.

### Adaptation

One cannot study the costs of climate change impact without also studying, or at least making assumptions about the costs of adaptation [109]. Studies focusing on costs of the impacts make widely differing assumptions about the amount of adaptation that will take place. While some studies completely ignore adaptation, other studies consider arbitrary levels of adaptation, or assume optimal adaptation. No studies use realistic models of adaptation [109]. There is little research that shows how adaptation costs compare to the potential damages of not adapting. The impacts of climate change and the capacity to adapt would be affected with the level of development and flexibility of the economy [123]. Hence, the future success and nature of adaptation depends on the assumed socio-economic scenario.

### Aggregation: Temporal

Climate change is a slow process. Today's emissions will affect the climate for decades to centuries, and sea level for centuries to millennia. As cause and effect are separated in time, so are costs (of emission reduction) and benefits (of avoided climate change). The procedure to make commensurate costs and benefits at different points in time is called discounting. Discounting is as common as it is controversial. See [3] for an excellent discussion.

Individuals discount future gains or losses because of two reasons. (People may also discount the future because it is more uncertain than the present, but in this case discounting is used as a shortcut for an uncertainty analysis.) First, money earns interest. Second, people are impatient. The first reason is widely accepted. Davidson [26] is one of the few exceptions. On the second reason, there is virtual consensus too. All ethical arguments show that people should not discount (e. g., [12]). All empirical evidence shows that people do nonetheless (e. g., [79,80]).

Climate change is a large-scale problem. Therefore, the discount rate of society is more relevant than the individual discount rate. The appropriate measure of the growth rate of money is the average growth rate of per capita consumption. Again, there is little dispute on this. But should the social rate of discount also include a measure of impatience? Again, philosophers agree: Impatience is immoral. However, this implies that a government would deviate from the will of the people. This may be defended with the argument that the government is the guardian of future, yet unborn people. However, the empirical evidence is clear in this case too: Governments are impatient [31].

Discounting is more profound over long periods than over short ones. Discounting implies that climate change damages that occur in a century or so are largely irrelevant. This realization has led people to rethink the fundamental principles of discounting, particularly

- (a) the notion that the procedure of discounting results from the intertemporal allocation of resources of an individual agent; and
- (b) the assumption that discounting is exponential.

To start with the individual perspective, Lind [56] and Lind and Schuler [55] argue that *earmarked* investment is a crucial assumption in discounting. The discount factor measures the trade-off between consumption now and consumption later, where consumption later is contingent on a specific investment plan. As the current generation cannot commit near-future generations to maintain their investments for the benefit of far-future generations, discounting breaks down between generations. Schelling [95] agrees. The alternative is to decide explicitly on the resource allocation between generations. Chichilnisky [19] shows that discounting coincides with a dictatorship of the present generation over future generations. Gerlagh and Keyzer [38] show that discounting is equivalent to the present generation owning all future resources. This is objectionable from a moral standpoint, but it is reality. This line of research has not led to practical alternatives to discounting.

Conventional discounting is exponential: The discount factor is  $(1 + r)^{-t}$ , where  $r$  is the discount rate and  $t$  is time. Some people argue that the functional specification of conventional discounting is wrong. The first component is empirical. Conventional exponential discounting has that the relative difference between two years is always equal, regardless of their distance from the present. That is, the difference between year 10 and 11 is the same as the distance between year 100 and 101. However, many people would in fact argue that the difference between year 10 and 11 is equal to the difference between year 100 and 110. Such hyperbolic discounting [22] is very similar to exponential discounting for short periods, but the difference is substantial for long periods. The similarity between exponential and hyperbolic discounting in the short run is important, because a switch to hyperbolic discounting would imply a drastic overhaul of long-term decisions only.

There are two further arguments for hyperbolic discounting cf. Dasgupta and Maskin [25]. The first is due to Weitzman [118]. He shows that, if one is uncertain what discount rate to use, then the lowest discount rate becomes increasingly dominant over time. The certainty-equivalent discount rate falls with time, and the difference between years shrinks in the more distant future. Consider the following example. After one year, the average of a 1% and a 10% discount rate is

$$1 - \left( \frac{(1.01^{-1} + 1.10^{-1})}{2} \right)^{1/1} = 5.0 \%$$

(and not 5.5%). After 100 years,

$$1 - \left( \frac{(1.01^{-100} + 1.10^{-100})}{2} \right)^{1/100} = 1.7 \%$$

That is, the average approaches the minimum as time progresses. One may criticize this as a short cut for a full uncertainty analysis. However, Gollier [39,40] shows that the same is true if a government somehow aggregates the individual discount rates of its citizens. In the long run, the preferences of the person with the lowest discount rate become increasingly important, and the discount rate declines over time.

Guo et al. [41] and Newell and Pizer [71] show that hyperbolic discounting leads to higher estimates of the social cost of carbon. However, the quantitative effect is limited by the fact that hyperbolic discount rates are high for the first decades.

### Aggregation: Spatial

Climate change is a global problem. Carbon dioxide and other greenhouse gases mix uniformly in the atmosphere.

This implies most of the impacts of one country's emissions fall on other countries. The same is true for the benefits of emission reduction. The impacts on different countries need to be aggregated somehow.

Two methods dominate the literature. In the first and oldest method, regional impacts are quantified in local currencies, converted to dollars, say, and added up [33,103]. This is simple, but the disadvantage is that similar impacts are treated differently. Most disturbingly, climate-change-induced deaths in rich countries receive a greater weight than climate-change-induced deaths in poor countries. The second method, known as equity weighing, corrects for this [6,36]. Rather than simply adding regional estimates, the regional utility-equivalents are added and then converted back to money according to an assumed global welfare function. A big disadvantage of this method is that climate-change-induced deaths are treated differently than deaths by other, national causes. The reason for this discrepancy is that equity weighing, as practiced in the literature, explicitly assumes a global decision maker.

In the meta-analysis of Tol [107], the median estimate of the marginal damage costs of carbon dioxide is \$ 10/tC without equity weights, and \$ 54/tC with equity weights. So, equity weighing is obviously important. The reason is simple. Poor countries are more vulnerable to climate change. Poor countries have little economic weight. Equity weights correct for this.

Morally, this may be the right thing to do. However, national governments also have a certain obligation to defend the interests of their citizens. A narrow interpretation of self-interest would suggest that impacts abroad be ignored (unless they spill over, e. g., through international migration). Then, climate change policy would be very limited, as most impacts will be abroad. However, the principle of good neighborhood is well established, both morally and legally. This entails that one should avoid doing harm to others; and should pay compensation if harm is done nonetheless (e. g., [113]).

A rational actor would avoid doing harm if that is cheaper than the compensation paid. From a national perspective, the relevant damages are then the impacts on the own country plus the compensation paid to other countries. Schelling [93] forcefully argues that compensation should equal the welfare loss of the victim rather than the welfare loss that the culprit would have experienced had she been the victim. This argues for aggregation of monetized impact estimates without equity weighing.

However, compensation would need to be paid only once. Furthermore, a country would also reasonably expect to be compensated itself. This implies that the dam-

age to a country equals the global damage times its share in causing the problem. Defining the latter is a thorny issue, as the cause-effect chain is long, complex, and uncertain. One would need to make arbitrary decisions on cause, effect and their connection.

### Uncertainty

Climate change is plagued by uncertainty [16]. Partly, this is because our understanding of climate change and its impacts is incomplete. For the larger part, however, this is because climate change will take place in the future, partly driven by future emissions, and impacting a future world. Future research and observations may reduce the uncertainty, although surprises may increase the uncertainty just as well, but uncertainty will never disappear. Learning and irreversibility play a crucial role in how to deal with uncertainty. Events that may or may not occur in some distant future, but whose consequences can be alleviated once it becomes clear if they would occur, should not worry us too much. On the other hand, if an effect is irreversible (e. g., species extinction), we may want to prevent it regardless of how uncertain it is and regardless of what future research will show (according to the "precautionary principle"). Another crucial part of dealing with uncertainty is risk aversion. Essentially, this determines how much weight we place on negative surprises. A risk neutral decision maker would cancel negative surprises against positive ones, but a risk adverse decision maker would not. Recent work has shown that the marginal damage costs of carbon dioxide are indeed very sensitive to the assumed degree of risk aversion. Although uncertainty and risk are often emphasized – often in a casual way – only few studies seek to quantify its implications (e. g., [51]).

In a recent paper, Weitzman [120] shows that, under a wide range of standard assumptions, the uncertainty about climate change is so large that the expected value of the social costs of climate change is infinite. Earlier, Tol [106] showed this for a specific model. This implies that uncertainty should take central stage in the analysis of climate policy. The Weitzman result throws up a number of methodological issues that will need to be resolved before the policy implications of this work become clear.

### Completeness

The impacts of climate change that have been quantified and monetized include the impacts on agriculture and forestry, water resources, coastal zones, energy consumption, air quality, and human health. Obviously, this list is incomplete. Also within each impact category, the as-

assessment is incomplete. Studies of the impacts of sea level rise on coastal zones, for instance, typically omit saltwater intrusion in groundwater [72]. Furthermore, studies typically compare the situations before and after climate change, but ignore that there will be a substantial period during which adaptation is suboptimal – the costs of this are not known.

Some of the missing impacts are most likely negative. Diarrhoea impacts have been quantified recently [57]. Like malaria, diarrhoea is a disease that is driven by poverty but sensitive to climate. Including diarrhoea tightens the link between development and climate policy. Increasing water temperatures would increase the costs of cooling power plants [101]. Redesigning urban water management systems, be it for more or less water, would be costly [4], as would implementing the safeguards against the increased uncertainty about future circumstances. Roads and bridges would suffer from weather conditions for which they were not designed; this would imply either disruption of traffic or expensive retrofits. Extratropical storms may well increase, leading to greater damage and higher building standards [27]. Expenditures on these things are relatively small. Even if climate change would double or triple the cost, the impact would be small. Ocean acidification would reduce marine biodiversity, and may well harm fisheries [52]. Ocean fisheries are only a small, and declining fraction of GDP, while there are ready substitutes for wild fish protein (notably fish farming). The value of biodiversity is unclear (see below).

Other missing impacts are probably positive. Higher wind speeds in the mid-latitudes would decrease the costs of wind and wave energy [11,44]. Less sea ice would improve the accessibility of arctic harbours, would reduce the costs of exploitation of oil and minerals in the Arctic, and may even open up new transport routes between Europe and East Asia [121]. Warmer weather would reduce expenditures on clothing and food, and traffic disruptions due to snow and ice [15]. Also in these cases, the impact of climate change is likely to be small relative to the economy.

Some missing impacts are positive in some places, and negative in others. Tourism is an example. Climate change may well drive summer tourists towards the poles and up the mountains [42,43]. People, however, are unlikely to change the time and money spent on holiday making. The effect is a redistribution of tourist revenue [8]. The global impact is close to zero, but regional impacts are measured in tens of billions of dollars – positive in temperate, rich countries, and negative in tropical, poor countries. This exacerbates the already skewed distribu-

tion of climate impacts. Some ski resorts may go out of business, and others would need expensive snowmaking equipment [29,97]. Other ski resorts would profit from the reduced competition. Although regional impacts may be substantial, at the global scale positives and negatives cancel.

Other impacts are simply not known. Some rivers may see an increase in flooding, and others a decrease [53]. At the moment, only a limited number of rivers have been studied in detail, and it is unclear how to extrapolate to other rivers. It is clear though, that land use and water management may greatly increase or reduce impacts. Although river floods wreak substantial havoc and damages of a single event can reach substantial numbers, average flood damage is in fact small relative to the economy [112]. Tropical storms do more damage, although a substantial share of the impact is due to bad planning rather than bad weather [14]. Nonetheless, tropical storms may prevent capital accumulation and the plantation of lucrative crops such as banana [30,69]. Unfortunately, it is not known how climate change would alter the frequency, intensity, and spread of tropical storms [63,89].

The missing impacts discussed above are probably small. There are also bigger gaps in the coverage of climate change impact studies. Climate change is likely to have a profound impact on biodiversity, but quantitative predictions are rare [13]. Although the economic impact of a small change in biodiversity is known to be small [88], the value of large biodiversity changes is unknown but could well be substantial [18]. There is a small but unknown chance that climate change will be more dramatic than is typically assumed in the impacts literature. This may be because of shutdown of the thermohaline circulation [61], a collapse of the Greenland or West-Antarctic Ice Sheet [84], or a release of large amounts of methane [45]. The economic analysis of such scenarios has only just begun [57]. It may be that climate change would lead to large-scale migration [64] and violent conflict, although there is only weak empirical support for this [49,124]. Finally, climate change impact studies stop at the end of the 21st century. In 2100, impacts are negative, and getting more negative at an accelerating pace. It is not known how rapidly things would get worse in the 22nd century without emission abatement.

Although the sign of the aggregate unknown impacts is not known, risk aversion would lead one to conclude that greenhouse gas emission reduction should be more stringent than suggested by a cost-benefit analysis based on the quantified impacts only. However, the size of the bias is unknown too – so the main policy implication is that more research is needed.



## Total Costs

The first studies of the welfare impacts of climate change were done for the USA [21,74,98,102]. Although Nordhaus [74] (see also Ayres and Walter [5]) extrapolated his US estimate to the world, the credit for the first serious study of the global welfare impacts goes to Fankhauser [32,33], although Hohmeyer and Gaertner [48] earlier published some low quality estimates. Other global estimates include those by Nordhaus [76,77], Tol [103], Nordhaus and Yang [82], Plambeck and Hope [90], Nordhaus and Boyer [81], Mendelsohn et al. [66,68], Tol [105], Maddison [59], Hope [50], Rehdanz and Maddison [92] and Nordhaus [78]. Note that Stern et al. [100] is based on Hope [50].

This is a rather short list of studies, and an even shorter list of authors. This problem is worse if one considers that Nordhaus and Mendelsohn are colleagues; that Fankhauser, Maddison and Tol are students of Pearce; and that Rehdanz is a student of Maddison and Tol; while Hope's (and Stern's) estimates are averages of Fankhauser's and Tol's. Although most fields are dominated by a few people, dominance is here for want of challengers. The effect of this is hard to gauge. The reasons are lack of funding (this work is too applied for academic sources, while applied agencies do not like the typical results and pre-empt this by not funding it), lack of daring (this research requires making many assumptions, and taking on well-entrenched incumbents), and lack of reward (the economics profession frowns on the required interdisciplinarity). In addition, many people, including many economists, would argue that climate change is beyond cost-benefit analysis and that monetary valuation is unethical.

Table 1 shows some characteristics of these studies. A few insights emerge. First, the welfare impact of a doubling of the atmospheric concentration of carbon dioxide on the current economy is relatively small. Although the estimates differ, impacts are not more than a few percent of GDP. The estimates of Hope [50], Mendelsohn et al. [66,68] and Tol [105] even point to initial benefits of climate change. (Studies published after 1995 all have regions with net gains and net losses due to global warming, whereas earlier studies only find net losses.) With such estimates, it is no surprise that cost-benefit analyses of climate change recommend only limited greenhouse gas emission reduction – for instance, Nordhaus [75] argues that the optimal rate of emission reduction is 10–15%, one of the more contentious findings of the climate economics literature.

Second, although the impact is relatively small, it is

not negligible. A few per cent of GDP in annual damage is a real concern.

Third, climate change may initially have positive impacts. This is partly because the higher ambient concentration of carbon dioxide would reduce water stress in plants and may make them grow faster – although this effect is now believed to be weaker [58]. Another reason is that the global economy is concentrated in the temperate zone, where a bit of warming may well be welcomed because of reductions in heating costs and cold-related health problems. At the same time, the world population is concentrated in the tropics, where the impacts of initial climate change are probably negative. Even though initial *economic* impacts are positive, it does not necessarily follow that greenhouse gas emissions should be subsidized. The climate responds rather slowly to changes in emissions, so the initial impacts cannot be avoided. Impacts start falling – that is, additional climate change reduces global welfare – roughly at the same time as climate change can be influenced by present and future emission reduction [47].

The fourth insight is that relative impacts are higher in poorer countries (see also Yohe and Schlesinger [122]). This is because poorer countries have a lower adaptive capacity [2], particularly in health [108], and have a greater exposure to climate change, particularly in agriculture and water resources. Furthermore, poorer countries tend to be hotter and therefore closer to temperature limits and short on spatial analogues should it get warmer still. At the same time, there are fewer studies on the impacts of climate change on developing countries than on developed countries. Although research is scarce [83], there is little reason to assume that climate change impacts would be homogeneous within countries; certainly, certain economic sectors (e.g., agriculture), regions (e.g., the coastal zone) and age groups (e.g., the elderly) are more heavily affected than others. This has two policy implications. Firstly, recall that greenhouse gas mix uniformly in the atmosphere. It does not matter where they are emitted or by whom, the effect on climate change is the same. Therefore, any justification of stringent emission abatement is an appeal to consider the plight of the poor and the impacts imposed on them by the rich [94,95]. While this makes for wonderful rhetoric and fascinating research (e.g., [104]), reality shows little compassion for the poor by the rich. Secondly, if poverty is the root cause for vulnerability to climate change, one may wonder whether stimulating economic growth or emission abatement is the better way to reduce impacts. Indeed, Tol and Yohe [115] argue that the economic growth foregone by stringent abatement more than offsets the avoided impacts of climate change, at least for malaria,

Climate Change, Economic Costs of, Table 1  
Economic impact estimates of climate change; numbers in brackets are either standard deviations or confidence intervals

Study	Warming	Impact	Minimum	Region	Maximum	Region
Nordhaus [76]	3.0	-1.3				
Nordhaus [77]	3.0	-4.8 (-30.0 to 0.0)				
Fankhauser [33]	2.5	-1.4	-4.7	China	-0.7	Eastern Europe and the former Soviet Union
Tol [103]	2.5	-1.9	-8.7	Africa	-0.3	Eastern Europe and the former Soviet Union
Nordhaus and Yang [82] <sup>a</sup>	2.5	-1.7	-2.1	Developing countries	0.9	Former Soviet Union
Plambeck and Hope [60] <sup>a</sup>	2.5	-2.5 (-0.5 to -11.4)	-8.6 (-0.6 to -39.5)	Asia (w/o China)	0.0 (-0.2 to 1.5)	Eastern Europe and the former Soviet Union
Mendelsohn et al. [66] <sup>a,b,c</sup>	2.5	0.0 0.1	-3.6 -0.5	Africa	4.0 1.7	Eastern Europe and the former Soviet Union
Nordhaus and Boyer [81]	2.5	-1.5	-3.9	Africa	0.7	Russia
Tol [105]	1.0	2.3 (1.0)	-4.1 (2.2)	Africa	3.7 (2.2)	Western Europe
Maddison [59] <sup>a,d,e</sup>	2.5	-0.1	-14.6	South America	2.5	Western Europe
Rehdanz and Maddison [92] <sup>a,c</sup>	1.0	-0.4	-23.5	Sub-Saharan Africa	12.9	South Asia
Hope [50] <sup>a</sup>	2.5	0.9 (-0.2 to 2.7)	-2.6 (-0.4 to 10.0)	Asia (w/o China)	0.3 (-2.5 to 0.5)	Eastern Europe and the former Soviet Union
Nordhaus [78]	2.5	-0.9 (0.1)				

<sup>a</sup>Note that the global results were aggregated by the current author.

<sup>b</sup>The top estimate is for the "experimental" model, the bottom estimate for the "cross-sectional" model.

<sup>c</sup>Note that Mendelsohn et al. only include market impacts.

<sup>d</sup>Note that the national results were aggregated to regions by the current author for reasons of comparability.

<sup>e</sup>Note that Maddison only considers market impacts on households.

while Tol [108] shows that development is a cheaper way of reducing climate-change-induced malaria than is emission reduction. Moreover, richer countries may find it easier and cheaper to compensate poorer countries for the climate change damages caused, than to reduce greenhouse gas emissions. Such compensation may be explicit and financial, but would more likely take the shape of technical and financial assistance with adaptation (cf. [85]).

The agreement between the studies is remarkable if one considers the diversity in methods. The studies of Fankhauser, Hope, Nordhaus, and Tol all use the enumerative method: 'physical' impact estimates are obtained one by one, from 'natural science' papers based on 'process-based' models or 'laboratory experiments'. These physical impacts are multiplied with their respective prices, and added up. The 'prices' are obtained by benefit transfer. In contrast, Mendelsohn's work is based on direct, empirical estimates of the welfare impacts, using observed variations in prices and expenditures to discern the effect of climate (e. g., [67]). Mendelsohn estimates are done

per sector and then added up, but physical modelling and benefit transfer are avoided. Nordhaus [78] uses empirical estimates of the *aggregate* climate impact on income, while Maddison [59] looks at patterns of *aggregate* household consumption. Like Mendelsohn, Nordhaus and Maddison rely exclusively on observations, but they assume that all climate effects are aggregated by the economy into incomes and expenditures. Rehdanz and Maddison [92] also empirically estimate the aggregate impact, but use self-reported happiness as an indicator; their approach is similar to that of Nordhaus and Maddison, but the indicator is subjective rather than objective. The enumerative studies of Fankhauser etc rely on controlled experiments (albeit with detailed, process-based models in most cases). This has the advantages of ease of interpretation and physical realism, but the main disadvantage is that certain things are kept constant that would change in reality; adaptation is probably the key element. The statistical studies of Mendelsohn etc rely on uncontrolled experiments. This has the advantage that everything varies as

in reality, but the disadvantages are that the assessment is limited to observed variations (which may be small compared to projected changes, particularly in the case of carbon dioxide concentration) and that effects may be spuriously attributed to climate. Therefore, the variety of methods enhances confidence, not in the individual estimates, but in the average.

The shortcomings of the estimates are at least as interesting. Welfare losses are approximated with direct costs, ignoring general equilibrium and even partial equilibrium effects (see below). In the enumerative studies, impacts are assessed independently of one another, even if there is an obvious overlap as between water resources and agriculture. Estimates are often based on extrapolation from a few detailed case studies, and extrapolation is to climate and levels of development that are very different from the original case study. Valuation is based on benefit transfer, driven only by difference in per capita income. Realistic modelling of adaptation is problematic, and studies either assume no adaptation or perfect adaptation. Many impacts are unquantified, and some of these may be large (see below). The uncertainties are unknown – only 4 of the 14 estimates in Table 1 have some estimate of uncertainty. These problems are gradually solved, but progress is slow. Indeed, the above list of caveats is similar to that in Fankhauser and Tol [34].

### Marginal Costs

Although the number of studies of the *total* costs of climate change is small, a larger number of studies estimate the *marginal* costs. The marginal damage cost of carbon dioxide is defined as the net present value of the incremental damage due to an infinitesimally small increase in carbon dioxide emissions. If this is computed along the optimal trajectory of emissions, the marginal damage cost equals the Pigou tax. Marginal damage cost estimates derive from total cost estimates – the fact that there are more estimates available, does not imply that we know more about the marginal costs than we do about the total costs. In fact, some of the total cost estimates [59,66,68,78,92] have yet to be used for marginal cost estimation, so that the empirical basis is actually smaller.

Tol [110] gathers 211 estimates of the SCC from 47 studies. The studies were grouped in those that were peer-reviewed and those that were not. Some studies are based on original estimates of the total costs of climate change, while other studies borrow total costs estimates from other studies. Most studies use incremental or marginal calculus to estimate the SCC, as they should, while a few others use average impacts or an unspecified method. Some studies

assume that climate changes but society does not, while other studies include a dynamic model of vulnerability. A few studies use entirely arbitrary assumptions about future climate change, while most studies are based on internally consistent scenarios. These classifications are used as quality indicators. More recent studies were given a higher weight. Many studies report multiple estimates. Most of the estimates are sensitivity analyses around a central estimate, and some estimates are only included to (approximately) reproduce an earlier study. Tol [110] introduces additional weights to account for this.

Tol [110] adjusts a Fisher–Tippett kernel density estimator to 211 data points, weighted as describe above. The 211 estimates provide the modes. Only a few of the studies provide an estimate of the uncertainty. Therefore, the standard deviation is set equal to the sample standard deviation.

Table 2 shows selected characteristics of the kernel distribution for the whole sample and selected sub-samples.

Splitting the sample by discount rate used has the expected effect: A higher discount rate implies a lower estimate of the SCC and a thinner tail. Table 2 also shows that estimates in the peer reviewed literature are lower and less uncertain than estimates in the gray literature.

Splitting the sample by publication date, shows that the estimates of the SCC published before AR2 [87] were larger than the estimates published between AR2 and AR3 [99], which in turn were larger than the estimates published since. Note that these differences are not statistically significant if one considers the means and standard deviation. However, the kernel distribution clearly shifts to the left. Therefore, AR4 [96] were incorrect to conclude that the economic estimates of the impact of climate change have *increased* since 2001. In their words (pp. 781): “There is some evidence that initial new market benefits from climate change will peak at a lower magnitude and sooner than was assumed for the TAR, and it is likely that there will be higher damages for larger magnitudes of global mean temperature increases than was estimated in the TAR.” It is unclear how Schneider et al. [96] reached this conclusion, but it is not supported by the data presented here.

The SCC estimate by Stern et al. [100] is almost an outlier in the entire sample (excluding, of course, the Stern estimate itself). Depending on the kernel density, the Stern estimate lies between the 90th and the 94th percentile. It fits in better with estimates that use a low discount rate and were not peer-reviewed – characteristics of the Stern Review – but even in comparison to those studies, Stern et al. [100] are on the high side. The Stern estimate also fits in better with the older studies. This is no sur-

Climate Change, Economic Costs of, Table 2

Selected characteristics (mode, mean, standard deviation, median, 90-percentile, 95-percentile, 99-percentile, percentile of the Stern estimate) of the joint probability density of the social cost of carbon for the whole sample (all) and selected subsamples (pure rate of time preference, review process, and publication date)

	All	PRTP			Review		Publication date		
		0%	1%	3%	peer	gray	<1996	96–01	>2001
Mode	35	129	56	14	20	53	36	37	27
Mean	127	317	80	24	71	196	190	120	88
St.Dev.	243	301	70	21	98	345	392	179	121
Median	74	265	72	21	48	106	88	75	62
90%	267	722	171	51	170	470	397	274	196
95%	453	856	204	61	231	820	1555	482	263
99%	1655	1152	276	82	524	1771	1826	867	627
Stern	0.92	0.56	1.00	1.00	0.97	0.84	0.86	0.92	0.96

prise, as the PAGE model (e. g., [50]) is calibrated to [87] and [99]. Other criticism of the Stern Review can be found in [24,65,79,80,115,119,120].

### Policy Implications

The policy implications of the above findings are several, and not necessarily in line with the conceived wisdom of climate policy. First and foremost, the economic impacts literature points out that climate change is a problem. Initial climate change may be beneficial, but it cannot be avoided. This is a sunk benefit. The avoidable part of climate change is in all likelihood negative. This justifies greenhouse gas emission reduction.

Second, the estimates of the marginal damage costs justify some emission abatement, but not too much. For instance, the future price of carbon dioxide emission permits in the European Trading System is around \$ 100/tC. Using the market rate of discount, the expected social cost of carbon is only \$ 24/tC. This climate policy has a benefit-cost ratio of 0.24. EU climate policy is therefore too stringent. Of course, European climate policy does pass the cost-benefit test according to CEC [17], but this study does not meet conventional standards of academic quality [111]. Earlier, Pearce [89] similarly concluded that the UK cost-benefit analysis [20] is deficient, while also the *Stern Review* [100] has been criticized in the academic literature (e. g., [80,119]).

Third, climate policy is about ethics rather than about economics [37,116]. The judgment what to do about greenhouse gas emissions rests on the values one attaches to far-flung countries and distant futures. The ethics are not straightforward, however. If one places a lot of weight on the future, one should make a trade-off between increasing investment in capital goods, education, emission

reduction, or technology. If one places a lot of weight on people in poor countries, one should make a trade-off between adaptation, development, emission reduction, and trade reform.

Fourth, the uncertainties about the economic impact of climate change are profound. Partly, this is because the subject is complex. A large share of the uncertainty can be explained, however, by the dearth of research funding. Although climate change is often said to be the largest (environmental) challenge of our times, very few researchers are funded to substantiate or refute that claim.

### Future Directions

Further research is therefore needed. Several problems with past and present research are identified above. Firstly, research into the economic impact of climate change is rightly classified as “applied research”. This implies, however, that research funding comes from bodies with a stake in the result, and that quality and independence are not necessarily overriding concerns. The *Stern Review* is the most prominent example in the recent past of a study that started with the conclusions and worked back to identify the required assumptions. The new *Centre for Climate Change Economics and Policy* at the *London School of Economics* may fall into the same trap. Because the stakes in climate policy are large, academic quality of research must be guaranteed.

Secondly, there are only two groups of independent academics who study the economic impact of climate change. These groups are not sufficiently funded. More importantly, these groups are rarely challenged. Combined with the first problem, it is therefore important to establish a third group of independent, academic economists to study the impact of climate change.

Thirdly, research on the impact of climate change, economic and otherwise, has been lamp-posting. After the groundbreaking work in the early 1990s, researchers have refined previous estimates. Little attention has been paid to those impacts for which no previous estimates exist. While this is the normal procedure of gradual progress in scientific research, the study of the impact of climate change is still in its formative stages. Not just *more*, but particularly *different* research is needed – into the economic effects of climate change on biodiversity, on violent conflict, on ice shelves and ocean current, and on economic development in the long term.

### Acknowledgments

My thoughts on this subject were shaped by discussions with Hadi Dowlatabadi, Tom Downing, Sam Fankhauser, David Maddison, Rob Mendelsohn, Bill Nordhaus, David Pearce, Steve Schneider, Joel Smith, and Gary Yohe.

### Bibliography

- Acemoglu D, Johnson S, Robinson JA (2001) The colonial origins of comparative development: an empirical investigation. *Am Econ Rev* 91:1369–1401
- Adger WN (2006) Vulnerability. *Glob Environ Chang* 16:268–281
- Arrow KJ, Cline WR, Maeler KG, Munasinghe M, Squitieri R, Stiglitz JE (1996) Intertemporal equity, discounting, and economic efficiency. In: Bruce JP, Lee H, Haites EF (eds) *Climate change 1995: economic and social dimensions – contribution of working group iii to the second assessment report of the intergovernmental panel on climate change*. Cambridge University Press, Cambridge, pp 125–144
- Ashley RM, Balmford DJ, Saul AJ, Blanksby JD (2005) Flooding in the future – predicting climate change, risks and responses in urban areas. *Water Sci Technol* 52(5):265–273
- Ayres RU, Walter J (1991) The greenhouse effect: damages, costs and abatement. *Environ Resour Econ* 1:237–270
- Azar C, Sterner T (1996) Discounting and distributional considerations in the context of global warming. *Ecol Econ* 19:169–184
- Berrens RP, Bohara AK, Jenkins-Smith HC, Silva CL, Weimer DL (2004) Information and effort in contingent valuation surveys: application to global climate change using national internet samples. *J Environ Econ Manag* 47:331–363
- Berrittella M, Bigano A, Roson R, Tol RSJ (2006) A general equilibrium analysis of climate change impacts on tourism. *Tour Manag* 27(5):913–924
- Bosello F, Roson R, Tol RSJ (2006) Economy-wide estimates of the implications of climate change: human health. *Ecol Econ* 58:579–591
- Bosello F, Roson R, Tol RSJ (2007) Economy-wide estimates of the implications of climate change: sea level rise. *Environ Resour Econ* 37:549–571
- Breslow PB, Sailor DJ (2002) Vulnerability of wind power resources to climate change in the continental united states. *Renew Energy* 27(4):585–598
- Broome J (1992) *Counting the cost of global warming*. White Horse Press, Cambridge
- Burkett VR, Wilcox DA, Stottlemeyer R, Barrow W, Fagre D, Baron J, Price J, Nielson JL, Allen CD, Peterson DL, Ruggerone G, Doyle T (2005) Nonlinear dynamics in ecosystem response to climate change: case studies and policy implications. *Ecol Complex* 2(4):357–394
- Burton I, Kates RW, White GF (1993) *The environment as hazard*, 2nd edn. The Guilford Press, New York
- Carmichael CG, Gallus Jr WA, Temeyer BR, Bryden MK (2004) A winter weather index for estimating winter road maintenance costs in the midwest. *J Appl Meteorol* 43(11):1783–90
- CBO (2005) *Uncertainty in analyzing climate change: policy implications*. congress of the united states. Congressional Budget Office, Washington
- CEC (2005) *Winning the battle against global climate change – background paper*. Commission of the European Communities, Brussels
- Champ PA, Boyle KJ, Brown TC (eds) (2003) *A primer on non-market valuation*. Kluwer, Dordrecht
- Chichilnisky G (1996) An axiomatic approach to sustainable development. *Soc Choice Welf* 13(2):219–248
- Clarkson R, Deyes K (2002) *Estimating the social cost of carbon emissions*. Working Paper 140. The Public Enquiry Unit – HM Treasury, London
- Cline WR (1992) *The economics of global warming*. Institute for International Economics, Washington
- Cropper ML, Aydede SK, Portney PR (1992) Rates of time preference for saving lives. *Am Econ Rev* 82(2):469–472
- Darwin RF (1999) A FARMer's view of the ricardian approach to measuring agricultural effects of climatic change. *Clim Chang* 41(3–4):371–411
- Dasgupta P (2007) *Commentary: The stern review's economics of climate change*. *Natl Inst Econ Rev* 199:4–7
- Dasgupta P, Maskin E (2005) Uncertainty and hyperbolic discounting. *Am Econ Rev* 95(4):1290–1299
- Davidson MD (2006) A social discount rate for climate damage to future generations based on regulatory law. *Clim Chang* 76:55–72
- Dorland C, Tol RSJ, Palutikof JP (1999) Vulnerability of the netherlands and northwest europe to storm damage under climate change. *Clim Chang* 43:513–535
- Easterly W, Levine R (2003) Tropics, germs, and crops: how endowments influence economic development. *J Monet Econ* 50:3–39
- Elsasser H, Buerki R (2002) Climate change as a threat to tourism in the alps. *Clim Res* 20(3):253–257
- Ennos AR (1997) Wind as an ecological factor. *Trends Ecol Evol* 12(3):108–111
- Evans DJ, Sezer H (2004) Social discount rates for six major countries. *Appl Econ Lett* 11:557–560
- Fankhauser S (1994) The economic costs of global warming damage: a survey. *Glob Environ Chang* 4(4):301–309
- Fankhauser S (1995) *Valuing climate change – the economics of the greenhouse*, 1st edn. EarthScan, London
- Fankhauser S, Tol RSJ (1996) Recent advancements in the economic assessment of climate change costs. *Energy Policy* 24(7):665–673
- Fankhauser S, Tol RSJ (2005) On climate change and economic growth. *Resour Energy Econ* 27:1–17
- Fankhauser S, Tol RSJ, Pearce DW (1997) The aggregation of

- climate change damages: a welfare theoretic approach. *Environ Resour Econ* 10:249–266
37. Gardiner SM (2006) A perfect moral storm: climate change, intergenerational ethics and the problem of moral corruption. *Environ Values* 15:397–413
  38. Gerlagh R, Keyzer MA (2001) Sustainability and the intergenerational distribution of natural resource entitlements. *J Public Econ* 79:315–341
  39. Gollier C (2002) Discounting an uncertain future. *J Public Econ* 85:149–166
  40. Gollier C (2002) Time horizon and the discount rate. *J Econ Theor* 107:463–473
  41. Guo JK, Hepburn C, Tol RSJ, Anthoff D (2006) Discounting and the social cost of carbon: a closer look at uncertainty. *Environ Sci Policy* 9(5):203–216
  42. Hamilton JM, Maddison DJ, Tol RSJ (2005) Climate change and international tourism: a simulation study. *Glob Environ Chang* 15(3):253–266
  43. Hamilton JM, Maddison DJ, Tol RSJ (2005) The effects of climate change on international tourism. *Clim Res* 29:255–268
  44. Harrison GP, Wallace AR (2005) Sensitivity of wave energy to climate change. *IEEE Trans Energy Convers* 20(4):870–877
  45. Harvey D, Huang Z (1995) Evaluation of the potential impact of methane clathrate destabilization on future global warming. *J Geophys Res* 100:2905–2926
  46. Hersch J, Viscusi WK (2006) The generational divide in support for environmental policies: european evidence. *Clim Chang* 77:121–136
  47. Hitz S, Smith JB (2004) Estimating global impacts from climate change. *Glob Environ Chang* 14:201–218
  48. Hohmeyer O, Gaertner M (1992) The costs of climate change – a rough estimate of orders of magnitude. Fraunhofer-Institut für Systemtechnik und Innovationsforschung, Karlsruhe
  49. Homer-Dixon TF (1994) Environmental scarcities and violent conflict: evidence from cases. *Int Secur* 19(1):5–40
  50. Hope CW (2006) The marginal impact of CO<sub>2</sub> from PAGE2002: an integrated assessment model incorporating the IPCC's five reasons for concern. *Integr Assess J* 6(1):19–56
  51. Hope CW, Maul P (1996) Valuing the impact of CO<sub>2</sub> emissions. *Energy Policy* 24(3):211–219
  52. Kikkawa T, Kita J, Ishumatsu A (2004) Comparison of the lethal effect of CO<sub>2</sub> and acidification on red sea bream (*pagrus major*) during the early developmental stages. *Marine Pollut Bull* 48(1–2):108–110
  53. Kundzewicz ZW, Graczyk D, Maurer T, Pinskiwar I, Radziejewski M, Svensson C, Szwed M (2005) Trend detection in river flow series: 1, annual maximum flow. *Hydrol Sci J* 50(5):797–810
  54. Li H, Berrens RP, Bohara AK, Jenkins-Smith HC, Silva CL, Weimer DL (2004) Exploring the beta model using proportional budget information in a contingent valuation study. *Ecol Econ* 28:329–343
  55. Lind RC, Schuler RE (1998) Equity and discounting in climate change decisions. In: Nordhaus WD (ed) *Economics and policy issues in climate change*. Resources for the Future, Washington, pp 59–96
  56. Lind RC (1995) Intergenerational equity, discounting, and the role of cost-benefit analysis in evaluating global climate policy. *Energy Policy* 23(4/5):379–389
  57. Link PM, Tol RSJ (2004) Possible economic impacts of a shut-down of the thermohaline circulation: an application of FUND. *Port Econ J* 3:99–114
  58. Long SP, Ainsworth EA, Leakey ADB, Noesberger J, Ort DR (2006) Food for thought: lower-than-expected crop yield stimulation with rising CO<sub>2</sub> concentrations. *Science* 312: 1918–1921
  59. Maddison DJ (2003) The amenity value of the climate: the household production function approach. *Resour Energy Econ* 25:155–175
  60. Manne AS, Richels RG (2004) US rejection of the kyoto protocol: the impact on compliance costs and CO<sub>2</sub> emissions. *Energy Policy* 32:447–454
  61. Marotzke J (2000) Abrupt climate change and thermohaline circulation: mechanisms and predictability. *Proc Natl Acad Sci* 97:1347–1350
  62. Masters WA, McMillan MS (2001) Climate and scale in economic growth. *J Econ Growth* 6:167–186
  63. McDonald RE, Bleaken DG, Cresswell DR, Pope VD, Senior CA (2005) Tropical storms: representation and diagnosis in climate models and the impacts of climate change. *Clim Dyn* 25(1):19–36
  64. McLeman R, Smit B (2006) Migration as an adaptation to climate change. *Clim Chang* 76:31–53
  65. Mendelsohn RO (2006) A critique of the stern report. *Regulation* (Winter 2006–2007):42–46
  66. Mendelsohn RO, Morrison W, Schlesinger ME, Andronova NG (2000) Country-specific market impacts of climate change. *Clim Chang* 45:553–569
  67. Mendelsohn RO, Nordhaus WD, Shaw D (1994) The impact of climate on agriculture: a ricardian analysis. *Am Econ Rev* 84(4):753–771
  68. Mendelsohn RO, Schlesinger ME, Williams LJ (2000) Comparing impacts across climate models. *Int Assess* 1:37–48
  69. Mulcahy M (2004) Weathering the storms: hurricanes and risk in the british greater caribbean. *Bus Hist Rev* 78(4):635–663
  70. Nakicenovic N, Swart RJ (eds) (2001) *IPCC special report on emissions scenarios*. Cambridge University Press, Cambridge
  71. Newell RG, Pizer WA (2003) Discounting the distant future: how much do uncertain rates increase valuations? *J Environ Econ Manag* 46:52–71
  72. Nicholls RJ, Tol RSJ (2006) Impacts and responses to sea-level rise: a global analysis of the SRES scenarios over the 21st Century. *Phil Trans Royal Soc A Math Phys Eng Sci* 361(1841):1073–1095
  73. Nordhaus WD (1982) How fast should we graze the global commons? *Am Econ Rev* 72(2):242–246
  74. Nordhaus WD (1991) To Slow or Not to Slow: The economics of the greenhouse effect. *Econ J* 101:920–937
  75. Nordhaus WD (1993) Rolling the 'DICE': An optimal transition path for controlling greenhouse gases. *Resour Energy Econ* 15:27–50
  76. Nordhaus WD (1994) *Managing the global commons: the economics of climate change*. The MIT Press, Cambridge
  77. Nordhaus WD (1994) Expert opinion on climate change. *Am Sci* 82(1):45–51
  78. Nordhaus WD (2006) Geography and macroeconomics: new data and new findings. *Proc Natl Acad Sci* 103(10):3510–3517. [www.pnas.org/cgi/doi/10.1073/pnas.0509842103](http://www.pnas.org/cgi/doi/10.1073/pnas.0509842103)
  79. Nordhaus WD (2007) Critical assumptions in the stern review on climate change. *Science* 317:201–202
  80. Nordhaus WD (2007) A review of the stern review on the economics of climate change. *J Econ Lit* 45(3):686–702

81. Nordhaus WD, Boyer JG (2000) *Warming the world: economic models of global warming*. The MIT Press, Cambridge
82. Nordhaus WD, Yang Z (1996) RICE: A regional dynamic general equilibrium model of optimal climate-change policy. *Am Econ Rev* 86(4):741–765
83. O'Brien KL, Sygna L, Haugen JE (2004) Vulnerable or resilient? A multi-scale assessment of climate impacts and vulnerability in norway. *Clim Chang* 64:193–225
84. Oppenheimer M, Alley RB (2005) Ice sheets, global warming, and article 2 of the UNFCCC. *Clim Chang* 68:257–267
85. Paavola J, Adger WN (2006) Fair adaptation to climate change. *Ecol Econ* 56:594–609
86. Pearce DW (2003) The social cost of carbon and its policy implications. *Oxford Rev Econ Policy* 19(3):1–32
87. Pearce DW, Cline WR, Achanta AN, Fankhauser S, Pachauri RK, Tol RSJ, Vellinga P (1996) The social costs of climate change: greenhouse damage and the benefits of control. In: Bruce JP, Lee H, Haites EF (eds) *Climate Change 1995: Economic and Social Dimensions – Contribution of Working Group III to the Second Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, pp 179–224
88. Pearce DW, Moran D (1994) *The economic value of biodiversity*. EarthScan, London
89. Pielke Jr RA, Landsea C, Mayfield M, Laver J, Pasch R (2005) Hurricanes and global warming. *Bull Am Meteorol Soc* 86(11):1571–1575
90. Plambeck EL, Hope CW (1996) PAGE95 – An updated valuation of the impacts of global warming. *Energy Policy* 24(9):783–793
91. Ready R, Navrud S, Day B, Dubourg R, Machado F, Mourato S, Spaninks F, Rodriguez MXV (2004) Benefit transfer in europe: How reliable are transfers between countries? *Environ Resour Econ* 29(1):67–82
92. Rehdanz K, Maddison DJ (2005) Climate and happiness. *Ecol Econ* 52:111–125
93. Schelling TC (1984) *Choice and consequence*. Harvard University Press, Cambridge
94. Schelling TC (1992) Some economics of global warming. *Am Econ Rev* 82:1–14
95. Schelling TC (1995) Intergenerational discounting. *Energy Policy* 23(4/5):395–401
96. Schneider SH, Semenov S, Patwardhan A, Burton I, Magadya CHD, Oppenheimer M, Pittock AB, Rahman A, Smith JB, Suarez A, Yamin F (2007) Assessing key vulnerability and the risk from climate change. In: Parry ML et al (eds) *Climate Change 2007: Impacts, Adaptation and Vulnerability – Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, pp 779–810
97. Scott D, McBoyle G, Mills B (2003) Climate change and the skiing industry in southern ontario (Canada): exploring the importance of snowmaking as a technical adaptation. *Clim Res* 23:171–181
98. Smith JB (1996) Standardized estimates of climate change damages for the united states. *Clim Chang* 32(3):313–326
99. Smith JB, Schellnhuber HJ, Mirza MMQ, Fankhauser S, Leemans R, Lin E, Ogallo L, Pittock B, Richels RG, Rosenzweig C, Tol RSJ, Weyant JP, Yohe GW (2001) Vulnerability to climate change and reasons for concern: a synthesis. In: Mccarthy JJ, Canziani OF, Leary NA, Dokken DJ, White KS (eds) *Climate change 2001: Impacts, adaptation, and vulnerability*, vol 19. Cambridge University Press, Cambridge, pp 913–967
100. Stern NH, Peters S, Bakhshi V, Bowen A, Cameron C, Catovsky S, Crane D, Cruickshank S, Dietz S, Edmonson N, Garbett SL, Hamid L, Hoffman G, Ingram D, Jones B, Patmore N, Radcliffe H, Sathiyarajah R, Stock M, Taylor C, Vernon T, Wanjie H, Zenghelis D (2006) *Stern review: The economics of climate change*. Cambridge University Press, Cambridge
101. Szolnoky C, Buzas K, Clement A (1997) Impacts of the climate change on the operation of a freshwater cooled electric power plant. *Periodica Polytechnica: Civil Eng* 41(2):71–94
102. Titus JG (1992) The costs of climate change to the united states. In: Majumdar SK et al (eds) *Global climate change: implications, challenges and mitigation measures*. Pennsylvania Academy of Science, Easton, pp 384–409
103. Tol RSJ (1995) The damage costs of climate change – towards more comprehensive calculations. *Environ Resour Econ* 5:353–374
104. Tol RSJ (2001) Equitable cost-benefit analysis of climate change. *Ecol Econ* 36(1):71–85
105. Tol RSJ (2002) New estimates of the damage costs of climate change, Part I: Benchmark Estimates. *Environ Resour Econ* 21(1):47–73
106. Tol RSJ (2003) Is the uncertainty about climate change too large for expected cost-benefit analysis? *Clim Chang* 56(3):265–289
107. Tol RSJ (2005) The marginal damage costs of carbon dioxide emissions: an assessment of the uncertainties. *Energy Policy* 33(16):2064–2074
108. Tol RSJ (2005) Emission abatement versus development as strategies to reduce vulnerability to climate change: an application of FUND. *Environ Dev Econ* 10:615–629
109. Tol RSJ (2005) Adaptation and mitigation: trade-offs in substance and methods. *Environ Sci Policy* 8:572–578
110. Tol RSJ (2007) The social cost of carbon: trends, outliers and catastrophes, research unit sustainability and global change FNU-144. Hamburg University and Centre for Marine and Atmospheric Science, Hamburg
111. Tol RSJ (2007) Europe's long-term climate target: A critical evaluation. *Energy Policy* 35:424–432
112. Tol RSJ, van der Grijp NM, Olsthoorn AA, van der Werff PE (2003) Adapting to climate change: A case study of riverine flood risks in the netherlands. *Risk Analysis* 23(3):575–583
113. Tol RSJ, Verheyen R (2004) State responsibility and compensation for climate change damages – a legal and economic assessment. *Energy Policy* 32:1109–1130
114. Tol RSJ, Yohe GW (2006) Of dangerous climate change and dangerous emission reduction. In: Schellnhuber HJ, Cramer W, Nakicenovic N, Wigley T, Yohe G (eds) *Avoiding dangerous climate change*. Cambridge University Press, Cambridge, pp 291–298
115. Tol RSJ, Yohe GW (2006) A review of the stern review. *World Econ* 7(4):233–250
116. Toman M (2006) Values in the economics of climate change. *Environ Values* 15:365–379
117. Viscusi WK, Zeckhauser RJ (2006) The perception and valuation of the risks of climate change: A rational and behavioral blend. *Clim Chang* 77:151–177
118. Weitzman ML (2001) Gamma discounting. *Am Econ Rev* 91(1):260–271

119. Weitzman ML (2007) A review of the stern review on the economics of climate change. *J Econ Lit* 45(3):703–724
120. Weitzman ML (2008) On modeling and interpreting the economics of catastrophic climate change. *Rev Econ Stat*
121. Wilson KJ, Falkingham J, Melling H, de Abreu R (2004) Shipping in the canadian arctic: Other possible climate change scenarios. *Int Geosci Remote Sens Symp* 3:1853–1856
122. Yohe GW, Schlesinger ME (2002) The economic geography of the impacts of climate change. *J Econ Geogr* 2:311–341
123. Yohe GW, Tol RSJ (2002) Indicators for social and economic coping capacity – moving towards a working definition of adaptive capacity. *Glob Environ Chang* 12(1):25–40
124. Zhang DD, Jim CY, Lin GCS, He YQ, Wang JJ, Lee HF (2006) Climatic change, wars and dynastic cycles in China over the last millennium. *Clim Chang* 76:459–477



## Climate Modeling, Global Warming and Weather Prediction, Introduction to

HARTMUT GRASSL  
Max Planck Institute for Meteorology,  
Hamburg, Germany

All systems operating away from thermodynamic equilibrium develop structures. The planet Earth will always be far away from thermodynamic equilibrium because of the strongly differing solar radiation input as a function of latitude and season. Hence, the differential heating of the surface and also the atmosphere must lead to temperature gradients, in turn causing pressure gradients that create currents in the ocean and the wind in the atmosphere. On a rotating sphere (in reality the geoid, which is close to a rotational ellipsoid) these flows form low and high pressure systems both in the ocean and the atmosphere which are able to reduce latitudinal gradients but never come close to thermodynamic equilibrium because of continuing differential heating. The average temperatures and flow fields, as well as their strong spatial and temporal variability, are a function of land/sea distribution and atmospheric composition, especially depending on water and ice in clouds. The strongly climatically relevant gases in the atmosphere are, to a large extent, a consequence of life on Earth.

Ranking all radiatively active gases in the atmosphere according to their influence on weather and climate shows the exceptional composition of the atmosphere: Water vapor ( $\text{H}_2\text{O}$ ) in all three phases but largely as a gas, carbon dioxide ( $\text{CO}_2$ ), ozone ( $\text{O}_3$ ), nitrous oxide ( $\text{N}_2\text{O}$ ) and methane ( $\text{CH}_4$ ) constitute only three thousandths of the atmospheric mass, yet they largely determine how much solar radiation reaches the surface, e. g., through clouds, and how much thermal or terrestrial radiation leaves from there to space, again a strong function of clouds and the above-mentioned gases. The average surface temperature is thus strongly depending on the concentration of the gases mentioned, which are all greenhouse gases. They do not strongly absorb solar radiation but do absorb terrestrial radiation, thereby forcing the surface and the lower atmosphere to warm in order to reach nearly equilibrium between absorbed and emitted energy. Any growth or reduction of greenhouse gas concentrations increases or decreases average surface temperature and thus changes climate.

The climate system, i. e., its interacting components at atmosphere, ocean, land, vegetation, soils and crust, shows

both a remarkable stability and high sensitivity. Over many million years the greenhouse effect of about 30 K has varied only by about  $\pm 5$  K with respect to present interglacial temperatures, thus has been stable in terms of temperature varying only from 283 to 293 K.  $+5$  K however meant melting of all inland ice sheets and  $-5$  K a new strong glaciation with major ice sheets reaching 40 to  $50^\circ$  N (see ► [Cryosphere Models](#)).

The observed strong increases of all long-lived naturally occurring greenhouse gases ( $+35\%$  for  $\text{CO}_2$ ,  $+120\%$  for methane, and  $+10\%$  for  $\text{N}_2\text{O}$  since 1750) have stimulated a mean global warming which now has emerged from strong climate variability. In 2007, the Fourth Assessment Report of the Intergovernmental Panel on Climate Change concluded: “The understanding of anthropogenic warming and cooling influences has improved . . . leading to very high confidence that the global average net effect of human activities since 1750 has been one of warming”. On the other hand, the high sensitivity of the climate system is demonstrated in so called abrupt climate change events (see ► [Abrupt Climate Change Modeling](#)).

The strongest global one of these – besides the impact of celestial bodies – is deglaciation after an intense glaciation in about 5000 to 10 000 years with a concomitant temperature increase of 4 to 5 K, caused by the slow latitudinal redistribution of solar radiation due to Earth orbit parameter changes as a consequence of slowly changing gravitational forcing by the neighboring planets, mainly Venus, Jupiter and Saturn. Because climate is a key natural resource for plants, animals and humans, any rapid climate change threatens life on Earth. Therefore agriculture (see ► [Climate Change and Agriculture](#)), forestry and all economic activities (see ► [Climate Change, Economic Costs of](#)) will be impacted by anthropogenic climate change in the 21st century, leading to strong consequences in societal behavior vis-a-vis this challenge, e. g., the one caused by growing inequity between those societies causing climate change, the industrialized countries, and those suffering first or more strongly such as subsistence farmers in semi-arid tropical areas. Climate Models developed so far include many physical processes (see ► [Single Column Modeling of Atmospheric Boundary Layers and the Complex Interactions with the Land Surface](#) as an example), parts of atmospheric chemistry, and vegetated land surface atmosphere interactions, but still lack reaction of ocean biomass production to enhanced  $\text{CO}_2$  levels. Finally, it is apparent that global warming will have an effect on human health and that this will include effects on food crops and animals (see ► [Climate Change and Human Health](#)).

The low horizontal resolution of global climate models has stimulated nested regional climate models (see ► [Re-](#)

gional Climate Models: Linking Global Climate Change to Local Impacts) delivering enhanced output in areas with strong topography or sea/land contrasts. The slowly emerging Earth System Models are no longer driven by changed atmospheric composition alone but by emissions, i. e., they can calculate resulting greenhouse gas concentrations. However they are still not advanced enough to answer the question: When will growing climate change stress turn the present uptake of anthropogenic CO<sub>2</sub> into forests through the CO<sub>2</sub>-fertilization into an additional CO<sub>2</sub> source for the atmosphere due to a generally weakened vegetation? It is common knowledge that weather, for example the passage of a coldfront at a certain location, can be forecast only for up to two weeks because of the intrinsically chaotic behavior of atmospheric flow. The accuracy of present day weather forecast models, which are very similar to the atmospheric component of climate

models, has advanced strongly, driven by higher spatial resolutions and better parametrizations of sub-grid scale processes in the models, assimilation of more (and especially satellite data) into the models and ensemble forecasting. This has recently led to the same forecast accuracy up to about 10 days in the southern hemisphere, where an in situ observing system for the starting fields of the model is largely lacking.

The increased attention the topic climate change has finally attracted in the political arena will certainly accelerate progress in this field, despite the complex nature of the functioning of the Earth system.

There are two additional articles on climate change which were recruited for other sections. These articles are: ► [Dynamic Games with an Application to Climate Change Models](#) and ► [System Dynamics Models of Environment, Energy and Climate Change](#).

## Complexity in Earthquakes, Tsunamis, and Volcanoes, and Forecast, Introduction to

WILLIAM H. K. LEE  
US Geological Survey (Retired), Menlo Park, USA

### Article Outline

[Introduction](#)  
[Earthquakes](#)  
[Tsunamis](#)  
[Volcanoes](#)  
[Discussions](#)

This Introduction is intended to serve as a ‘road map’ for readers to navigate through the 42 Encyclopedia articles on earthquakes, tsunamis, and volcanoes. Selecting the topics and authors was somewhat subjective, as it is not possible to cover the vast existing literature with only 42 articles. They are, however, representative of the wide range of problems investigated in connection with these natural phenomena. I will introduce these articles by grouping them into sections and then into subsections. However, some articles belong to more than one section or one subsection, reflecting the inter-related nature of earthquakes, tsunamis and volcanoes. For the benefit of the readers, I will point to certain issues discussed in some of the articles which, in my view, have not been settled completely. I have also taken the liberty of quoting or paraphrasing sentences from many of these articles when introducing them, but I do not claim to be accurate. It is best for these articles to speak for themselves.

I wish to thank Bernard Chouet for helping me in planning and reviewing the manuscripts of the volcanoes section. I am grateful to Bernard Chouet, Edo Nyland, Jose Pujol, Chris Stephens, and Ta-liang Teng for their helpful comments that greatly improved this manuscript.

### Introduction

Earthquakes, tsunamis, and volcanic eruptions are complex and often inter-related natural phenomena with disastrous impact to society rivaling those caused by the worst floods or storms. The 1556 Huaxian earthquake in the Shansi province of China claimed over 830,000 lives. The total economic loss of the 1995 Kobe earthquake in Japan was estimated at US \$200 billion. The 2004 Indian Ocean tsunami (triggered by the Sumatra–Andaman earthquake of December 26) brought devastation thousands of miles away with fatalities exceeding 280,000. The 79 AD erup-

tion of Mount Vesuvius near Naples, Italy buried the towns of Pompeii and Herculaneum. The 1902 eruption of Mount Pelée, Martinique, totally destroyed the town of St. Pierre.

Insurance companies classify major natural catastrophes as storms, floods, or earthquakes (including tsunamis, and volcanic eruptions). Since 1950, about 2.5 million people have died due to these catastrophes and overall economic losses have totaled about US \$2 trillion in current dollar values. Earthquakes, tsunamis, and volcanic eruptions have accounted for about half of the fatalities and more than one third of the total economic losses. Geoscientists have attempted to predict these events, but with limited success. There are many reasons for such slow progress: (1) systematic monitoring of earthquakes, tsunamis and volcanoes requires large capital investment for instruments and very long-term support for operation and maintenance; (2) catastrophic earthquakes, tsunamis and volcanic eruptions occur rarely, and (3) politicians and citizens are quick to forget these hazards in the face of other more frequent and pressing issues. But with continuing rapid population growth and urbanization, the loss potential from these natural hazards in the world is quickly escalating.

With advances in nonlinear dynamics and complexity studies, geoscientists have applied modern nonlinear techniques and concepts such as chaos, fractal, critical phenomena, and self-organized criticality to the study of earthquakes, tsunamis and volcanoes. Here we sample these efforts, mainly in seismicity modeling for earthquake prediction and forecast, along with articles that review recent progress in studying earthquakes, tsunamis, and volcanoes. Although predictability is desirable, it is also possible to reduce these natural hazards with more practical approaches, such as early warning systems, hazard analysis, engineering considerations, and other mitigation efforts. Several articles in this Encyclopedia discuss these practical solutions.

### Earthquakes

When a sudden rupture occurs in the Earth, seismic waves are generated. When these waves reach the Earth’s surface, we may feel them as a series of vibrations, which we call an earthquake. Instrumental recordings of earthquakes have been made since the latter part of the 19th century by seismographic stations and networks from local to global scales. The observed data have been used, for example, (1) to compute the source parameters of earthquakes, (2) to determine the physical properties of the Earth’s interior, (3) to test the theory of plate tectonics,

(4) to map active faults, (5) to infer the nature of damaging ground shaking, (6) to carry out seismic hazard analyzes, and (7) to predict and forecast earthquakes. A satisfactory theory of the complex earthquake process has not yet been achieved, and realistic equations for modeling earthquakes do not exist at present. There is, however, good progress towards a physical foundation for the earthquake source process, partly as a result of research directed toward earthquake prediction.

### **Earthquake Monitoring, and Probing the Earth's Interior**

Earthquakes are complex natural phenomena, and their monitoring requires an interdisciplinary approach, including using tools from other scientific disciplines and engineering. In ► **Earthquake Monitoring and Early Warning Systems**, W.H.K. Lee and Y.M. Wu presented a summary of earthquake monitoring, a description of the products derived from the analysis of seismograms, and a discussion of the limitations of these products. The basic results of earthquake monitoring are summarized in earthquake catalogs, which are lists of origin time, hypocenter location, and magnitude of earthquakes, as well as other source parameters. Lee and Wu describe the traditional earthquake location method formulated as an inverse problem. In ► **Earthquake Location, Direct, Global-Search Methods**, Lomax et al. review a different approach using direct-search over a space of possible locations, and discuss other related algorithms. Direct-search earthquake location is important because, relative to the traditional linearized method, it is both easier to apply to more realistic Earth models and is computationally more stable. Although it has not been widely applied because of its computational demand, it shows great promise for the future as computer power is advancing rapidly.

The most frequently determined parameter after 'location' is 'magnitude', which is used to characterize the 'size' of an earthquake. A brief introduction to the quantification of earthquake size, including magnitude and seismic moment, is given in ► **Earthquake Monitoring and Early Warning Systems** by Lee and Wu. Despite its various limitations, magnitude provides important information concerning the earthquake source. Magnitude values have an immense practical value for realistic long-term disaster preparedness and risk mitigation efforts. A detailed review, including current practices for magnitude determinations, appears in ► **Earthquake Magnitude** by P. Bormann and J. Saul.

Besides computing earthquake source parameters, earthquake monitoring also provides data that can be used

to probe the Earth's interior. In ► **Tomography, Seismic**, J. Pujol reviews a number of techniques designed to investigate the interior of the Earth using arrival times and/or waveforms from natural and artificial sources. The most common product of a tomographic study is a seismic velocity model, although other parameters, such as attenuation and anisotropy, can also be estimated. Seismic tomography generally has higher resolution than that provided by other geophysical methods, such as gravity and magnetics, and furnishes information (1) about fundamental problems concerning the internal structure of the Earth on a global scale, and (2) for tectonic and seismic hazard studies on a local scale.

In ► **Seismic Wave Propagation in Media with Complex Geometries, Simulation of**, H. Igel et al. present the state-of-the-art in computational wave propagation. They point to future developments, particularly in connection with the search for efficient generation of computational grids for models with complex topography and faults, as well as for the combined simulation of soil-structure interactions. In addition to imaging subsurface structure and earthquake sources, 3-D wave simulations can forecast strong ground motions from large earthquakes. In the absence of deterministic prediction of earthquakes, the calculation of earthquake scenarios in regions with sufficiently well-known crustal structures and faults will play an important role in assessing and mitigating potential damage, particularly those due to local site effects.

In addition to the classical parametrization of the Earth as a layered structure with smooth velocity perturbation, a new approach using scattered waves that reflect Earth's heterogeneity is introduced by H. Sato in his article on ► **Seismic Waves in Heterogeneous Earth, Scattering of**. For high-frequency seismograms, envelope characteristics such as the excitation level and the decay gradient of coda envelopes and the envelope broadening of the direct wavelet are useful for the study of small-scale inhomogeneities within the Earth. The radiative transfer theory with scattering coefficients calculated from the Born approximation and the Markov approximation for the parabolic wave equation are powerful mathematical tools for these analyzes. Studies of the scattering of high-frequency seismic waves in the heterogeneous Earth are important for understanding the physical structure and the geodynamic processes that reflect the evolution of the solid Earth.

### **Earthquake Prediction and Forecasting**

A fundamental question in earthquake science is whether earthquake prediction is possible. Debate on this question

has been going on for decades without clear resolution. Are pure observational methods without specific physical understanding sufficient? Earthquakes have been instrumentally monitored continuously for about 100 years (although not uniformly over the Earth), but reliable and detailed earthquake catalogs cover only about 50 years. Consequently, it seems questionable that earthquakes can be predicted solely on the basis of observed seismicity patterns, given that large earthquakes in a given region have recurrence intervals ranging from decades to centuries or longer. Despite progress made in earthquake physics, we are still not able to write down all the governing equations for these events and lack sufficient information about the Earth's properties. Nevertheless, many attempts have been and are being made to predict and forecast earthquakes. In this section, several articles based on empirical and physics-based approaches will be briefly introduced.

In ► **Geo-complexity and Earthquake Prediction**, V. Keilis-Borok et al. present an algorithmic prediction method for individual extreme events having low probability but large societal impact. They show that the earthquake prediction problem is necessarily intertwined with problems of disaster preparedness, the dynamics of the solid Earth, and the modeling of extreme events in hierarchical complex systems. The algorithms considered by Keilis-Borok et al. are based on premonitory seismicity patterns and provide alarms lasting months to years. Since the 1990s, these alarms have been posted for use in testing such algorithms against newly occurred large earthquakes. Some success has been achieved, and although the areas for the predicted earthquakes are very large and the predicted time windows are very long, such predictions can be helpful for the officials and the public to undertake appropriate preparedness.

Stochastic models are a practical way of bridging the gap between the detailed modeling of a complex system and the need to fit models to limited data. In ► **Earthquake Occurrence and Mechanisms, Stochastic Models for**, D. Vere-Jones presents a brief account of the role and development of stochastic models of seismicity, from the first empirical studies to current models used in earthquake probability forecasting. The author combines a model of the physical processes generating the observable data (earthquake catalogs) with a model for the errors, or uncertainties, in our ability to predict those observables.

D.A. Yuen et al. propose the use of statistical approaches and data-assimilation techniques to earthquake forecasting in their article on ► **Earthquake Clusters over Multi-dimensional Space, Visualization of**. The nature of the spatial-temporal evolution of earthquakes may be assessed from the observed seismicity and geodetic mea-

surements by recognizing nonlinear patterns hidden in the vast amount of seemingly unrelated data. The authors endeavor to bring across the basic concept of clustering and its role in earthquake forecasting, and conclude that the clustering of seismic activity reflects both the similarity between clusters and their correlation properties.

In ► **Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space**, G. Zoeller et al. present a combined approach to understanding seismicity and the emergence of patterns in the occurrence of earthquakes based on numerical modeling and data analysis. The discussion and interpretation of seismicity in terms of statistical physics leads to the concept of 'critical states', i. e. states in the seismic cycle with an increased probability for abrupt changes involving large earthquakes. They demonstrate that numerical fault models are valuable for understanding the underlying mechanisms of observed seismicity patterns, as well as for practical estimates of future seismic hazard.

D. Sornette and M.J. Werner in ► **Seismicity, Statistical Physics Approaches to** stress that the term 'statistical' in 'statistical physics' has a different meaning than as used in 'statistical seismology'. Statistical seismology has been developed as a marriage between probability theory, statistics, and earthquake occurrences without considerations of earthquake physics. In statistical physics approaches to seismicity, researchers strive to derive statistical models from microscopic laws of friction, damage, rupture, etc. Sornette and Werner summarize some of the concepts and tools that have been developed, including the leading theoretical physical models of the space-time organization of earthquakes. They then present several examples of the new metrics proposed by statistical physicists, underlining their strengths and weaknesses. They conclude that a holistic approach emphasizing the interactions between earthquakes and faults is promising, and that statistical seismology needs to evolve into a genuine physically-based statistical physics of earthquakes.

In ► **Earthquake Networks, Complex**, S. Abe and N. Suzuki discuss the construction of a complex earthquake network obtained by mapping seismic data to a growing stochastic graph. This graph, or network, turns out to exhibit a number of remarkable physical and mathematical behaviors that share common traits with many other complex systems. The scale-free and small-world natures are typical examples in complex earthquake networks.

Electromagnetic phenomena associated with earthquakes, such as earthquake light have been reported throughout almost all human history. Until rather recently, however, most such observations were unreliable and best described as folklore. In ► **Earthquakes, Electro-**

**magnetic Signals of**, S. Uyeda et al. summarize the scientific search for electromagnetic precursors for earthquake prediction. The presumption is that since earthquakes occur when slowly increasing tectonic stress in the Earth's crust reaches a critical level; the same stress may give rise to some electromagnetic phenomena. Research on possible relationships was initiated in several countries around the world in the 1980s. Two main approaches are (1) the monitoring of possible emissions from focal regions in a wide range of frequency from DC to VHF, and (2) the monitoring of possible anomalies in the transmission of man-made electromagnetic waves of various frequencies over focal regions. Despite much circumstantial evidence, earthquake-related electromagnetic signals, in particular those at a pre-seismic stage are not yet widely accepted to be associated with earthquakes.

Observational programs focused on searching for reliable precursory phenomena in seismicity, seismic velocities, tilt and strain, electromagnetic signals, chemical emissions and animal behavior, claim some successes but no systematic precursors have been identified. In ► **Earthquake Forecasting and Verification**, J.R. Holliday et al. stress that reliable earthquake forecasting will require systematic verification. They point out that although earthquakes are complex phenomena, systematic scaling laws such as the Gutenberg–Richter frequency-magnitude relation have been recognized. The Gutenberg–Richter relation is given by:  $\log N(M) = a - bM$ , where  $M$  is the earthquake magnitude,  $N(M)$  is the number of earthquakes with magnitude greater than or equal to  $M$ , and  $a$  and  $b$  are constants. Since  $b \approx 1$ , this means that the number of earthquakes increase tenfold for each decrease of one magnitude unit. This suggests that *large* earthquakes occur in regions where there are large numbers of *small* earthquakes. On this basis, the regions where large earthquakes will occur can be forecast with considerable accuracy, but the Gutenberg–Richter relation provides no information about the precise occurrence times.

### **Earthquake Engineering Considerations and Early Warning Systems**

Since seismic hazards exist in many regions of the world, three major strategies are introduced to reduce their societal impacts: (1) to avoid building in high seismic-risk areas, (2) to build structures that can withstand the effects of earthquakes, and (3) to plan for earthquake emergencies. The first strategy is not very practical because, with rapid population growth, many economically productive activities are increasingly located in high seismic-risk areas. However, by mapping active faults and by studying

past earthquakes, we may estimate the risk potential from earthquakes and plan our land use accordingly. The second strategy depends on the skills of engineers, and also requires seismologists to provide realistic estimates of the ground motions resulting from expected earthquakes. The third strategy includes attempting to predict earthquakes reliably well in advance to minimize damage and casualties, and also requires the cooperation of the entire society. Although we are far from being able to predict earthquakes reliably, earthquake early warning systems can provide critical information to reduce damage and casualties, as well as to aid rescuing and recovery efforts.

Accurate prediction of the level and variability of near-source strong-ground motions in future earthquakes is one of the key challenges facing seismologists and earthquake engineers. The increasing number of near-source recordings collected by dense strong-motion networks exemplifies the inherent complexity of near-field ground shaking, which is governed by a number of interacting physical processes. Characterizing, quantifying, and modeling ground-motion complexity requires a joint investigation of (1) the physics of earthquake rupture, (2) wave-propagation in heterogeneous media, and (3) the effects of local site conditions. In ► **Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures**, P.M. Mai discusses briefly the beginnings of strong-motion seismology and the recognition of ground-motion complexity. Using two well-recorded recent earthquakes, the author introduces the observational aspects of near-field ground shaking and describes the basic mathematical tools used in the computation of ground motion. The key elements for characterizing and modeling ground-motion complexity are also explained, supplemented by a concise overview of the underlying physical processes.

With increasing urbanization worldwide, earthquake hazards pose ever greater threats to lives, property, and livelihoods in populated areas near major active faults on land or near offshore subduction zones. Earthquake early-warning (EEW) systems can be useful tools for reducing the impact of earthquakes, provided that the populated areas are favorably located with respect to earthquake sources and their citizens are properly trained to respond to the warning messages. Under favorable conditions, an EEW system can forewarn an urban population of impending strong shaking with lead times that range from a few seconds to a few tens of seconds. A lead time is the time interval between issuing a warning and the arrival of the  $S$ - and surface waves, which are the most destructive due to their large amplitudes. Even a few seconds of advance warning is useful for pre-programmed emergency measures at various critical facilities, such as the deceler-

ation of rapid-transit vehicles and high-speed trains, the orderly shutoff of gas pipelines, the controlled shutdown of some high-technological manufacturing operations, the safe-guarding of computer facilities, and bringing elevators to a stop at the nearest floor.

Recent advances in early warning methodologies are summarized by W.H.K Lee and Y.M. Wu in the second part of their article, ► [Earthquake Monitoring and Early Warning Systems](#). In ► [Earthquake Early Warning System in Southern Italy](#), A. Zollo et al. analyze and illustrate the main scientific and technological issues related to the implementation and management of the earthquake early warning system under development in the Campania region of southern Italy. The system is designed to issue alerts to distant coastal targets using data from a dense seismic network deployed in the Apennine belt region. The authors note that earthquake early warning systems can also help mitigate the effects of earthquake-induced disasters such as fires, explosions, landslides, and tsunamis. Earthquake early warning systems can be installed at relatively low cost in developing countries, where even moderate-size earthquakes can cause damage comparable to that caused by much larger earthquakes in developed countries.

Nonlinear problems in structural earthquake engineering deal with the dynamic response of meta-stable, man-made buildings subjected to strong earthquake shaking. During earthquakes, structures constructed on soft sediments and soils deform together with the underlying soil. Strong shaking forces the soil-structure systems to evolve through different levels of nonlinear response, with continuously changing properties that depend upon the time history of excitation and on the progression and degree of damage. In ► [Earthquake Engineering, Non-linear Problems in](#), M.D. Trifunac first briefly discuss the literature on complex and chaotic dynamics of simple mechanical oscillators, and then introduces the dynamic characteristics and governing equations of the meta-stable structural dynamics in earthquake engineering. He describes the nature of the solutions of the governing equations in terms of both the vibrational and the wave representations. The author also addresses the dynamic instability, material and geometric nonlinearities, and complexities of the governing equations associated with nonlinear soil-structure interaction.

Structural health monitoring and structural damage detection refers to the processes of determining and tracking the structural integrity and assessing the nature of damage in a structure. An important and challenging problem is being able to detect the principal components of damage in structures (as they occur during or soon af-

ter the earthquake) before physical inspection. In the article, ► [Earthquake Damage: Detection and Early Warning in Man-Made Structures](#), M.I. Todorovska focuses on global methods and intermediate-scale methods, which can point to the parts of the structure that have been damaged. Recently, structural identification and health monitoring of buildings based on detecting changes in wave travel time through the structure has received renewed attention and has proven to be very promising.

### Earthquake Physics

Brittle deformation, which is the primary mode of deformation of the Earth's crust in response to tectonic stress, is manifested by faulting at the long timescale and by earthquakes at the short timescale. It is one of the best-known examples of a system exhibiting self-organized criticality. A full understanding of this system is essential for evaluating earthquake hazards, but our current understanding is sketchy. In ► [Brittle Tectonics: A Non-linear Dynamical System](#), C.H. Scholz shows that an earthquake dynamic system has two characteristic length scales,  $W^*$  and  $W^{**}$ . An earthquake nucleates within the seismogenic zone and initially propagates in all directions along its perimeter, acting as a 3D crack. When its dimension exceeds  $W^*$ , the rupture is restricted to propagating in the horizontal direction, acting as a 2D crack. Thus a symmetry breakage occurs at the dimension  $W^*$ . *Small* earthquakes, with dimensions smaller than  $W^*$ , are not self-similar with *large* earthquakes, those with lengths larger than  $W^*$ . The same occurs for suprafaults at the dimension  $W^{**}$  (a suprafault is the shear relaxation structure that includes a fault and its associated ductile shear zone).

Earthquake prediction is desirable for reducing seismic hazards, but we lack an understanding of how and why earthquakes begin and grow larger or stop. Theoretical and laboratory studies show that a quasi-static rupture growth precedes dynamic rupture. Thus, detecting the quasi-static rupture growth may lead to forecasting the subsequent dynamic rupture. In ► [Earthquake Nucleation Process](#), Y. Iio reviews studies that analyze the early portions of observed waveforms, and summarizes what we presently understand about earthquake nucleation process. An earthquake initiates over a small patch of a fault, and then their rupture fronts expand outward until they stop. Some large earthquakes have a rupture extent greater than 1000 km, while fault lengths of small microearthquakes range over only a few meters. Surprisingly, the concept that earthquakes are self-similar is widely accepted despite fault length ranging over 6 orders of magnitude. One example of such similarity is the proportionality

of average fault slip to fault length, which implies a constant static stress drop, independent of earthquake size.

The self-similarity law raises a fundamental question, namely what is the difference between large and small earthquakes? One end-member model represents earthquakes as ruptures that grow randomly and then terminate at an earlier stage for smaller earthquakes, but continue longer for larger earthquakes. This type of model has been proposed mainly to explain the frequency–magnitude distribution of earthquakes (the Gutenberg–Richter relation), but it implies that it is impossible to forecast the final size of an earthquake at the time the rupture initiates. However, the other end-member model predicts that larger earthquakes have a larger ‘seed’ than smaller earthquakes, and that large and small earthquakes are different even at their beginnings.

Geoscientists have long sought an understanding of how earthquakes interact. Can earthquakes trigger other earthquakes? The answer is clearly yes over short time and distance scales, as in the case of mainshock–aftershock sequences. Over increasing time and distance scales, however, this question is more difficult to answer. In ► [Earthquakes, Dynamic Triggering of](#), S.G. Prejean and D.P. Hill explore the most distant regions over which earthquakes can trigger other earthquakes. This subject has been the focus of extensive research over the past twenty five years, and offers a potentially important key to improving our understanding of earthquake nucleation. In this review, the authors discuss physical models and give a description of documented patterns of remote dynamic triggering.

Models of the earthquake source have been successfully used in predicting many of the general properties of seismic waves radiated from earthquakes. These general properties can be derived from a simple omega-squared spectral shape. In ► [Earthquake Scaling Laws](#), R. Madariaga derives general expressions for energy, moment and stress in terms of measured spectral parameters, and shows that earthquake sources can be reduced to a single family with the three parameters of moment, corner frequency and radiated energy. He suggests that most of the properties of the seismic spectrum and slip distribution can be explained by a simple crack model. Whether an earthquake is modeled as a simple circular crack or as a complex distribution of such cracks, the result is the same.

In ► [Earthquake Source: Asymmetry and Rotation Effects](#), R. Teisseyre presents a consistent theory describing an elastic continuum subjected to complex internal processes, considers all the possible kinds of the point-related motions and deformations, and defines a complex rotation field including spin and twist. Also included in the discus-

sion is a new description of the source processes, including the role of rotation in source dynamics, an explanation of co-action of the slip and rotation motions, and a theory of seismic rotation waves. Rotational seismology is an emerging field, and a progress report is provided in the Appendix in ► [Earthquake Monitoring and Early Warning Systems](#) by W.H.K. Lee and Y.M. Wu.

### Some New Tools to Study Earthquakes

The Global Positioning System (GPS) is a space-based Global Navigation Satellite System. Using signals transmitted by a constellation of GPS satellites, the positions of ground-based receivers can be calculated to high precision, making it possible to track relative movements of points on the Earth’s surface over time. Unlike older geodetic surveying methods (which involved periodically but infrequent measuring angles, distances, or elevations between points), GPS can provide precise 3-D positions over a range of sampling rates and on a global scale. GPS equipment is easy to use and can be set up to collect data continuously. Since its early geophysical applications in the mid-1980s, this versatile tool, which can be used to track displacements over time periods of seconds to decades, has become indispensable for crustal deformation studies, leading to many important insights and some surprising discoveries. In ► [GPS: Applications in Crustal Deformation Monitoring](#), J. Murray-Moraleta focuses on applications of GPS data to the studies of tectonic, seismic, and volcanic processes. The author presents an overview of how GPS works and how it is used to collect data for geophysical studies. The article also describes a variety of ways in which GPS data have been used to measure crustal deformation and investigate the underlying processes.

The concept of a seismic cycle involves processes associated with the accumulation and release of stress on seismogenic faults, and is commonly divided into three intervals: (1) the coseismic interval for events occurring during an earthquake, (2) the postseismic interval immediately following an earthquake, and (3) the interseismic period in between large earthquakes. In ► [Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of](#), R. Lohman explores how we can draw conclusions about fault zone slip at depths far greater than are directly accessible to us, based on how the Earth’s surface deforms during, before, and after earthquakes.

Atmospheric sound can be radiated by the displacement or rupture of the Earth’s surface induced by earthquakes, tsunamis, and volcanoes, and by the flow and excitation of fluids during volcanic eruptions. These complex and potentially cataclysmic phenomena share some



common physics, yet represent different ways of converting energy into atmospheric sound. In ► **Infrasound from Earthquakes, Tsunamis and Volcanoes**, M. Garces and A. LePichon discuss some of the signal features unique to earthquakes, tsunamis, and volcanoes captured by the present generation of infrasound arrays. They also discuss contemporary methods for the analysis, interpretation, and modeling of these diverse signals, and consider some of the associated geophysical problems that remain unsolved.

### Tsunamis

Tsunamis are oceanic gravity waves generated by seafloor deformation due to earthquakes, volcanic eruptions, landslides, or asteroid impacts. Earthquake tsunamis, such as the 2004 Indian Ocean tsunami (caused by the Sumatra–Andaman earthquake of December 26), are the most frequent type of tsunamis. However, large volcanic eruptions, such as the 1883 Krakatau eruption (in the Sunda strait between the islands of Java and Sumatra) also cause ocean-wide tsunamis. Landslides (which are often triggered by earthquakes) cause large tsunamis locally, but their effects are usually limited to the immediate vicinity of the source.

### Modeling: Forward and Inverse Approaches

Forward-modeling of a tsunami starts from given initial conditions, computes its propagation in the ocean, and calculates the tsunami arrival times and/or water run-up heights along the coasts. Once the initial conditions are provided, the propagation and coastal behavior can be numerically computed for an actual bathymetry. These calculations are useful for early tsunami warning and for detailed hazard estimations. However, the initial conditions associated with tsunami generation processes are still poorly known, because large tsunamis are rare and the tsunami generation in the open ocean is not directly observable. Currently, the tsunami source is estimated indirectly, mostly on the basis of seismological analysis, but a more direct estimation of the tsunami source is essential to better understand the tsunami generation process and to more accurately forecast the effects of a tsunami along the coasts.

In ► **Tsunamis, Inverse Problem of**, K. Satake reviews inverse methods used in the quantification of tsunami sources from the observations. The author describes the tsunami generation by earthquakes, with an emphasis on the fault parameters and their effects on tsunami propagation, including shallow water theory and numerical computation. The author then summarizes the tsunami observations, including instrumental sea-level data and run-

up height estimates for modern, historical and prehistoric tsunamis. He also describes methods for modeling and quantifying a tsunami source, and for analyzing tsunami travel times, amplitudes and waveforms. He concludes with an estimation of earthquake fault parameters derived from waveform inversion of tsunami data, and a discussion of heterogeneous fault motion and its application for tsunami warning.

Tsunami inundation is the one of the final stages of tsunami evolution, when the wave encroaches upon and floods dry land. It is during this stage that a tsunami is most destructive and takes the vast majority of its victims. To gauge the near-shore impact of tsunami inundation, engineers and scientists rely primarily on three different methods: (1) field survey of past events, (2) physical experimentation in a laboratory, and (3) numerical modeling. In ► **Tsunami Inundation, Modeling of**, P.J. Lynett focuses on numerical simulations. He reviews tsunami generation and open ocean propagation, and discusses the physics of near-shore tsunami evolution, hydrodynamic modeling of tsunami evolution, moving shoreline algorithms, and effect of topographical features on inundation.

### Tsunami Forecasting and Warning

The original definition of ‘tsunami earthquake’ was given by H. Kanamori (Phys Earth Planet Inter 6:346–359, 1972) as “an earthquake that produces a large-size tsunami relative to the value of its surface wave magnitude ( $M_S$ )”. The true damage potential that a tsunami earthquake represents may not be recognized by conventional near real-time seismic analysis methods that utilize measurements of relatively high-frequency signals, and thus the threat may only become apparent upon the arrival of the tsunami waves on the local shores. Although tsunami earthquakes occur relatively infrequently, the effect on the local population can be devastating, as was most recently illustrated by the July 2006 Java tsunami earthquake, which was quickly followed by tsunami waves two to seven meters high, traveling as far as two kilometers inland and killing at least 668 people.

It is important to note that the definition of ‘tsunami earthquake’ is distinct from that of ‘tsunamigenic earthquake’. A tsunamigenic earthquake is any earthquake that excites a tsunami. Tsunami earthquakes are a specific subset of tsunamigenic earthquakes. In ► **Tsunami Earthquakes**, J. Polet and H. Kanamori describe the characteristics of tsunami earthquakes and the possible factors involved in the anomalously strong excitation of tsunamis by these events. They also discuss a possible model for these infrequent, but potentially very damaging events.

Tsunamis are among nature's most destructive hazards. Typically generated by large, underwater shallow earthquakes, tsunamis can cross an ocean basin in a matter of hours. Although difficult to detect, and not dangerous while propagating in open ocean, tsunamis can unleash awesome destructive power when they reach coastal areas. With advance warning, populations dwelling in coastal areas can be alerted to evacuate to higher ground and away from the coast, thus saving many lives.

Tsunami travels at about the same speed of a commercial airliner, however, seismic waves can travel at speeds more than 40 times greater. Because of this large disparity in speed, scientists rely on seismic methods to detect the possibility of tsunami generation and to warn coastal populations of an approaching tsunami well in advance of its arrival. The seismic P-wave for example, travels from Alaska to Hawaii in about 7 min, whereas a tsunami will take about 5.5 hours to travel the same distance. Although over 200 sea-level stations reporting in near-real time are operating in the Pacific Ocean, it may take an hour or more, depending on the location of the epicenter, before the existence (or not) of an actual tsunami generation is confirmed. In other ocean basins where the density of sea-level instruments reporting data in near real-time is less, the delay in tsunami detection is correspondingly longer. However, global, regional, and local seismic networks, and the infrastructure needed to process the large amounts of seismic data that they record, are well in place around the world. For these reasons, tsunami warning centers provide initial tsunami warnings to coastal populations based entirely on the occurrence of a large shallow offshore earthquake. It is well-known, however, that large shallow offshore earthquakes may or may not be tsunamigenic.

In ► **Tsunami Forecasting and Warning**, O. Kami-gaichi discusses the complexity problem in tsunami forecasting for large local events, and describes the Tsunami Early Warning System in Japan. Tsunami disaster mitigation can be achieved effectively by the appropriate combination of software and hardware countermeasures. Important issues for disaster mitigation includes: (1) improving people's awareness of the tsunami hazards, (2) imparting the necessity of spontaneous evacuation when people notice an imminent threat of tsunami on their own (feeling strong shaking near the coast, seeing abnormal sea level change, etc), (3) giving clear directions on how to respond to the tsunami forecast, and (4) conducting tsunami evacuation drills. The author notes that in tsunami forecasting, a trade-off exists between promptness and accuracy/reliability.

In ► **Earthquake Source Parameters, Rapid Estimates for Tsunami Warning**, B. Hirshorn and S. Weinstein de-

scribe the basic method used by the Pacific Tsunami Warning Center (PTWC) mainly for large teleseismic events. Software running at the PTWC processes in real time seismic signals from over 150 seismic stations worldwide provided by various seismic networks. Automatic seismic event detection algorithms page the duty scientists for any earthquake occurring worldwide over about Magnitude 5.5. Other automatic software locates these events, and provides a first estimate of their magnitude and other source parameters in near real time. Duty scientists then refine the software's automated source parameter estimates and issue a warning if necessary. The authors also describe their ongoing efforts to improve estimates of earthquake source parameters.

### **Wedge Mechanics, Submarine Landslides and Slow Earthquakes**

A study of the mechanics of wedge-shaped geological bodies, such as accretionary prisms in subduction zones and fold-and-thrust belts in collision zones, is interesting because they enable us to use the observed morphology and deformation of these bodies to constrain properties of the thrust faults underlying them. The fundamental process described in wedge mechanics is how gravitational force, in the presence of a sloping surface, is balanced by basal stress and internal stress. The internal state of stress depends on the rheology of the wedge. The most commonly assumed wedge rheology for geological problems is perfect Coulomb plasticity, and the model based on this rheology is referred to as the Coulomb wedge model.

The connection between wedge mechanics and great earthquakes and tsunamis at subduction zones is an emerging new field of study. In their article, ► **Wedge Mechanics: Relation with Subduction Zone Earthquakes and Tsunamis**, Wang et al. cover the topics of stable and critical Coulomb wedges, dynamic Coulomb wedge, stress drop and increase in a subduction earthquake, and tsunamigenic coseismic seafloor deformation. Better constraints are needed to quantify how stresses along different down-dip segments of the subduction fault evolve with time throughout an earthquake cycle and how the evolution impacts wedge and seafloor deformation. Submarine monitoring in conjunction with land-based monitoring at subduction zones that are currently in different phases of earthquake cycles will allow us to better understand the evolution of fault and wedge stresses during the inter-seismic period. In this regard, cabled seafloor monitoring networks including borehole observatories being designed or implemented at different subduction zones will surely yield valuable data in the near future.

The term ‘submarine landslide’ encompasses a multitude of gravitational mass failure features at areal scales from square meters to thousands of square kilometers. The term ‘slow earthquake’ describes a discrete slip event that produces millimeter to meter-scale displacements identical to those produced during earthquakes but without the associated seismic shaking. Recently, a GPS network on the south flank of Kilauea volcano, Hawaii, recorded multiple slow earthquakes on the subaerial portion of a large landslide system that extends primarily into the submarine environment. Since catastrophic failure of submarine landslides can cause a tsunami they represent significant hazards to coastal zones. Because submarine landslide systems are among the most active as well as spatially confined deforming areas on Earth, they are excellent targets for understanding the general fault failure process. In ► **Submarine Landslides and Slow Earthquakes: Monitoring Motion with GPS and Seafloor Geodesy**, B.A. Brooks et al. present a review of this interdisciplinary topic of interest in geodesy, seismology, tsunamis, and volcanology.

## Volcanoes

About 1,500 volcanoes have erupted one or more times during the past 10,000 years, and since A.D. 1600, volcanic disasters have killed about 300,000 people and resulted in property damage and economic loss exceeding hundreds of millions of dollar. Articles in this section are intended to summarize recent research in: (1) volcano seismology, (2) physical processes involved in volcanoes, and (3) modeling volcanic eruptions and hazards warning.

## Volcano Seismology

Magma transport in a volcano is episodic due to the inherent instability of magmatic systems at all time scales. This episodicity is reflected in seismic activity, which originates in dynamic interactions between gas, liquid and solid along magma transport paths that involve complex geometries. The description of the flow processes is governed by the nonlinear equations of fluid dynamics. In volcanic fluids, further complexity arises from the strong nonlinear dependence of magma rheology on temperature, pressure, and water and crystal content, and nonlinear characteristics of associated processes underlying the physico-chemical evolution of liquid-gas mixtures constituting magma.

In ► **Volcanoes, Non-linear Processes in**, B. Chouet presents a brief review of volcano seismology and addresses basic issues in the quantitative interpretation of processes in active volcanic systems. Starting with an introduction of the seismic methodology used to quantify

the source of volcano seismicity, the author then focuses on sources originating in the dynamics of volcanic fluids. A review of some of the representative source mechanisms of Long-Period (LP) and Very Long-Period (VLP) signals is followed by a description of a mesoscale computational approach for simulating two-phase flows of complex magmatic fluids. Refined understanding of magma and hydrothermal transport dynamics therefore requires multidisciplinary research involving detailed field measurements, laboratory experiments, and numerical modeling. Such research is fundamental to monitoring and interpreting the subsurface migration of magma that often leads to eruptions, and thus would enhance our ability to forecast hazardous volcanic activity.

Volcano seismicity produces a wide variety of seismic signals that provide glimpses of the internal dynamics of volcanic systems. Quantitative approaches to analysis and interpret volcano-seismic signals have been developed since the late 1970s. The availability of seismic equipments with wide frequency and dynamic ranges since the early 1990s has revealed a variety of volcano-seismic signals over a wide range of periods. Quantification of the sources of volcano-seismic signals is crucial to achieving a better understanding of the physical states and dynamics of magmatic and hydrothermal systems. In ► **Volcano Seismic Signals, Source Quantification of**, H. Kumagai provides the theoretical basis for a quantification of the sources of volcano-seismic signals. The author focuses on the phenomenological representation of seismic sources, waveform inversion to estimate source mechanisms, spectral analysis based on an autoregressive model, and physical properties of fluid-solid coupled waves.

Among various eruptive styles, Strombolian activity is easier to study because of its repetitive behavior. Since Strombolian activity offers numerous interesting seismic signals, a growing attention has been devoted to the application of waveform inversion for imaging conduit geometry and retrieving eruption dynamics from seismological recordings. Quantitative models fitting seismological observations are a powerful tool for interpreting seismic recordings from active volcanoes. In ► **Slug Flow: Modeling in a Conduit and Associated Elastic Radiation**, L. D’Auria and M. Martini discuss the mechanism of generation of Very-Long Period (VLP) signals accompanying Strombolian explosions. This eruptive style, occurring at many basaltic volcanoes worldwide, is characterized by the ascent and the bursting of large gas slugs. The mechanism of formation, ascent and explosion of bubbles and slugs and their relation with eruptive activity has been studied theoretically and by analogue simulations. The authors report results from numerical simulations, focusing on the

seismic signals generated by pressure variations applied to the conduit walls.

### Physical Processes in Volcanoes

The dynamics of solid-liquid composite systems are relevant to many problems, including how melts or aqueous fluids migrate through the mantle and crust toward the surface, how deformation and fracture in these regions are influenced by the existence of fluids, and also how these fluids can be observed in seismic tomographic images. In ► **Earth's Crust and Upper Mantle, Dynamics of Solid-Liquid Systems** in, Y. Takei introduces a general continuum mechanical theory for macroscopic dynamics of solid-liquid composite systems, and emphasizes on how such interactions with pore geometry can be studied. The author then discusses the determinability of porosity and pore geometry from seismic tomographic images, and presents a practical method to assess porosity and pore geometry from tomographic  $V_p$  and  $V_s$  images.

A volcano consists of solids, liquids, gases, and intermediate materials of any two of these phases. Mechanical and thermodynamical interactions between these phases are essential in the generating a variety of volcanic activities. In particular, the gas phase is mechanically distinct from the other phases and plays an important role in dynamic phenomena in volcanoes. In ► **Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation**, M. Ichihara and T. Nishimura discuss several bubble dynamics phenomena from the viewpoint that a bubbly fluid acts as an impulse generator of observable signals, such as earthquakes, ground deformations, airwaves, and an eruption itself. The authors focus on the notion that the impulse is excited by non-linear coupling between internal processes in a bubbly fluid and an external perturbation. The importance of these processes has recently become noticed as a possible triggering mechanism of eruptions, earthquakes, and inflation of a volcano.

Our capability to mitigate volcano hazards relies in large part on forecasting explosive events, a process which requires a high degree of understanding of the physicochemical factors operating during explosive volcanism. The approaches taken to gain an understanding of explosive volcanism have relied on a combination of field observations, theoretical models and laboratory models of materials and mechanisms. In ► **Volcanic Eruptions, Explosive: Experimental Insights**, S.J. Lane and M.R. James first review aspects of the volcanic materials literature, with the aim of illustrating the nature of molten rock, the complexity of which underpins most explosive volcanic processes. Experimental modeling of these processes can then

build on the materials understanding. Such experiments involve investigation of the behavior of natural volcanic products at laboratory time and length scales, including the response of magma samples to rapid changes in pressure and temperature, the fall behavior of silicate particles in the atmosphere, and the generation and separation of electrostatic charge during explosive eruptions.

In ► **Volcanic Eruptions: Cyclicity During Lava Dome Growth**, O. Melnik et al. consider the process of slow extrusion of very viscous magma that forms lava domes. Dome-building eruptions are commonly associated with hazardous phenomena, including pyroclastic flows generated by dome collapses, explosive eruptions, and volcanic blasts. These eruptions commonly display fairly regular alternations between periods of high and low or no activity with time scales from hours to years. Usually hazardous phenomena are associated with periods of high magma discharge rate. Hence, understanding the causes of pulse activity during extrusive eruptions is an important step towards forecasting volcanic behavior, and especially the transition to explosive activity when magma discharge rate increases by a few orders of magnitude. In recent years the risks have escalated because the population density in the vicinity of many active volcanoes has increased.

### Modeling Volcanic Eruptions and Hazards Warning

While a wide range of complex deterministic models exists to model various volcanic processes, these provide little in the way of information about future activity. Being the (partially) observed realization of a complex system, volcanological data are inherently stochastic in nature, and need to be modeled using statistical models. In ► **Volcanic Eruptions: Stochastic Models of Occurrence Patterns**, M.S. Bebbington considers models of eruption occurrence, omitting techniques for forecasting the nature and effect of the eruption. As the track record of a potentially active volcano provides the best method of assessing its future long-term hazards, the author first briefly reviews the provenance and characteristics of the data available, and then discusses various taxonomies for stochastic models. The examples of Mount Etna and Yucca Mountain are selected for more detailed examination partly because many, somewhat contradictory, results exist. Different models make different assumptions, and vary in how much information they can extract from data. In addition, the data used often varies from study to study, and the sensitivity of models to data is important, but too often ignored.

In ► **Volcanic Hazards and Early Warning**, R.I. Tilling highlights the range in possible outcomes of volcano un-

rest and reviews some recent examples of the actual outcomes documented for several well-monitored volcanoes. The author also discusses the challenge for emergency-management authorities, as well as challenges in achieving refined predictive capability. To respond effectively to a developing volcanic crisis, timely and reliable early warnings are absolutely essential; they can be achieved only by a greatly improved capability for eruption prediction. This in turn depends on the quantity and quality of volcano-monitoring data and the diagnostic interpretation of such information.

### Discussions

The terms ‘Prediction’ and ‘forecasting’ are often used interchangeably. However, the commonly accepted definition of an ‘earthquake prediction’ is a concise statement, in advance of the event, of the time, location, and magnitude of a future earthquake. To be practically useful to the society, the time window must be short (in days or months), the location extent small (within tens of kilometers), and the magnitude precise ( $\pm 0.5$  unit). A ‘forecast’, on the other hand, is more loosely defined as the probability of an occurrence of a large earthquake in a given region (e.g., southern California) during the coming decades or centuries. The broader time intervals associated with forecasts allows society to consider and implement mitigation efforts over a large region.

As an observational seismologist and on a personal note, I am skeptical that earthquakes can be reliably predicted before (1) we have collected accurate data of their occurrences over a sufficiently long period of time, and (2) we have a good understanding of the physical processes that create them. Although earthquakes have been

known from antiquity, accurate earthquake catalogs exist only since about the 1960s. Since the recurrence of a damaging earthquake in a given area is often more than 100 years (some even thousands of years), it is obvious that we lack the necessary observed data. C. Lanczos (*Linear Differential Operators*, Van Nostrand-Reinhold, 1961) said it well in general: “a lack of information cannot be remedied by any mathematical trickery”. Nevertheless, we must apply new concepts and tools to extract as much useful information as possible from the existing data. Indeed, we must thank many pioneers for enlightening us with many interesting and tentative results about earthquakes, tsunamis and volcanoes that they managed to extract from inadequate and insufficient data.

Because most tsunamis are generated by earthquakes, successes in predicting tsunamis depend on predicting earthquakes and recognizing them as tsunamigenic. Predicting a volcanic eruption is a little easier, as the location is known and there are often some observable phenomena preceding it. However, the exact time and the intensity and extent of an eruption are difficult to predict because the volcanic processes are very complex involving gas, liquid and solid phases.

Fortunately, earthquake, tsunami and volcano hazards can be reduced by employing sound engineering practices, early warning systems, and hazard analysis, using many of the tools and concepts that were developed for prediction. Since fatalities and economic loss from a single catastrophic event can reach 100,000 or more, and \$100 billion or more, respectively, it is imperative that governments should support long-term monitoring with modern instruments and research, including complexity studies of earthquakes, tsunamis and volcanoes.

# Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of

ROWENA LOHMAN  
Cornell University, Ithaca, USA

## Article Outline

Glossary  
Definition of the Subject  
Introduction  
Highlights of Earthquake Geodesy  
Modeling of Geodetic Observations  
Future Directions  
Bibliography

## Glossary

**Aseismic** Occurring without detectable radiated seismic energy.

**Cascadia** The region of the Pacific Northwest dominated by the Cascade Range and affected by subduction of the Juan de Fuca plate beneath North America.

**Coseismic** Occurring during an earthquake.

**Elastic** A form of behavior of a solid when subjected to stress. Elastic solids deform in response to stress by an amount proportional to a constant known as the “rigidity”. When any applied stress is removed, an elastic solid recovers its original shape.

**Geodesy** The study of the shape and area of the Earth, including large-scale variations that affect the rotation dynamics of the planet of the whole, down to smaller length scales of earthquakes, landslides, etc.

**Forward model** A description of what a model of some process would predict about behavior of the system, e. g., how a given distribution of subsurface slip on a fault during an earthquake should affect observations of ground deformation at the surface.

**GPS** Global Positioning System. A network of satellites that transmit a signal that can be used by receivers (small transportable and/or permanent affixed to the ground) to infer three-dimensional positions.

**InSAR** Interferometric Synthetic Aperture Radar. The combination of Synthetic Aperture Radar imagery (generally acquired from airborne or satellite-based platforms) to infer changes in ground deformation, digital elevation models, variations in atmospheric water vapor, etc.

**Interseismic deformation** Occurs in the time period between earthquakes, usually associated with gradual in-

crease in elastic stress to be released in future earthquakes.

**Inverse theory** The approach to determining the values for parameters of a given physical model that best describe observations of the system of interest.

**Leveling** The field of geodesy involved in the determination of variations in angle from horizontal between nearby fixed points on the Earth’s surface, usually converted to changes in elevations.

**Locked zone** The portion of the fault zone that does not slip during the interseismic period, therefore accumulating stress and eventually rupturing coseismically.

**Paleoseismology** The study of individual earthquakes that occurred in the past, usually before the advent of instrumental recordings of seismic events.

**Plate tectonics** The theory governing how discrete plates on the Earth’s surface move relative to each other over geologic time.

**Postseismic deformation** Deformation occurring in the hours to years following an earthquake.

**Seismic cycle** The combination of strain build-up and release that occurs on plate margins and along faults within plates, accommodated by processes within the coseismic, postseismic and interseismic time scales.

**Seismogenic** The region of a fault zone that is capable of producing earthquakes. Also refers to effects caused by an earthquake.

**Subduction** The process by which one tectonic plate descends beneath another, usually accompanied by volcanism and seismicity.

**Triangulation** The field of geodesy related to measuring horizontal angles and changes in angles between networks of fixed points.

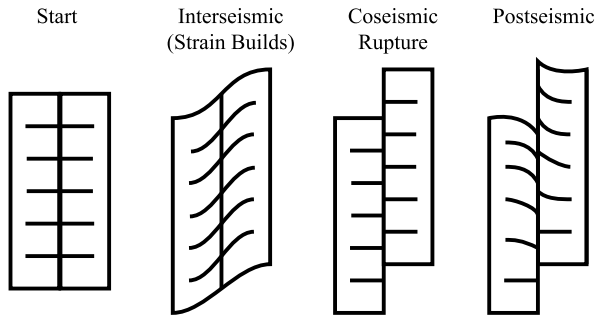
**Trilateration** The field of geodesy related to measuring distance and changes in distance between networks of fixed points.

**Viscoelastic** A material behavior that is a combination of viscous and elastic behavior, resulting in some permanent deformation when the material is subjected to changes in stress.

**Viscosity** A material property describing its ability to flow in response to an applied stress. A measure of the response of a material to a stress, resulting in permanent deformation. The deformation rate of a viscous material depends on both the viscosity and applied stress.

## Definition of the Subject

The seismic cycle consists of the processes associated with the accumulation and release of stress on seismogenic faults. The cycle is commonly divided into 3 periods: the



Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 1

**Cartoon illustrating map view (view from overhead) of ground deformation during the seismic cycle for a strike-slip fault (after [73]). Similar models exist for other fault types**

coseismic interval for events occurring during an earthquake, the postseismic interval after an earthquake, and the interseismic period in between earthquakes (Fig. 1). Some of this deformation during these different periods is related directly to the motion on the fault during the earthquake – the ground is translated in one direction or another, there is crushing of rock, rotation or heaving of blocks of earth, and landslides triggered by the shaking. There is also ground deformation that results because of secondary effects – movement of ground water within the crushed and strained rock, cascades of earthquakes triggered by stress changes during the first event, continued slip on the fault interface, as well as flow of deeper, more ductile layers of the crust and mantle in response to the changes in stress. By examining these different behaviors and studying crustal deformation we can learn more about the underlying cause.

Earthquakes and other motion along fault zones are some of the ways that the Earth's crust accommodates far-field forcings due to plate tectonic motions. One of the prime features of interest as we study the seismic cycle is the magnitude of motion along these fault zones at different temporal and length scales. However, the only place where we can directly observe this motion is within the shallowest (< several meters) parts of the fault zone, through maps of features that are offset across the fault. The field of paleoseismology involves the search for information about previous earthquakes on a fault zone, often through trenches dug across the fault and the use of radiogenic dating to determine how frequently earthquakes have occurred in the past.

In order to infer what is occurring throughout the whole seismogenic zone (often the upper ~15 km within the continental crust), we rely on an arsenal of tools that

include seismology (the study of how seismic waves travel through the earth) and geodesy (the study of changes to the shape of the earth's surface). In this chapter, we will explore how we can draw conclusions about fault zone slip at depths far greater than are directly accessible to us, based on how the earth's surface deforms during, before and after earthquakes.

## Introduction

Ground shaking due to earthquakes can be felt over most of the Earth's surface, both on land and underwater. Earthquakes are also associated with volcanic activity, landslides and tsunami – they are often concentrated near discrete tectonic plate boundaries but also occupy diffuse zones where the Earth's crust is deforming [12,57]. The seismic energy that is released during an earthquake passes through deep portions of the Earth and helps us to learn about material properties we could not illuminate any other way, including details of the structure of the Earth's crust, core and everything in between. However, the destructive cost of large earthquake requires that the primary goal of earthquake research is that of determining when and where damaging earthquakes will occur [99]. To achieve this goal, we first have to understand what happens during each earthquake, i. e., where it was located, how big it was, how it's location relates to the distribution of previous earthquakes, etc.

One of the most common methods that we use to study earthquakes relies on the feature most apparent to humans – the rapid and often destructive movements of the ground that occur during the earthquake. Seismology studies how the ground shakes during earthquakes and how we can use that information to better understand seismogenesis. However, this ground shaking is also always accompanied by some amount of permanent deformation of the ground. After all, the primary driving force behind most earthquakes is the slow motion of tectonic plates relative to each other. In this chapter we will not cover seismology, although combinations of seismic data and observations of geodetic displacements are often very powerful tools. Instead, we cover the geodetic observations of ground deformation during earthquakes and some of the methods that we use to interpret this deformation in terms of what actually happened on and around the fault zone.

## Highlights of Earthquake Geodesy

Geodesy is defined as the branch of mathematics concerned with the shape and area of the Earth. In this chapter we will examine how the use of geodesy to quantify changes in the shape of the Earth can help us learn about

earthquakes. We begin with a brief history of how the use of geodesy to study earthquakes has evolved, along with a description of some of the more interesting individual earthquakes and other seismic cycle behaviors that have been observed geodetically. We then explain the mechanics of how we use these measurements of surface deformation to understand processes deep beneath the Earth's surface.

### Observations of Coseismic Displacements

The history of earthquake geodesy begins even before we really understood what earthquakes were. The key observations began when scientists began to associate earthquakes with observable deformation of the terrain. In Charles Lyell's *Principles of Geology* [46], he noted that earthquakes often accompany abrupt changes in the ground surface. The 1819 Ran of Cutch, India [64] and 1855 Wairarapa, New Zealand earthquakes were some of the first events where the accompanying ground deformation was observed. In 1835, during Charles Darwin's voyage on the *H.M.S. Beagle*, he experienced a large earthquake near Concepcion, Chile. During his reconnaissance of the area, he noted that the coastline had risen several meters in areas, exposing barnacles that had previously been underwater. He also found fossils hundreds of meters above sea level, indicating that numerous earthquakes had raised the cliffs over many millennia.

Several large earthquakes in the second half of the 19th century helped to advance the theory that earthquakes were caused by motion on faults. The 1872 Owens Valley, California, earthquake [19], the 1888 Amuri/Marlborough, New Zealand, earthquake [50], the 1891 Nobi, Japan, earthquake [36] and the 1893 Baluchistan earthquake [21] were each accompanied by visible deformation of the ground surface. These observations, and the fact that much of the deformation was consistent with regional, long-term topographic relief, helped counter arguments that earthquakes were primarily caused by volcanic activity [48].

Around the same time interval, the practice of surveying was coming into more widespread use, notably in India where Britain was involved in mapping the areas under its control. Surveyors installed permanent "monuments" or benchmarks, which they could then return to and re-survey at a later date. The May 17, 1892 Tapanuli, Sumatra, earthquake occurred during a triangulation survey by JJA Muller [58]. They noted changes in their surveyed angles between benchmarks that were consistent with two meters of deformation on a branch of the great Sumatran fault. In the foothills of the Himalaya, triangulation/leveling surveys measured deformation due to the

1897 Assam earthquake [64] and the 1905 Kangra earthquake [53]. In Italy, the 1915 Avezzano earthquake [94] was also spanned by early geodetic surveys.

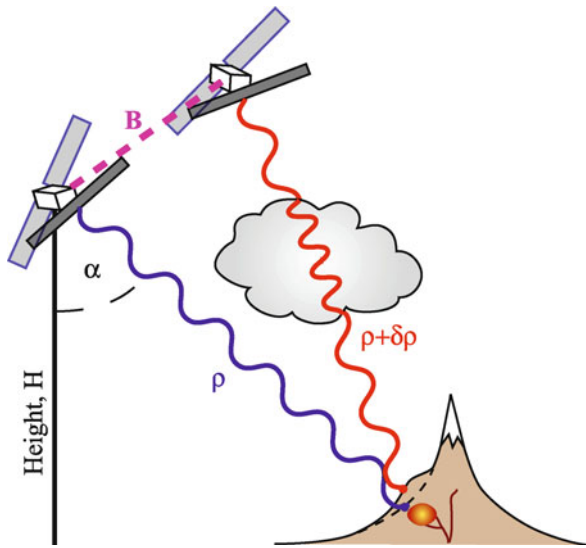
In North America, the first great earthquake to be surveyed was the 1906 earthquake that destroyed most of the city of San Francisco, CA [40]. HF Reid [72] used three sets of triangulation surveys across San Andreas Fault to show that there had been approximately 3.2 meters of slip across the fault. During the 1920s and 1930s, there was a great deal of leveling work done in Japan which captured deformation associated with the 1927 Tango earthquake, and the 1944 and 1946 earthquakes associated with oceanic plate subduction in southwestern Japan [63,93]. Tide gauges along the coasts have also proved useful in earthquake studies, especially near subduction zones.

There was a boom in the use of the three main ground-based geodetic techniques (triangulation, trilateration and leveling) in the middle of the twentieth century, with observations of the 1940 Imperial Valley and 1962 Tehachapi earthquakes in California and the first observation of shallow interseismic creep on faults in the San Francisco Bay Area and in the Salton Trough/Imperial Valley region in Southern California. Next, the development of space-based geodesy began to allow for more precise and spatially extensive surveys of areas before and after earthquakes. Very Long Baseline Interferometry (VLBI) stations scattered around the world placed strong constraints on the relative motion of individual tectonic plates, as did the widespread use of Global Positioning System (GPS) [22,83].

GPS can be used either in continuous mode at permanent stations, which allows for high precision observations that can catch temporal changes in deformation, or in survey/campaign mode, where researchers take GPS receivers out to fixed benchmark stations in the field (sometimes previously studied using earlier surveying methods). Some of the first large earthquakes to be studied using GPS were the Ms 6.6 Superstition Hills, CA, earthquake in 1987 [39] and the 1989 Mw 7.1 Loma Prieta, CA earthquake [3]. The precise locations and descriptions of ground deformation allowed researchers to begin to examine how individual earthquakes fit in with the long-term topography, including the estimation of possible recurrence intervals [35,41].

**InSAR** The advent of Interferometric Synthetic Aperture Radar (InSAR), and its application to the 1992 Landers, California, earthquake [17,49], ushered in a new era of geodesy with its unparalleled spatial density of deformation observations [23,76]. Synthetic Aperture Radar (SAR) images are acquired from a variety of platforms that send



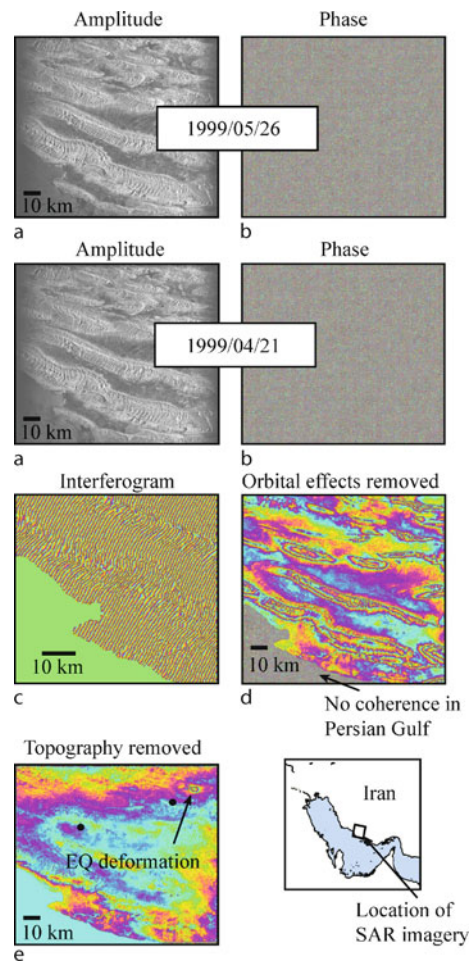


Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 2

**Formation of a Synthetic Aperture Radar (SAR) interferogram:** Two satellite radar images of the ground surface, separated by a spatial baseline,  $B$ , are combined to solve for the  $\delta\rho$ , the difference in line length ( $\rho$ ) between the satellite and the ground along the satellite viewing direction ( $\alpha$ ). Knowledge of the satellite orbital path and the relative distance between the two image acquisitions ( $B$ ) is necessary to convert  $\delta\rho$  to elevation. If the two images are obtained simultaneously, or over a small time interval, the interferogram will only reflect topographic relief throughout the imaged area. However, if changes in ground surface elevation (e.g., the volcanic inflation shown in this figure between the *blue* and *red*), or changes in the atmosphere/ionosphere (e.g., water vapor content) occur between the two image acquisitions, this will also be reflected in the final interferogram

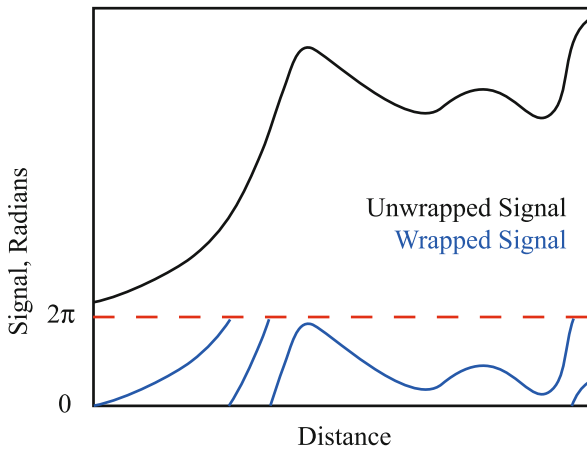
out radar signals, including satellites and airplanes (Fig. 2), and contain information about the phase and amplitude of the reflected return from the earth's surface.

A SAR interferogram is the difference in phase between two SAR images, and reflects a variety of factors that change the path length between the satellite and the ground. After correcting for some of these factors, such as topography (Fig. 3), the resulting interferometric phase is a function of any ground deformation that occurred between the two SAR acquisitions, in addition to variations in atmospheric water vapor, the ionosphere, etc. [11,20,100]. The high spatial resolution of InSAR (pixels commonly  $5\text{ m} \times 20\text{ m}$  or smaller) combined with its large areal coverage ( $\sim 100\text{ km}$ ) produces images with so many pixels that they can become unwieldy to deal with computationally, especially when multiple interferograms are studied at once. Most researchers use various down-



Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 3

**Overview of steps in forming and interpreting an interferogram.** An interferogram requires the combination of two sets of amplitude *a* and phase *b* observations from two separate image acquisitions (here, on two different dates separated by a month over an area along the Persian Gulf, Iran). While phase in the individual SAR images appear to be random noise, when the two images are combined, the phase changes vary coherently to form the interferogram in *c*. If we remove the effects of the satellite orbital geometry (the "curved earth effect"), we are left with *d*, which reflects both topography, ground surface deformation and any atmospheric changes between the two image acquisitions. The area in the *lower left* still looks like white noise due to the fact that water (here, the Persian Gulf) changes its reflectivity over short time scales and the phase does not remain coherent. If we remove the effects of topography *e* we are left with a map that clearly shows the effects of a small  $M_w 5$  earthquake that occurred during this time interval (*black dots* show the seismically-determined locations for this event). Other features are due primarily to changes in atmospheric water vapor



Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 4

Illustration of how the inherent  $2\pi$  ambiguity affects the observed deformation. The “true” ground deformation vs. distance profile (black signal) would appear as the segmented blue curve (“wrapped” signal) when viewed using InSAR, since interferometry only retains information about the relative phase within a  $2\pi$  cycle, not about the absolute phase, or number of cycles between the satellite and the ground. Reconstructing the original deformation field requires a process known as phase unwrapping, and there will also be an ambiguity as to the absolute value of the deformation field as a whole

sampling methods to reduce the number of data points considered without significantly reducing the amount of information retained [43].

The interferometric phase is only sensitive to ground deformation towards or away from the satellite, in a direction known as the satellite line of sight (LOS). Also, since the interferometric phase can only be measured as a fraction of the radar wavelength, an interferogram does not immediately give us an absolute measurement of the magnitude of ground deformation (Fig. 4). We need to add up the interferometric “fringes” to solve for the total deformation field. This “unwrapped” deformation field, along with information about the line-of-sight direction, can be combined in attempts to model the earthquake source parameters, or for models of deformation throughout other parts of the seismic cycle.

InSAR observations have improved the spatial complexity of these observations, allowing over a dozen large earthquakes to be studied in detail by a wide range of observers. Some noteworthy events that had various portions of their coseismic-postseismic activity covered by combinations of InSAR and GPS include the 1992 Landers, CA, earthquake [13,17,49], the 1995 Antofagasta, Chile, earthquake [70,77], the 1997 Hector Mine, CA earthquake [85],

the 1999 Chi-Chi, Taiwan earthquake [8,98], the 2003 Bam, Iran earthquake [18], and the 2005 Nias–Simeulue earthquake [28]. In other cases, InSAR can help to determine the location of small- to moderate-sized earthquakes in areas with little other geophysical or field-based information [44].

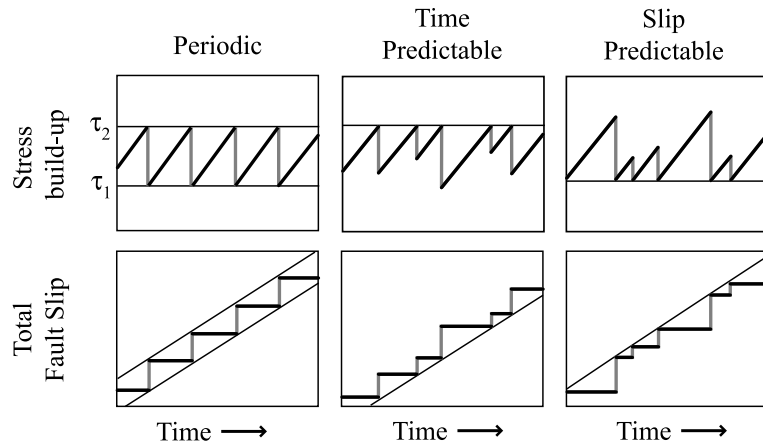
## Observations Throughout the Seismic Cycle

**Early Models** As the body of information about earthquakes grew, scientists were able to begin hypothesizing about the processes controlling their occurrence. Reid’s observations of the 1906 San Francisco earthquake led him to propose his elastic rebound theory [73], where he hypothesized that the crust behaves like an elastic solid driven from the far field at a constant rate, which ruptures in earthquakes at periodic intervals to allow the two blocks to slide past each other (Fig. 1). In the interseismic period, the ground deforms smoothly in a manner that depends on the relative plate velocities, the thickness of the elastic plate, the elastic plate rigidity, etc. The size of the largest potential earthquake on a fault would depend on the length and depth of the elastic zone, and the timing until the next earthquake would depend on how much strain had built up since the last one.

This simple model, where each earthquake releases all of the built-up stress along the fault zone at a regular interval, does not seem to hold in the real world. On most faults, the magnitude of fault slip and rupture area appears to vary significantly between earthquakes (Fig. 5). Additionally, large amounts of accelerated deformation and other types of postseismic behavior are observed to occur immediately after earthquakes, indicating that stress release is not accommodated in a simple manner. Observations such as these can help us improve and expand on the models that we utilize to explain earthquake occurrence.

**Postseismic Behavior** Earthquakes may primarily release built-up strain due to plate motions, but they also produce stress increases in both the near-field within the crust and within the underlying mantle. Observations of postseismic deformation, which is driven by the preceding coseismic stress changes [25,79,92], span a wide range of behaviors that may be explained by equally large range of constitutive properties. For some shallow strike-slip earthquakes, the observed postseismic deformation is as large as the fault slip during the earthquake [38,86].

Observed postseismic behaviors include poroelastic deformation [13,33], where fluid flow following gradients in coseismic stress changes results in ground deformation, frictional afterslip [5,28,55,81,82] and viscoelastic relax-



Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 5

Cartoon illustrating theories governing the temporal distribution of large earthquakes [88], each assuming a constant increase in stress vs. time in the interseismic interval (*black lines*), and coseismic behavior (*gray*) that depends on the accumulated stress in various ways. In the periodic model [73], earthquakes occur at constant intervals and always have the same magnitude, releasing the same amount of built-up stress. In the time-predictable model, earthquakes always occur when a maximum stress,  $\tau_2$ , is reached, but the slip in each event varies. The time of the next earthquake depends on how much stress was released in the previous event. In the slip-predictable model, earthquakes occur at varying times, but always release the amount of built-up stress down to the minimum level,  $\tau_1$ . The amount of slip, therefore, depends on the time since the previous earthquake

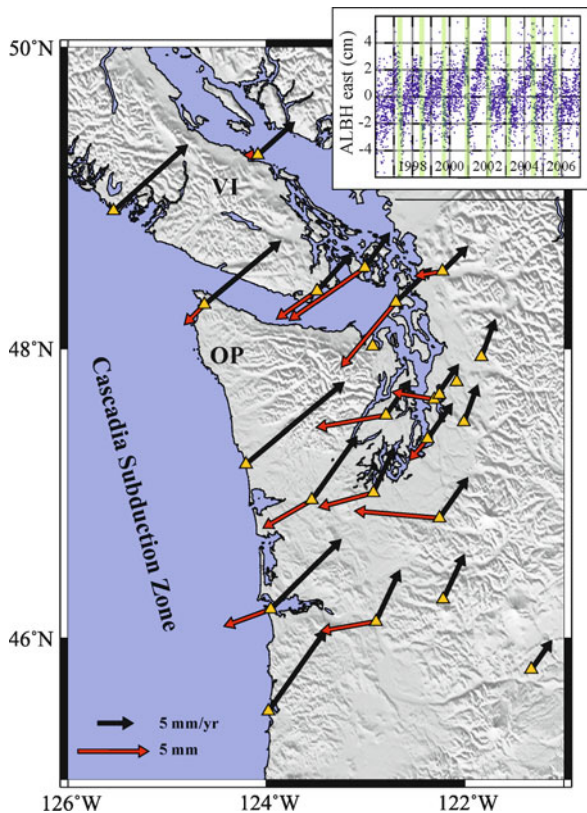
ation of the lower crust [25,26,68,74,79]. Afterslip is sometimes triggered not only on the fault that caused the earthquake, but also on surrounding faults within the stress field [1,12]. Another very noticeable consequence of an earthquake is the series of aftershocks following it, which may be triggered by combinations of stress changes as the lower crust relaxes after an earthquake as well as those due to motion during the earthquake itself [90].

In cases where the coseismic slip distribution (and it's associated stress change) is well-constrained by combinations of geodetic and seismic data, we can explore models of crustal and mantle properties, or place bounds on laboratory-derived rock mechanics laws [7], in order to fit the postseismic response to the coseismic stress change. For instance, the distribution of afterslip places constraints on the frictional behavior of the fault zone [32,55]. However, any such study requires good understand of the processes occurring during the earthquake itself.

**Interseismic Behavior** Although it is usually less dramatic than deformation occurring during and immediately after earthquakes, interseismic deformation can also tell us a great deal about the fault zone. The depth of the “locked” or “coupled” zone that will eventually rupture seismically [24,29], the rate at which stress is accumulating along the fault zone [59,92], and even variations in crustal elastic properties when rocks of different types are brought into contact across the fault zone [14], can all be addressed

by examination of interseismic deformation across a fault. In addition, data types that span a finite amount of time, such as InSAR or campaign GPS observations, will always contain some amount of interseismic strain that may need to be removed before studying measurements spanning an earthquake. The steadiest interseismic deformation is mainly due to flow of the lower crust and mantle beneath the elastic upper crust, but there are also observations of steady creep in the shallowest portions of several strike-slip fault zones [31,37,47,73].

**Transient Behavior** One of the most intriguing fault behaviors observed recently are isolated deformation events that are often not directly associated with an earthquake at all. These “slow” or “silent” earthquakes have been observed around the world in subduction zones, especially within the Cascadia subduction zone ([51,54], Fig. 6), Japan [27,56,65], Mexico [45], and Chile [64]. Along-strike variations in the amount of plate convergence that is taken up by these aseismic slip processes vs. large coseismic events may persist over very long timescales, as is indicated by the correlation of potentially aseismic “seismic gaps” in subduction zones and gravity anomalies [89,95]. Although most transient fault zone activity has been noticed along subduction zones, there are also observations of transient aseismic creep along the San Andreas Fault [60] and within the Imperial Valley, CA [42]. Further observations will help us understand how great of a contribution



Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 6  
 Episodic slow slip events in Cascadia. OP=Olympic Peninsula, VI=Vancouver Island. *Black arrows* indicate interseismic convergence, *red arrows* indicate motion during one of the slow slip events recorded by continuous GPS (*yellow triangles*). *Inset* shows detrended component of deformation recorded by stations ALBH on Vancouver Island, with more than a decade of regular slow slip events (*vertical green bars*). Figure after [54]

combinations of steady interseismic and transient aseismic behaviors make to the release of plate motion across many of the major tectonic boundaries around the world.

## Modeling of Geodetic Observations

### Overview of Inverse Theory

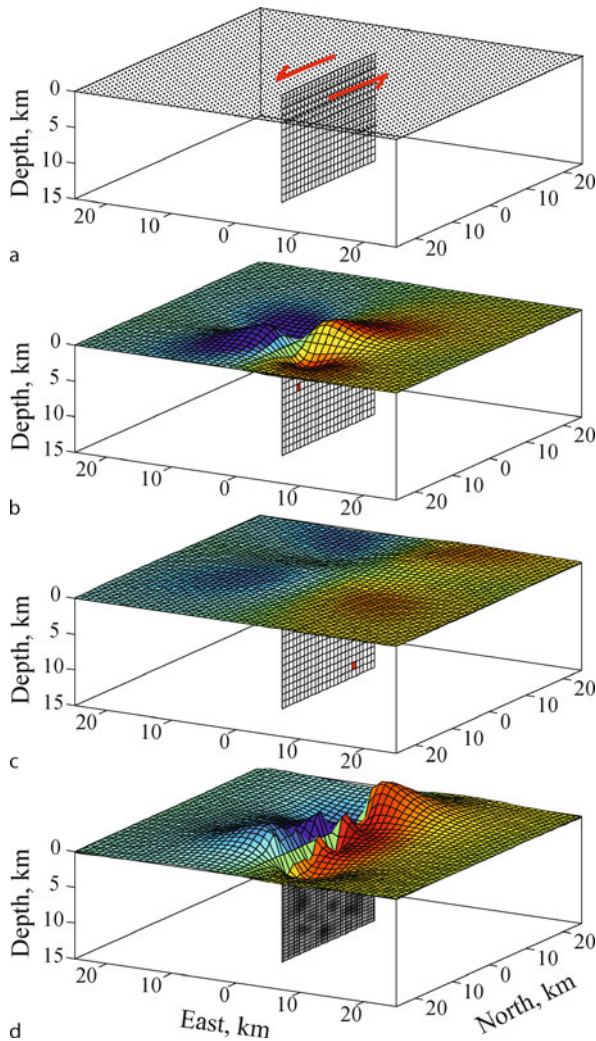
All of the observations discussed above involve measurements of surface deformation. We can learn nothing about processes occurring below the surface, within fault zones, without the ability to infer how those observations were generated. Inverse theory spans the range of problems relating to identifying which model of real-world behavior is consistent with observed data. Some of the simplest inverse problems include that of fitting a line to a series of

data points, or even calculating the mean value of a set of observations. Good treatments of the field of inverse theory and its applications to geophysical problems in general can be found in books by Parker [66] and Menke [52], as well as in key papers by Backus and Gilbert [4] and Tarantola and Valette [91].

Before a geophysical inversion can be performed (i. e., determining the size of an earthquake based on ground deformation) the “forward model” must be defined and understood. A forward model is a description of how we believe the system of interest, here, the Earth’s crust, will behave in response to the process we are studying (i. e., fault slip). For inversions for fault slip, the forward model is our best guess at how the deformation field produced by motion on the fault surface will propagate through the solid earth to the ground surface where we can observe it. Laboratory, seismological and field observations support the idea that the crust behaves as an elastic solid on the short timescales associated with earthquakes and their immediate aftermath, although parts of the crust deform viscously at longer time scales under prolonged stress. The response of the elastic crust surrounding a plane that undergoes fault slip can be described mathematically [62]. Figure 7 shows the predicted ground deformation for strike-slip motion on fault “patches” at varying depths. The ground deformation response is linear, i. e., the deformation from multiple sources is just the sum of deformation from the individual fault patches. Therefore, we can predict the expected ground deformation from arbitrarily complicated fault zone geometries and slip distributions.

In an inverse problem, we consider the behavior of a forward model or family of forward models from a number of possible sources that could potentially explain our data, and we find the combination these sources that best fits the data. In this section, the family of forward models is represented by slip on a fault plane that has been divided up into a number of fault patches, and the best-fit model would be the distribution of fault slip magnitude along this fault plane that best matches the observed ground deformation.

**Linear vs. Nonlinear** Inverse problems can be divided up into two major families: Linear and Nonlinear. In a linear problem, we solve for the sum of forward models that best fits the data, without any size constraints on the individual contributions from each forward model (fault patch). One example of a linear problem is where we find the best-fit fault slip distribution on a fixed fault plane, without any constraints such as requiring that the fault slip has a positive or negative sign (i. e., right-lateral vs. left lateral). Linear problems are essentially extensions of the



Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 7

**Predicted ground surface deformation for a vertical, left-lateral strike-slip fault a, using Okada [62]. Fault slip on a shallow fault patch (red, panel b) produces a larger amount of deformation over a smaller area than does fault slip on a deeper region of the fault plane c. Here, deformation of the grid indicates vertical deformation and color indicates the amount of northward (red) or southward (blue) deformation. The amount of total deformation from a complicated fault slip distribution is achieved by summing the components from all parts of the fault d**

problem of fitting a line to a set of data, and are quite easy to solve [52].

Nonlinear problems exist when the functional shape and form of the predicted data varies with the parameter choice. An inversion for the location of the fault plane itself, where quantities such as the  $x, y, z$ , location or fault plane strike and dip are varied, is very nonlinear. The

process of solving a nonlinear problem involves finding a model that best fits the data in some pre-defined way, usually by finding the model that produces the lowest sum of squared residuals with the data (least-squares fit). Nonlinear problems may have multi-dimensional misfit functions with poorly defined or multiple minima, and the only way to find the absolute or “global” minimum is by exhaustively searching the parameter space.

Methods for searching the parameter space in a nonlinear problem fall into two main camps. Gradient-based methods determine the slope and gradient of the misfit surface (often in many dimensions) and follow it downhill in an iterative way. These methods work well if the misfit surface has a single, well defined minimum. If the search begins near one of the local minima, it is possible that the true, global minimum will never be found. Global methods involve variations that range from simply sampling the parameter space very densely, to iterative methods that track many initial samples or families of samples in gradient-based searches. One method that performs very well in the  $\sim 9$ -dimensional parameter space of finding the best-fit fault plane and slip that fits data for a small earthquake is the Neighborhood Algorithm [44,78]. In the Neighborhood Algorithm (NA), the misfit values at many initial points are used to iteratively focus in on regions of the parameter space with lower misfit values. The NA method has the strength of being able to track and define multiple minima instead of just choosing one. A survey of several other nonlinear optimization methods applied to earthquake location problems can be found in Cervelli [6].

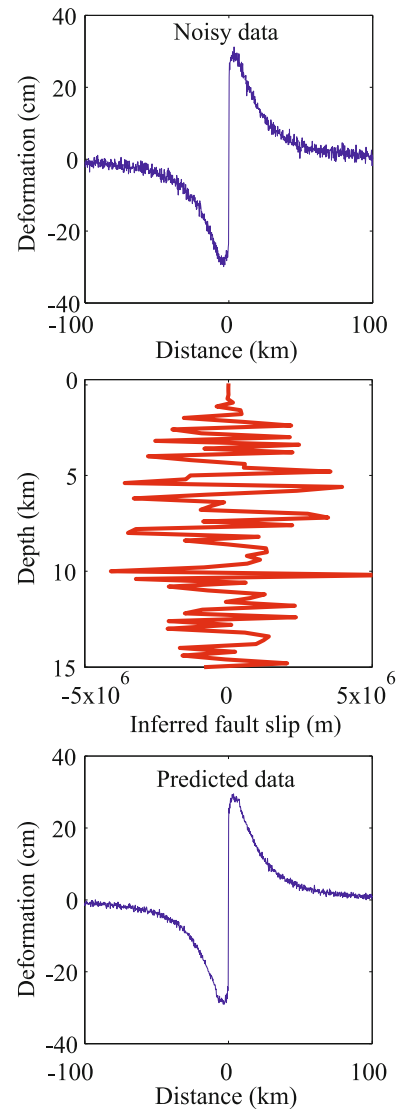
**Smoothing** The specific problem of inverting for fault slip on a particular fault plane has an interesting twist to it that is shared by many geophysical inverse problems. Just as an inversion for the best-fit line to data usually requires at least two data points, other inversions also effectively require the existence of more data than unknown quantities. For fault-slip problems the question does not boil down to simple numbers of data points vs. the number of fault patches – parts of the fault plane are better-resolved than others, and some data points contribute more information than is provided by their neighbors. For instance, if we had a thousand GPS receivers within 100 feet of the fault zone, but extended our target fault plane down to 100 km, the GPS data would be able to tell us little about what was going on at depth, even if we only divided the fault plane up into a few fault patches. Conversely, data points very close together and far from the fault contribute essentially the same information about what occurred – combining them can help to reduce noise, but it does not help us to isolate variations in fault slip along the fault.

Because of these inherent variations in resolution, both in model and data, an inversion can run into problems if we simply solve for the best-fit fault slip distribution on a finely-divided fault plane. Since the forward models for two fault patches at great depth are very similar on the surface, the inversion cannot determine the relative strength of fault slip that should be assigned to each. The difference between  $[0 \ 1]$  and  $[-1000 \ 1001]$  is very small when you propagate it up to the surface, often smaller than the noise in the data (or even machine precision!). Therefore, the “best-fit” model would have unrealistically large variations that have nothing to do with what really occurred during the earthquake (Fig. 8). This effect is often known as “checkerboarding”. While some of the features in the inferred slip distribution are related to the earthquake, much of the complexity is due to the attempt of the inversion to fit noise within the data.

There are several methods for dealing with this effect, all of which place some bounds on how large or spatially rough the variations in fault slip can be. The simplest method (although one of the hardest to optimize) is to parametrize the fault plane in a way that the fault patches are never so small that there are large tradeoffs in their predicted surface deformation [70]. The extreme of this would be to only use one fault patch. In most cases where we have a lot of data, this solution would not fit the data well nor tell us much about the earthquake.

Another method is to place a penalty during the least squares inversion on fault slip solutions that are very large (usually involving large variations) or that are spatially rough. These “regularized” inversions result in spatially smooth slip distributions that usually fit the data almost as well as the rough, unregularized inversions. Regularization always involves some choice of how much weight needs to be placed on the roughness penalty vs. the fit to the data, which can be a difficult procedure. Too large of a penalty weight and the slip distribution will be too smooth and will not fit the data (the logical extension of this is the single fault patch model discussed above). Too small of a weighting and the slip distribution will be arbitrarily rough, with unrealistic changes in sign throughout the fault zone.

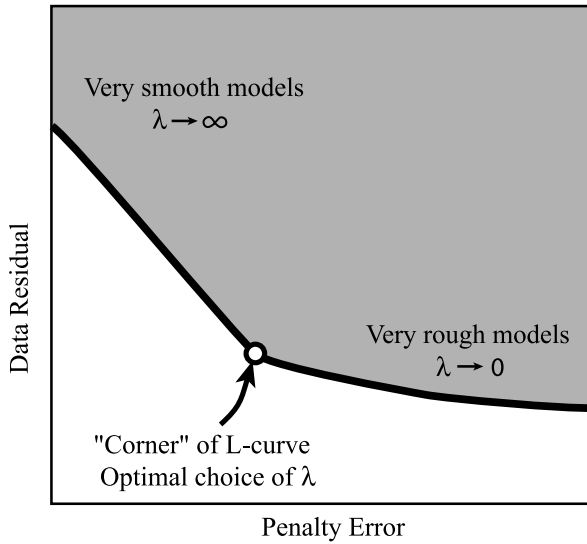
There are two main families of methods for choosing the appropriate smoothing weight (apart from random guessing, which often performs surprisingly well!). In general, as the amount of regularization is increased, the fault slip distribution becomes smoother and the fit to the data worsens. Ideally, we would choose the value where the proportion of the slip distribution that is just fitting data noise is as small as possible, without smoothing so much that we fit neither the noise nor the underlying signal due to the earthquake. The first method re-



Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 8

**Example of how noisy data can affect fault slip inversion. *Top*: Deformation from a vertical strike slip fault, with added noise. *Middle*: Best fit slip distribution inferred from noisy data. The large fluctuations mostly cancel each other out at the surface, producing (*bottom*) predicted data that matches both the underlying deformation signal and the noise**

quires choosing the smoothing weight from a plot of data fit vs. model roughness, often referred to as an “L-curve” because of the characteristic shape of such curves (Fig. 9). The smoothing value at the corner of the “L” generally fits the data well without being “too rough”. This type of parameter choice is slightly arbitrary and not easy to auto-



Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 9

**Typical L-curve.** Each point on the curve corresponds to a slip distribution inferred using a different value of  $\lambda$ . In most cases, the "penalty error" will be a measure of model roughness

mate, but it does allow the researcher to include some intuition about characteristics of the earthquake.

The other family relies on the concept that a good choice of smoothing would be able to reproduce another, independent set of data spanning the earthquake fairly well. Too much smoothing and the slip model would fit neither the original nor the additional data sets well. If the choice of smoothing were too small, the resulting complex slip distribution would mainly be fitting noise in the original data set and would not, therefore, fit the independent noise in the second data set.

Of course, we rarely have the luxury of multiple data sets – if they do exist, we should use them for the main inversion! Data resampling procedures, known variously as the bootstrap, jackknife or cross-validation [10] are used to simulate the existence of multiple, independent data sets in cases where only one exists. Du et al. [9] compared cross-validation and other techniques for choosing smoothing parameters for geodetic data spanning the 1983 Borah Peak earthquake. Another powerful parameter choice method is the Akaike Bayesian Information Criterion (ABIC) that can be used to choose smoothing weights or other inversion characteristics [1,30,97].

**Noise** Although finding the best-fit model to our data is certainly important, knowing how confident we are in that estimate is just as crucial. Studies of postseismic behavior, for instance, rely on good estimates of the coseismic

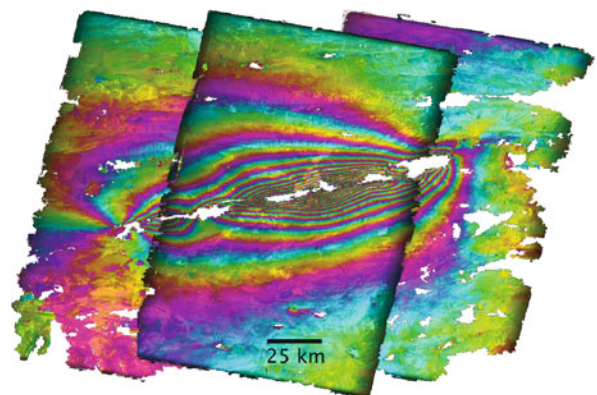
slip distribution [25], and are improved further when we know what constraints we can place on that slip distribution. A variety of techniques exist for estimating these confidence limits – for linear problems, and when we know the character of the noise (magnitude, spatial correlation), it is quite simple to propagate data errors through to error bounds on the inversion results [52]. However, for nonlinear problems, and for cases where we are not quite sure how much of our signal is noise (usually the case with InSAR data), we need to rely on other methods for estimating the noise.

The same data resampling procedures described above can be used to generate multiple sets of inversion results that should reflect the noise structure of the data, even when it is not understood ahead of time. For nonlinear problems, such as the 3-D location of the best fit fault plane and earthquake mechanism for a particular earthquake, knowledge about the data noise can be used to construct multiple synthetic data sets that can then each be inverted using the nonlinear method of choice. Any conclusions about the problem (i. e., error bounds on fault slip) should be drawn from the ensemble as a whole [44].

### Case Examples

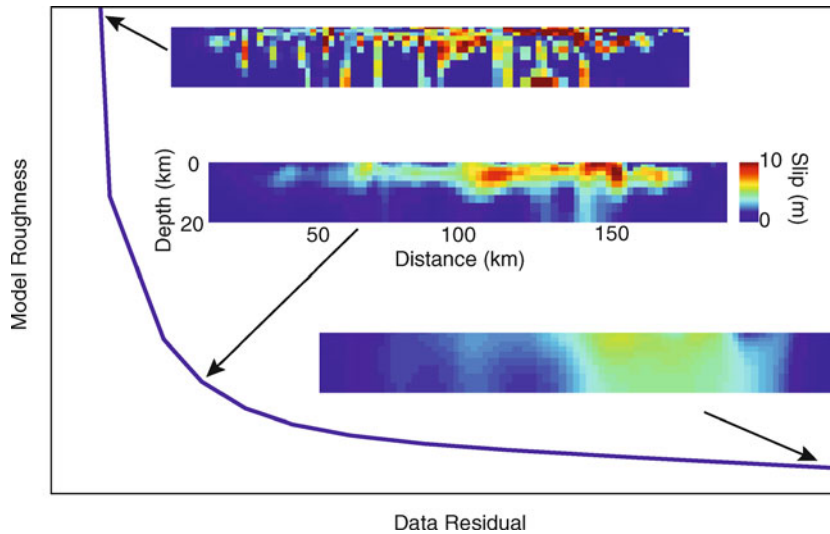
Here we examine case examples for two earthquakes. In each case, we review the data and discuss characteristics of the inferred slip distributions.

**Mw 7.6 Manyi, Tibet, Earthquake** The 1997 strike-slip earthquake that occurred Tibet produced a very long rupture (170 km) with up to 7 meters of offset. It was such a long rupture that it takes three overlapping SAR tracks



Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 10

**Three overlapping interferograms spanning the 1997 Manyi earthquake**



Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 11

L-curve for the Manyi EQ, showing slip distributions inferred using 3 different values of smoothing, each with the same color scale. Small  $\lambda$  (top panel) produces an unrealistically rough slip distribution, whereas very large  $\lambda$  (bottom panel) produces a smooth model that doesn't fit the data

(Fig. 10) to cover its full length! Here, we show the effects of variations in the spatial smoothing weights ( $\lambda$ ) on an inversion of the deformation field for fault slip. Figure 11 illustrates the differences in inferred slip distributions for various magnitudes of  $\lambda$ , resulting in a spectrum between very smooth slip distributions that do not fit the data very well, to impossibly rough slip distributions that only fit the data marginally better than the “optimal” model (center panel). Our optimal model predicts up to 10 meters of fault slip with most slip occurring in the upper 10 km. Here, we chose the optimal value of  $\lambda$  using cross-validation, but the other methods discussed above can also be applied to this problem.

Asymmetries in the profiles of deformation across the fault during the Manyi earthquake have led some researchers to think that the elastic behavior of the crust may be different on one side of the fault vs. the other [67]. The magnitude of horizontal ground motion for a straight, vertical, strike-slip fault should be symmetric across the fault, if there are no other complications. However, variations in fault plane dip can affect the resulting ground deformation as much as variations in crustal properties, requiring that such studies carefully consider the fault geometry used in their inversions.

**Mw 7.1 Hector Mine, CA, Earthquake** The 1999 Hector Mine earthquake [16,85] occurred in the Mojave Desert of Southern California, and was widely studied due to its

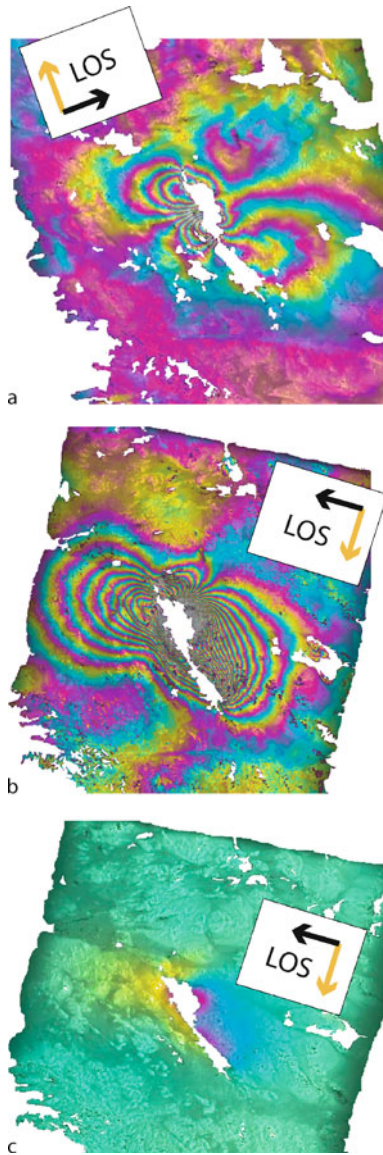
proximity to large population centers and the San Andreas fault, it's very complete data coverage (GPS, InSAR, seismic data), and the fact that it occurred only a few years after another very large earthquake on a neighboring fault (1992 Landers earthquake). The combination of the two earthquakes within such a short time interval suggests that series of earthquakes may occur separated by long intervals of quiescence instead of events always occurring at semi-regular time intervals of strain buildup and release [61,69].

In Fig. 12, we show a subset of the available InSAR data spanning the Hector Mine earthquake. Both continuous and campaign GPS data were also recorded on either side of the Y-shaped rupture. In Fig. 13, we show the results of inverting the GPS and all the InSAR data for the best-fit fault slip distribution, with up to 6 meters of fault slip. Note how the maximum fault slip is not at the surface, but peaks at a few km depth. If we plot the average slip vs. depth (Fig. 14), the profile clearly shows how the shallow slip deficit is robust even given the amount of noise in the data.

### Future Directions

The spatial and temporal coverage of geodetic data sets such as InSAR and GPS is increasing to the point where we can observe new types of fault zone behavior and solve for characteristics on a scale and precision that could hardly have been imagined a few decades ago. As we increase

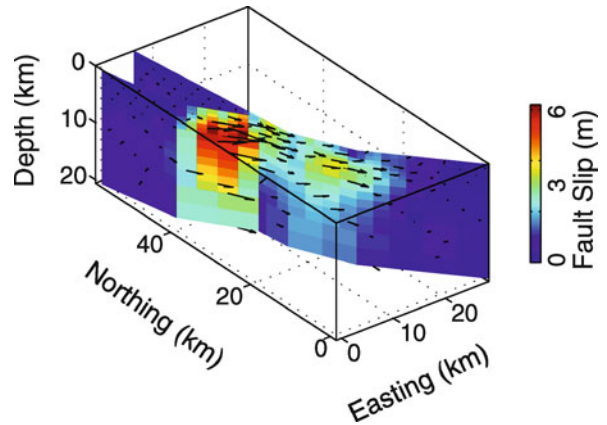




Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 12

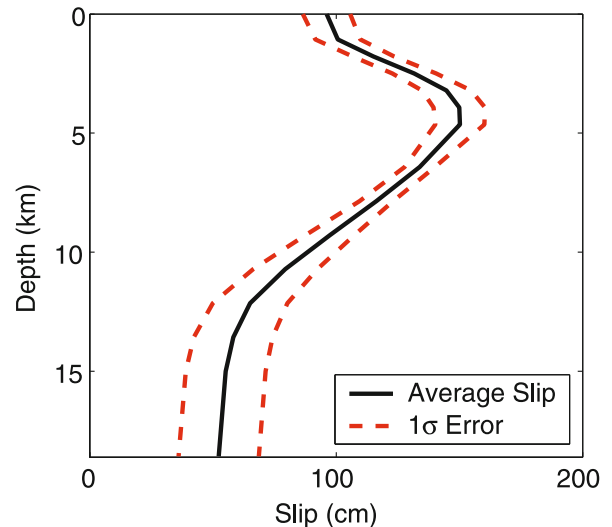
Three representations of the Hector Mine EQ. **a** "Ascending" and **b** "descending" wrapped interferograms, where the satellite is moving in the direction indicated by the *yellow arrow* and viewing in the direction of the *black arrow*. Note how the same deformation field looks very different depending on the viewing direction. Color cycles correspond to  $\sim 3$  cm of deformation. **c** Unwrapped version of interferogram in **b**. Now the color scale corresponds to 3 meters and spans the entire dynamic range of the data

our ability to understand and describe the kinematics of crustal deformation, we can begin to explore the dynamic processes driving seismic and volcanic deformation in tec-



Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 13

3D representation of inferred fault slip that occurred during the Hector Mine earthquake



Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 14

Fault slip during the Hector Mine earthquake, averaged along strike and plotted vs. depth. *Red dashed lines* indicate the expected error bounds introduced by atmospheric noise

tonically active regions around the world. Three of the main fields that show the most promise in the near future are discussed below:

### Transient Deformation and Slow, Silent Earthquakes

The episodic slow slip events that show up with amazing regularity ( $\sim 14$  months) in Cascadia ([51,54], Fig. 6), and that are also observed in Japan [27,56,65] and Mexico [45], are some of the most exciting new types of fault

zone behavior that scientists have observed in recent years. The idea that earthquakes contained energy that was released by slower ruptures was noticed using seismic data far earlier than it was ever observed geodetically (1960 Chilean earthquake [34]). The slow slip events are often associated with heightened levels of spatially-correlated, low-level seismicity, or seismic tremor [87]. Currently, the process that allows these slow earthquakes to initiate and propagate at the observed speeds is still unknown, although they may be occurring at depths where dehydration of subducted sediments becomes important.

Apart from their intrinsic interest (since we always seek to explain processes that we do not understand), these slow/silent events release a not insignificant amount of the accumulated strain across the fault boundaries. In some ways this is a boon to people living nearby – any plate tectonic motion that is accommodated aseismically means that there is that much less accumulated strain that could be released in a destructive earthquake. However, there are indications that aseismic slip events are often closely followed by earthquakes [42,75], indicating that the changes in stress during the slow event can push other, nearby, regions past the brink until they rupture coseismically. This, also, can be seen in a positive light – perhaps close monitoring of aseismic deformation along active plate boundaries can serve as an early warning system, or at least a signal to raise the forecasted earthquake hazard whenever heightened activity is observed.

### Reduction or Modeling of Atmospheric Noise

One of the largest hurdles for researchers approaching InSAR data is the atmospheric noise present in all interferograms. GPS data is affected by this problem as well, but the atmospheric signal can be solved for or averaged out due to the long observation times allowed by GPS. An area of active research is the modeling and removing of the atmospheric noise signal from InSAR data, using a variety of tools that range from other satellite-based observations of atmospheric water vapor content to simplistic models of correlations between elevation and interferometric phase [96].

The first-order layering of the atmosphere results in profiles of water vapor that tend to decrease vs. elevation in a manner that often appears essentially linear when two SAR images are combined in an interferogram. However, lateral variations with distance from water bodies and gradients from one side of a mountain belt to the other, result in the fact that the appropriate elevation vs. signal can vary quite a bit across a typical SAR image. For some target problems, such as the location of a small earthquake or

the fault slip distribution for an earthquake in a relatively flat area, these elevation-dependent signals will likely affect the inversion to only a limited extent. However, in studies where the signal of interest is correlated with elevation, such as fold growth or subsidence caused by lake loading, great care should be taken when removing overly simple models of atmospheric water vapor from the data.

A more promising, albeit still problematic, approach is the modeling of water vapor content based on observations from satellite-based platforms. Water vapor measurements are made by a variety of satellites, but two in particular have the spatial scale and temporal resolution that make them potentially very useful in InSAR applications. The Moderate Resolution Imaging Spectroradiometer (MODIS) instruments on both the Terra and Aqua satellites acquire almost daily observations over most of the Earth's surface at fairly high resolution (1 km spacing). They are not acquired at exactly the same time as the SAR images, but they can still be used to track seasonal trends or to seek individual features that may persist between the SAR and MODIS image acquisition times. While these cannot be used to quantitatively remove the atmospheric signal, they can be very useful from a qualitative standpoint in determining whether a particular interferometric signal may or may not be tectonic in origin.

The Medium Resolution Imaging Spectrometer (MERIS) is physically located on the same satellite (ENVISAT) that acquires much of the SAR imagery used today. Since MERIS observations are made at essentially the same time as the SAR imagery, they see the same atmosphere and can not only be used in the same qualitative way as MODIS, but also show promise of allowing the actual removal of atmospheric contributions to the interferometric phase. Some difficulties lie in the fact that MERIS and MODIS measurements of atmospheric water vapor can only be made in cloud-free images, and that “double-bounce” effects of the signal bouncing off the base of both visible and invisible clouds can bias the observations and introduce spurious water vapor features. Still, these satellite-based methods show great promise. The most robust method for reducing the contribution from atmospheric noise to our inversions remains the use of multiple independent data sets, whenever available.

### Data Assimilation

As mentioned in Sect. “**Highlights of Earthquake Geodesy**”, the vast quantity of data now available to us can prove troublesome – if it takes hours to perform one forward model of a possible earthquake scenario with the available data, then the inverse problem quickly becomes

unmanageable (or at least can begin to extend past the length of a normal graduate thesis). Parallel computing methods and the rapid decrease in cost for computing resources now makes the operation of large, multi-processor machines feasible within individual research departments. These large machines allow us to approach inverse problems with the “big hammer” of stochastic, or Monte Carlo, methods, which rely on the use of many randomly generated simulations of a system [6]. However, the expected continuing increase of both spatial and temporal coverage of deformation observations requires that we need to continue developing new tools that allow us to capitalize on data time series as well as individual snapshots of ground deformation.

Kalman filters and related methods are one very powerful tool now in widespread use, especially within the GPS community [51,84]. The incorporation of InSAR data into such methods is slightly more difficult, in part because of the large number of data points, but also because of the varying character of noise between interferograms and difficulties in dissociating the description of noise vs. modeling of the geophysical signal of interest. Data assimilation methods developed in the atmospheric sciences, which also deal with data sets of varying spatial and temporal scales and resolutions, are a potentially rich source of tools that the geodetic community can explore in the near future.

## Bibliography

### Primary Literature

- Akaike H (1980) Bayesian statistics. In: Bernardo JM, DeGroot MH, Lindley DV, Smith AFM (eds) Likelihood and the Bayes procedure. University Press, Valencia, pp 143–166
- Allen CR et al (1972) Displacements on the Imperial, Superstition Hills, and San Andreas faults triggered by the Borrego Mountain Earthquake. US Geol Surv Prof Pap 787:87–104
- Arnadottir T, Segall P (1994) The 1989 Loma Prieta earthquake imaged from inversion of geodetic data. J Geophys Res 99:21835–21855
- Backus G, Gilbert F (1970) Uniqueness in the inversion of inaccurate gross earth data. Phil Trans R Soc Lond 266:123–192
- Burgmann R et al (2002) Time-dependent afterslip on and deep below the Izmit earthquake rupture. Bull Seism Soc Amer 92:126–137
- Cervelli P et al (2001) Estimating source parameters from deformation data, with an application to the March 1997 earthquake swarm off the Izu Peninsula, Japan. J Geophys Res 106:11217–11237
- Dieterich JH (1992) Earthquake nucleation on faults with rate- and state- dependent strength. Tectonophysics 211:115–134
- Dominguez S, Avouac JP, Michel R (2003) Horizontal coseismic deformation of the 1999 Chi-Chi earthquake measured from SPOT satellite images: Implications for the seismic cycle along the western foothills of central Taiwan. J Geophys Res 108. doi: [10.1029/2001JB000951](https://doi.org/10.1029/2001JB000951)
- Du Y, Aydin A, Segall P (1992) Comparison of various inversion techniques as applied to the determination of a geophysical deformation model for the 1983 Borah Peak earthquake. Bull Seism Soc Amer 82:1840–1866
- Efron B, Tibshirani R (1993) An introduction to the bootstrap. In: Monographs on statistics and applied probability, vol 83. Chapman and Hall, London
- Emardson TR, Simons M, Webb FH (2003) Neutral atmospheric delay in interferometric synthetic aperture radar applications: Statistical description and mitigation. J Geophys Res 108. doi: [10.1029/2002JB001781](https://doi.org/10.1029/2002JB001781)
- England P, Jackson J (1989) Active deformation of the continents. Annu Rev Earth Planet Sci 17:197–226
- Fialko Y (2004) Evidence of fluid-filled upper crust from observations of post-seismic deformation due to the 1992 Mw 7.3 Landers earthquake. J Geophys Res 109. doi: [10.1029/2004JB002985](https://doi.org/10.1029/2004JB002985)
- Fialko Y (2006) Interseismic strain accumulation and the earthquake potential on the southern San Andreas fault system. Nature 441:968–971
- Fialko Y et al (2002) Deformation on nearby faults induced by the 1999 Hector Mine Earthquake. Science 297:1858–1862
- Fialko Y, Simons M, Agnew D (2001) The complete (3-D) surface displacement field in the epicentral area of the 1999 Mw 7.1 Hector Mine earthquake, California, from space geodetic observations. Geophys Res Lett 28:3063–3066
- Freymueller J, King NE, Segall P (1994) The Co-seismic slip distribution of the Landers earthquake. Bull Seism Soc Amer 84:646–659
- Funning GJ et al (2005) Surface displacements and source parameters of the 2003 Bam (Iran) earthquake from Envisat advanced synthetic aperture radar imagery. J Geophys Res 110. doi: [10.1029/2004JB003338](https://doi.org/10.1029/2004JB003338)
- Gilbert GK (1890) Lake Bonneville. In: US Geol Surv Monograph, vol 1. Washington
- Goldstein R (1995) Atmospheric limitations to repeat-track radar interferometry. Geophys Res Lett 22:2517–2520
- Griesbach CL (1893) Notes on the earthquake in Baluchistan on the 20th December 1892. Geol Survey India Rec 26
- Hager BH, King RW, Murray MH (1991) Measurement of crustal deformation using the global positioning system. Annu Rev Earth Planet Sci 19:351–382
- Hanssen RA (2001) Radar interferometry: Data interpretation and error analysis. Kluwer, Dordrecht
- Harris RA, Segall P (1987) Detection of a locked zone at depth on the Parkfield, California segment of the San Andreas Fault. J Geophys Res 92:7945–7962
- Hearn EH (2002) Dynamics of Izmit earthquake postseismic deformation and loading of the Duzce earthquake hypocenter. Bull Seism Soc Amer 92:172–193
- Hetland EA, Hager BH (2003) Postseismic relaxation across the central Nevada seismic belt. J Geophys Res 108. doi: [10.1029/2002JB002257](https://doi.org/10.1029/2002JB002257)
- Hirose H et al (1999) A slow thrust slip event following the two 1996 Hyuganada earthquakes beneath the Bungo Channel, southwest Japan. Geophys Res Lett 26:3237–3240
- Hsu YJ et al (2006) Frictional afterslip following the Mw 8.7, 2005 Nias-Simeulue earthquake, Indonesia. Science 312. doi: [10.1126/science.1126960](https://doi.org/10.1126/science.1126960)

29. Ito T, Hashimoto M (2004) Spatiotemporal distribution of interplate coupling in southwest Japan from inversion of geodetic data. *J Geophys Res* 109. doi: [10.1029/2002JB002358](https://doi.org/10.1029/2002JB002358)
30. Jackson DD, Matsuura M (1985) A Bayesian approach to non-linear inversion. *J Geophys Res* 90:581–591
31. Johanson IA, Burgmann R (2005) Creep and quakes on the northern transition zone of the San Andreas Fault from GPS and InSAR data. *Geophys Res Lett* 32. doi: [10.1029/2005GL023150](https://doi.org/10.1029/2005GL023150)
32. Johnson KM, Burgmann R, Larson K (2006) Frictional properties on the San Andreas fault near Parkfield, California, inferred from models of afterslip following the 2004 earthquake. *Bull Seism Soc Amer* 96:5321–5338
33. Jonsson S et al (2003) Post-earthquake ground movements correlated to pore-pressure transients. *Nature* 424:179–183
34. Kanamori H, Stewart GS (1972) A slow earthquake. *Phys Earth Planet Int* 18:167–175
35. King GCP, Stein RS, Rundle JB (1988) The growth of geological structures by repeated earthquakes 1: Conceptual framework. *J Geophys Res* 93:13307–13318
36. Koto B (1983) On the cause of the great earthquake in central Japan, 1891. *J Coll Sci Imp Univ Japan* 5:296–353
37. Langbein J, Gwyther RL, Hart RHG, Gladwin MT (1999) Slip-rate increase at Parkfield in 1993 detected by high-precision EDM and borehole tensor strainmeters Source. *Geophys Res Lett* 26(16):2529–2532
38. Langbein J, Murray JR, Snyder HA (2006) Coseismic and initial postseismic deformation from the 2004 Parkfield, California, Earthquake, observed by Global Positioning System, Electronic Distance Meter, Creep Meters, and Borehole Strainmeters. *Bull Seism Soc Amer* 96:304–320
39. Larsen S et al (1992) Global Positioning System measurements of deformations associated with the 1987 Superstition Hills earthquake – evidence for conjugate faulting. *J Geophys Res* 97:4885–4902
40. Lawson AC et al (1908) The California earthquake of April 18, 1906 – report of the state earthquake investigation committee. Carnegie Insitute, Washinton
41. Lin J, Stein RS (1989) Coseismic folding, earthquake recurrence and the 1987 source mechanism at Whittier Narrows, Los Angeles Basin, California. *J Geophys Res* 94:9614–9632
42. Lohman RB, McGuire JJ (2007) Earthquake swarms driven by aseismic creep in the Salton Trough, California. *J Geophys Res* 112. doi: [10.1029/2006JB004596](https://doi.org/10.1029/2006JB004596)
43. Lohman RB, Simons M (2005) Some thoughts on the use of InSAR data to constrain models of surface deformation: Noise structure and data downsampling. *Geochem Geophys Geosyst* 6. doi: [10.1029/2004GC000841](https://doi.org/10.1029/2004GC000841)
44. Lohman RB, Simons M (2005) Locations of selected small earthquakes in the Zagros mountains. *Geochem Geophys Geosyst* 6. doi: [10.1029/2004GC000849](https://doi.org/10.1029/2004GC000849)
45. Lowry AR et al (2001) Transient fault slip in Guerrero, southern Mexico. *Geophys Res Lett* 28:3753–3756
46. Lyell C (1837) *Principles of Geology*, 5th edn. Murray, London
47. Lyons S, Sandwell D (2003) Fault creep along the southern San Andreas from interferometric synthetic aperture radar, permanent scatterers and stacking. *J Geophys Res* 108. doi: [10.1029/2002JB001831](https://doi.org/10.1029/2002JB001831)
48. Mallet R (1862) *The first principles of observational seismology*. Chapman and Hall, London
49. Massonnet D, Rossi M, Carmona C, Adragna F, Pelzer G, Feigl K, Rabaute T (1993) The displacement field of the Landers earthquake mapped by radar interferometry. *Nature* 364: 138–142
50. McKay A (1890) On the earthquake of September 1888, in the Amuri and Marlborough districts of the South Island. *NZ Geol Surv Rep Geol Explor 1885–1889* 20:78–1007
51. McGuire J, Segall P (2003) Imaging of aseismic fault slip transients recorded by dense geodetic networks. *Geophys J Int* 155:778–788
52. Menke W (1989) *Geophysical data analysis: Discrete inverse theory*. Academic Press, London
53. Middlemiss CS (1910) The Kangra earthquake of 4th April, 1905. *Geol Surv India Mem* 37
54. Miller MM et al (2002) Periodic slow earthquakes from the Cascadia subduction zone. *Science* 295:2423
55. Miyazaki S et al (2004) Space time distribution of afterslip following the 2003 Tokachi-oki earthquake: Implications for variations in fault zone frictional properties. *Geophys Res Lett* 31. doi: [10.1029/2003GL019410](https://doi.org/10.1029/2003GL019410)
56. Miyazaki S, McGuire JJ, Segall P (2003) A transient subduction zone slip episode in southwest Japan observed by the nationwide GPS array. *J Geophys Res* 108. doi: [10.1029/2001JB000456](https://doi.org/10.1029/2001JB000456)
57. Molnar P, Tapponnier P (1975) Cenozoic tectonics of Asia: Effects of a continental collision. *Science* 189:419–425
58. Muller JJA (1895) De verplaatsing van eenige traingulatie pilaren in de residenti Tapanuli (Sumatra) tengevolge de aardbeving van 17 Mei 1892. *Natuurwet Tijdscht Ned Indie* 54: 299–307
59. Murray J, Segall P (2002) Testing time-predictable earthquake recurrence by direct measurement of strain accumulation and release. *Nature* 419:298–291
60. Nadeau RM, McEvilly TV (2004) Periodic pulsing of characteristic microearthquakes on the San Andreas Fault. *Science* 303:202–222
61. Nur A, Hagai R, Beroza G (1993) The nature of the Landers-Mojave earthquake line. *Science* 261:201–203
62. Okada Y (1985) Surface deformation due to shear and tensile faults in a half space. *Bull Seism Soc Amer* 75:1135–1154
63. Okudo T (1950) On the mode off the vertical land-deformation accompanying the great Nankaido earthquakes. *Bull Geogr Surv Inst* 2:37–59
64. Oldham RD (1928) The Cutch (Kacch) earthquake of 16th June 1819, with a revision of the great earthquake of 12th June 1897. *Geol Surv India Mem* 46:71–147
65. Ozawa S et al (2002) Detection and monitoring of ongoing aseismic slip in the Tokai region, central Japan. *Science* 298:1009–1012
66. Parker RL (1977) Understanding inverse theory. *Annu Rev Earth Planet Sci* 5:35–64
67. Peltzer G, Crampe F, King G (1999) Evidence of nonlinear elasticity of the crust from the Mw 7.6 Manyi (Tibet) earthquake. *Science* 286:272–276
68. Pollitz FF (2003) Transient rheology of the uppermost mantle beneath the Mojave desert, California. *Earth Plan Sci Lett* 215:89–104
69. Pollitz FF, Sacks IS (2002) Stress triggering of the 1999 Hector Mine earthquake by transient deformation following the 1992 Landers Earthquake. *Bull Seism Soc Amer* 92:1487–1496
70. Pritchard ME et al (2002) Co-seismic slip from the 1995 July 30 Mw=8.1 Antofagasta, Chile, earthquake as constrained by InSAR and GPS observations. *Geophys J Int* 150:362–376

71. Pritchard ME, Simons M (2006) An aseismic slip pulse in northern Chile and along-strike variations in seismogenic behavior. *J Geophys Res* 111. doi: [10.1029/2006JB004258](https://doi.org/10.1029/2006JB004258)
72. Reid HF (1910) The mechanics of the earthquake. In: Lawson AC (ed) *The California earthquake of April 18, 1906*. Carnegie Institute, Washington
73. Reid HG (1911) The elastic-rebound theory of earthquakes. *Univ Calif Pub Bull* 6:413–444
74. Reilinger R (1986) Evidence for postseismic viscoelastic relaxation following the 1959 M=7.5 Hebgen Lake, Montana, earthquake. *J Geophys Res* 91:9488–9494
75. Roeloffs EA (2006) Evidence for aseismic deformation rate changes prior to earthquakes. *Annu Rev Earth Planet Sci* 34:591–627
76. Rosen PA et al (2000) Synthetic aperture radar interferometry. *Proc IEEE* 88:333–382
77. Ruegg JC et al (1996) The Mw=8.1 Antofagasta (North Chile) earthquake July 30, 1995: First results from teleseismic and geodetic data. *Geophys Res Lett* 23:917–920
78. Sambridge M (1998) Geophysical inversion with a neighborhood algorithm – I: Searching a parameter space. *Geophys J Int* 138:479–494
79. Savage JC, Lisowski M, Svarc JL (1994) Postseismic deformation following the 1989 (M=7.1) Loma Prieta, California, earthquake. *J Geophys Res* 99:13757–13765
80. Schmidt DA et al (2005) Distribution of aseismic slip rate on the Hayward fault inferred from seismic and geodetic data. *J Geophys Res* 110. doi: [10.1029/2004JB003397](https://doi.org/10.1029/2004JB003397)
81. Scholz CH (1998) Earthquakes and friction laws. *Nature* 391: 37–42
82. Segall P, Burgmann R, Matthews M (2000) Time-dependent triggered afterslip following the 1989 Loma Prieta earthquake. *J Geophys Res* 105:S615–S634
83. Segall P, Davis JL (1997) GPS Applications for Geodynamics and Earthquake Studies. *Annu Rev Earth Planet Sci* 25:301–336
84. Segall P, Matthews M (1997) Time dependent inversion of geodetic data. *J Geophys Res* 102:22391–22409
85. Simons M, Fialko Y, Rivera L (2002) Coseismic deformation from the 1999 Mw 7.1 Hector Mine, California, earthquake as inferred from InSAR and GPS observations. *Bull Seism Soc Amer* 92:1390–1402
86. Sharp RV et al (1982) Surface faulting in the Imperial Valley. In: Sharp RV, Lienkaemper JJ, Bonilla MG, Burke DB, Cox BF, Herd DG, Miller DM, Morton DM, Ponti DJ, Rymer MJ, Tinsley JC, Yount JC, Kahle JE, Hart EW, Sieh K (eds) *The Imperial Valley, California, earthquake of October 15, 1979*. US Geol Surv Pap 1254:119–144
87. Shelley DR et al (2006) Low frequency earthquakes in Shikoku, Japan, and their relationship to episodic tremor and slip. *Nature* 442:188–191
88. Shimazaki K, Nakata T (1980) Time-Predictable Recurrence Model for Large Earthquakes. *Geophys Res Lett* 7(4):279–282
89. Song AT, Simons M (2003) Large trench-parallel gravity variations predict seismogenic behavior in subduction zones. *Science* 301:630–633
90. Stein R (1999) The role of stress transfer in earthquake occurrence. *Nature* 402:605–609
91. Tarantola A, Valette B (1982) Inverse problems = quest for information. *J Geophys* 50:159–170
92. Thatcher W (1984) The earthquake deformation cycle at the Nankai trough, southwest Japan. *J Geophys Res* 89:3087–3101
93. Tsuboi C (1932) Investigation on the deformation of the earth's crust in the Tango district connected with the Tango earthquake of 1927 (part 4). *Bull Earthq Res Inst Tokyo Univ* 10: 411–434
94. Ward S, Valensise GR (1989) Fault parameters and slip distribution of the 1915 Avezzano, Italy, earthquake derived from geodetic observations. *Bull Seism Soc Amer* 79:690–710
95. Wells RE et al (2003) Basin-centered asperities in great subduction zone earthquakes: A link between slip, subsidence and subduction erosion. *J Geophys Res* 108. doi: [10.1029/2002JB002072](https://doi.org/10.1029/2002JB002072)
96. Williams S, Bock Y, Fang P (1998) Integrated satellite interferometry: Tropospheric noise, GPS estimates and implications for interferometric synthetic aperture radar products. *J Geophys Res* 103:27051–27067
97. Yabuki T, Matsuura M (1992) Geodetic data inversion using a Bayesian information criterion for spatial distribution of fault slip. *Geophys J Int* 109:363–375
98. Yang M et al (2000) Geodetically observed surface displacements of the 1999 Chi-Chi, Taiwan earthquake. *Earth Planet Space* 52:403–413
99. Yeats RS, Sieh K, Allen CF (1997) *The geology of earthquakes*. Oxford University Press, New York
100. Zebker HA, Rosen PA, Hensley S (1997) Atmospheric effects in interferometric synthetic aperture radar surface deformation and topographic maps. *J Geophys Res* 102:7547–7563

## Books and Reviews

- Tse ST, Rice JR (1986) Crustal earthquake instability in relation to the depth variation of frictional slip properties. *J Geophys Res* 91:9452–9472

## Cryosphere Models

ROGER G. BARRY  
NSIDC, CIRES, University of Colorado, Boulder, USA

### Article Outline

Glossary  
 Definition of the Subject  
 Introduction  
 Snow Cover  
 Floating Ice  
 Glaciers  
 Ice Sheets  
 Frozen Ground and Permafrost  
 Future Directions  
 Bibliography

### Glossary

**Cryosphere** All forms of terrestrial snow and ice.  
**Newtonian viscous body** A body whose stress at each point is linearly proportional to its strain rate at that point.

### Definition of the Subject

The cryosphere comprises all terrestrial forms of snow and ice – snow cover, floating ice, glaciers, ice sheets, frozen ground and permafrost. It is a critical element of the climate system because of its high reflectivity, its insulating effects on the land and ocean, and its storage of water on short and long time scales. Numerical models of components of the cryosphere have been developed over the last 30 years or so, and some elements of these are now incorporated in coupled climate models and earth system models.

### Introduction

Currently there are no comprehensive models of the entire cryosphere. Rather there is a wide range of models of components of the cryosphere – snow cover, floating ice, glaciers, ice sheets, frozen ground and permafrost – and various components are treated with varying degrees of detail in coupled atmosphere-ocean-land models. Cryospheric processes are generally parametrized in such earth system models.

Models developed for each of the main cryospheric components are discussed in turn.

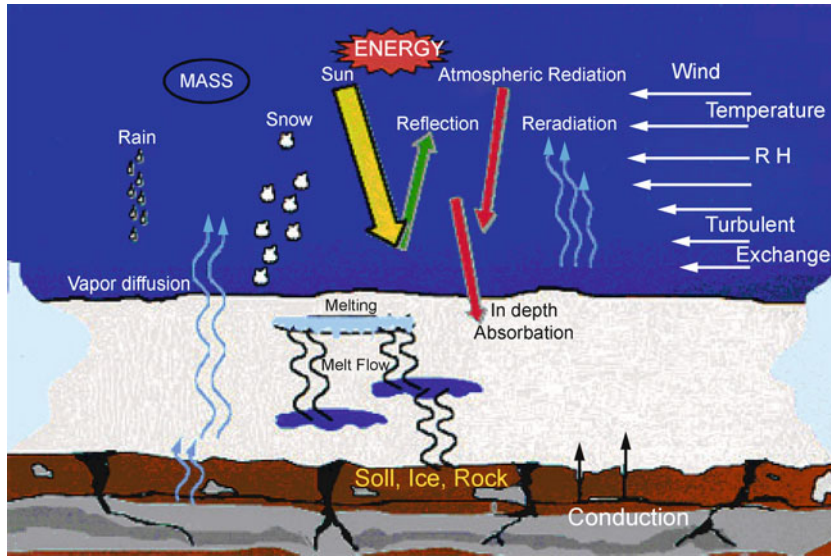
### Snow Cover

Snow cover is observed in situ at hydrometeorological stations, from daily depth measurements, (monthly) snow courses and in special automated networks such as the western United States Snow Telemetry (SNOTEL) network of snow pressure pillows. Its extent is also observed and mapped daily (since June 1999) over the Northern Hemisphere from operational satellites of the National Oceanic and Atmospheric Administration (NOAA) in the USA. Snow covers about 47 million km<sup>2</sup> at maximum in January and there is only a small area in South America in July. Hemispheric snow water equivalent estimates are routinely made from passive microwave data (1979-present) with a 25-km resolution (Armstrong et al. [4]).

There are numerous models of the formation and disappearance of snow cover. Many have a hydrological focus aimed at estimating seasonal runoff. Some use simple temperature degree-day formulations while others incorporate a full energy balance calculation. Dozier and Painter [14] examine the use of multispectral and hyperspectral remote sensing to estimate the snow's spectral albedo, along with other properties such as grain size, contaminants, temperature, liquid water content, and depth or water equivalent.

The U.S. Army Cold Regions Research and Engineering Laboratory Model SN THERM (SNow THERmal Model) is a 1-D energy balance model for snow and soil that is forced by meteorologically determined surface fluxes [41]. It simulates in-snow properties and processes, such as heat conduction, water flow, melt, vapor flow, compaction, grain growth, and solar absorption (see Fig. 1). The output provides snow depth, profiles of snow temperature, water content, density, grain size, and surface fluxes of sensible heat and evaporation. Surface boundary conditions require: incoming solar and longwave radiation; wind speed, air temperature and humidity at some reference height; and precipitation. The model will estimate solar and longwave radiation from cloud cover, if data on these variables are not available. Lower boundary conditions include soil textural properties (currently clay or sand used as defaults), wetness and temperature profile.

A comparative study of three snow models with different complexities was carried out by Jin et al. [39] to assess how a physically detailed snow model can improve snow modeling within general circulation models. The three models were (a) SN THERM; (b) a simplified three-layer model, Snow-Atmosphere-Soil Transfer (SAST), which includes only the ice and liquid-water phases; and (c) the snow submodel of the Biosphere-Atmosphere Transfer Scheme (BATS), which calculates snowmelt from the en-



Cryosphere Models, Figure 1  
The snowpack energy balance as characterized by SNTHERM (US Army Corps of Engineers)

ergy budget and snow temperature by the force–restore method. SNTHERM gave the best match to observations with the SAST simulation being close. BATS captured the major processes in the upper layers of a snow pack where solar radiation is the main energy source and gave satisfactory seasonal results.

CROCUS is a model of the Centre d'Etudes de la Neige, Grenoble [10]. It is a 1-D physical model that determines mass and energy balance for a snow cover and is used for operational avalanche forecasting. The snow cover is represented as a pile of layers parallel to the ground. Energy exchanges are projected orthogonally to the slope. The model describes the evolution of the internal state of the snow cover as a function of meteorological conditions. The variables describing the snow cover are temperature, density, liquid water content, and snow type of each layer. To match the natural layers, the thickness and number of layers are adjusted by the model. The model simulates the heat conduction, melting/refreezing of snow layers, settlement, metamorphism, and percolation. It simulates dry and wet snow metamorphism with experimental laws derived from laboratory data. Snow grains are characterized by their size and type. This allows an accurate albedo of the snow cover to be calculated.

Bartelt and Lehning [7] also present a 1-D physical model of the snow pack (SNOWPACK) with equations for heat transfer, water transport, vapor diffusion and mechanical deformation. New snow, snow drift and ablation are treated. The snow layers are treated in terms of height,

density and microstructure (grain size, shape and bonding). The model is used for avalanche warnings in Switzerland.

### Interception Models

A physically-based snowfall interception model that scales snowfall interception processes from branch to canopy is now available [25]. It takes account of the persistent presence and subsequent unloading of intercepted snow in cold climates. To investigate how snow is intercepted at the forest stand scale, measurements of wind speed, air temperature, above- and below-canopy snowfall, accumulation of snow on the ground and the load of snow intercepted by a suspended, weighed, full-size conifer were collected from spruce and pine stands in the southern boreal forest. Interception efficiency is found to be particularly sensitive to snowfall amount, canopy density and time since snowfall. Further work resulted in process-based algorithms describing the accumulation, unloading and sublimation of intercepted snow in forest canopies (Pomeroy et al. [70]). These algorithms are unique in that they scale up the physics of interception and sublimation from small scales, where they are well understood, to forest stand-scale calculations of intercepted snow sublimation.

### Blowing Snow Models

Physically-based treatments of blowing snow and wind are used to develop a distributed model of blowing snow

transport and sublimation over complex terrain for an Arctic tundra basin by Essery et al. [18]. A reasonable agreement with results from snow surveys is obtained when sublimation processes are included. Sublimation typically removes 15–45% of the seasonal snow cover. The model is able to reproduce the distributions of snow mass, classified by vegetation type and landform, which can be approximated by lognormal distributions. The representation used for the downwind development of blowing snow with changes in wind speed and surface characteristics is shown to have a moderating influence on snow redistribution.

Spatial fields of snow depth have power spectra in one and two dimensions that occur in two frequency intervals separated by a scale break between 7 and 45 m [84]. The break in scaling is controlled by the spatial distribution of vegetation height when wind redistribution is minimal and by the interaction of the wind with surface concavities and vegetation when wind redistribution is dominant.

In mountainous regions, wind plays a prominent role in determining snow accumulation patterns and turbulent heat exchanges, strongly affecting the timing and magnitude of snowmelt runoff. Winstral and Marks [90] use digital terrain analysis to quantify aspects of the upwind topography related to wind shelter and exposure. They develop a distributed time-series of snow accumulation rates and wind speeds used to force a distributed snow model. Terrain parameters were used to distribute rates of snow accumulation and wind speeds at an hourly time step for input to ISNOBAL, an energy and mass balance snow model. ISNOBAL forced with accumulation rates and wind fields generated from the terrain parametrizations accurately models the observed snow distribution (including the formation of drifts and scoured wind-exposed ridges) and snowmelt runoff. By contrast, ISNOBAL forced with spatially constant accumulation rates and wind speeds taken from the sheltered meteorological site at Reynolds Mountain in southwest Idaho, a typical snow-monitoring site, overestimated peak snowmelt inputs and tended to underestimate snowmelt inputs prior to the runoff peak.

### Avalanche Models

Avalanches range in size from sluffs with a volume of  $< 10 \text{ m}^3$  to extreme releases of  $10^7$ – $10^8 \text{ m}^3$ ; corresponding impact pressures range from  $< 10^3$ – $10^6 \text{ Pa}$ . There are two main types – loose snow avalanches and slab avalanches. Commonly, they begin with the failure of snow layers with densities less than  $300 \text{ kg m}^{-3}$ . An avalanche path comprises a starting zone, the track, and a runout–deposition

zone. Loose snow avalanches are initiated when the angle of repose is exceeded – about  $45^\circ$ . The angle increases as temperatures rise due to increased cohesion. Slush avalanches can occur on slopes  $< 10^\circ$ . Downslope propagation continues to the kinetic angle of repose at about  $17^\circ$  [52,69]. Slab avalanches occur when a cohesive slab is released over an extensive plane of weakness on slopes of  $35$ – $40^\circ$ . Slab thicknesses are  $0.1$ – $4 \text{ m}$  and have a mean density of  $\sim 200 \text{ kg m}^{-3}$ . Bed surface temperatures are near  $0^\circ\text{C}$ .

The variables of interest for forecasting are velocity, run-out distance and impact pressure. The acceleration of an avalanche is resisted by surface friction, air drag at the front and upper boundary, and ploughing at the advancing front and underneath surface. According to Perla [69], maximum velocities range from  $20$ – $30 \text{ m s}^{-1}$  for path lengths up to  $500 \text{ m}$  and slope angles of  $25$ – $35^\circ$ . The mean run-out length on 67 Colorado avalanche paths was  $380 \text{ m}$  (Bovis and Mears [9]). On occasion, the run-out may cross a valley floor and continue up the facing slope. Impact pressures are a maximum at  $1$ – $2 \text{ m}$  above the surface and range in value from about  $1$ – $10 \times 10^5 \text{ Pa}$  (Perla [69]).

### Land Surface Schemes in GCMs

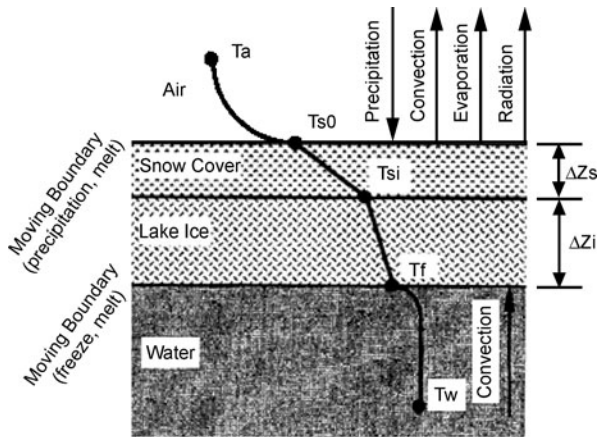
Snow cover is treated in Land Surface Models (LSMs), but snow and ice albedo parametrizations differ widely in their complexity [5]. A Snow Model Intercomparison was conducted using 24 snow cover models developed in ten different countries [17]. The models differ as to being multi-layer or not, the inclusion of a soil model, variable heat conductivity, variable snow density, and the treatment of liquid storage. Only four of the models met all five criteria.

27 atmospheric general circulation models (GCMs) were run under the auspices of the Atmospheric Model Intercomparison Project (AMIP). The AMIP models reproduce a seasonal cycle of snow extent similar to the observed cycle. However, GCMs tend to underestimate autumn and winter snow extent (especially over North America) and overestimate spring snow extent (especially over Eurasia). The majority of models displays less than half of the observed interannual variability. No temporal correlation is found between simulated and observed snow extent, even when only months with extremely high or low values are considered [20]. The second generation AMIP-II simulations gave better results [21].

### Floating Ice

Lake ice formation is dependent on the density characteristics of fresh water, which reaches a maximum den-





Cryosphere Models, Figure 2  
The elements of a model of lake ice [47]

sity at 4°C. As a water body cools in the autumn it becomes isothermal at 4°C. Further cooling of the surface allows a less dense layer to form and eventually frazil ice or sheet ice forms depending on the wind conditions. Snow accumulation on lake ice depresses the ice surface below the water level, causing the snow to become saturated and leading to the formation of white snow-ice (in contrast with the black lake water ice). In rivers the flow motion leads to frazil ice, which builds up into pancakes. A 1-D energy balance model of lake ice growth is described by Liston and Hall [47] that treats lake-ice freeze-up, break-up, total ice thickness and ice type (Fig. 2). The model is forced by daily atmospheric data on precipitation, wind speed and air temperature.

Sea ice grows thermodynamically by freezing of sea water at near  $-1.8^\circ\text{C}$  due to the salinity, by the accumulation of snow cover on its surface, and by dynamic processes such as ridging and rafting. It decays thermodynamically, by wave action and by export to areas of warmer ocean. Ice types include new ice, young ice, first year and multi-year ice (World Meteorological Organization [91]). Typical first year ice thickness in the spring in the Arctic is about 1.5–2 m; multi year unridged ice may be 3–4 m thick. Ridging produces keels that may extend to 20–30 m depth. Ice draft (below the sea surface) is measured by upward looking sonars that are moored or deployed on submarines in the Arctic. Ice extent and concentration – the fractional coverage of ice – are determined by aerial reconnaissance, and primarily by satellite remote sensing (optical, passive microwave, synthetic aperture radar, scatterometry and laser altimetry). Weekly or 10-day charts of ice conditions are produced by national operational ice services (see [94]).

Modeling sea ice in either a stand-alone model or a GCM involves the solution of the following equations [19]:

- for momentum, to obtain the ice velocity fields;
- for thermodynamic processes to obtain net ice growth/melt; and
- conservation equations including deformation and transport of ice, plus the thermodynamic sources and sinks.

Ice dynamics is based on five stresses: wind stress, water stress, internal ice stress, Coriolis force, and the stress from the tilt of the sea surface. The Coriolis force and the tilt term are an order of magnitude less than the other three terms. The air and water stresses assume a constant turning angle of  $25^\circ$  in the Arctic and  $-25^\circ$  in the Antarctic [28]. Internal ice stress is highly variable depending on ice conditions. It can be negligible when the ice cover is not compact and there are “free-drift” conditions, but it can be the largest force when there is thick, compact ice cover. The force due to ice resistance to deformation involves the relationship between stress and strain rate, which is termed the rheology (Flato [19]). Early work assumed that stress is linearly dependent on strain rate as in a linear viscous fluid [12]. Pritchard et al. [71] used an elastic-plastic rheology where the stress is linearly dependent on strain up to a yield strength where failure occurs. Hibler [27] developed a viscous-plastic model with an elliptical yield curve; the pre-yield stress states are linearly related to the strain rate. Advances by Hunke and Dukowicz [35] address the response of the ice on the timescales associated with wind forcing through an elastic viscoplastic (EVP) rheology. The model was modified so that it reduces to the viscous-plastic model at these timescales, whereas at shorter timescales the adjustment process takes place by a mathematically efficient elastic wave mechanism. Recently, Lagrangian [33,45] sea ice models using a granular rheology have been developed (Tremblay and Mysak [83]; Overland et al. [65]). They have advantages in that they model individual sea ice “floes”, but are also computationally intensive and are still in their infancy.

Ice dynamics have been extensively treated by Hibler [27]. He couples the dynamics to the ice thickness characteristics by allowing the ice interaction to become stronger as the ice becomes thicker and/or contains a lower area percentage of thin ice. The dynamics in turn causes high/low oceanic heat losses in regions of ice divergence/convergence. The ice is considered to interact in a plastic manner with the plastic strength depending on the ice thickness and concentration. These in turn evolve according to continuity equations that in-

clude changes in ice mass and percent of open water due to advection, ice deformation and thermodynamic effects. Anisotropic dynamic behavior of sea ice has also been investigated [13,30], though such approaches are computationally intensive and currently are not commonly used in models. The standard model treats sea ice as a viscoplastic material that flows plastically under typical stress conditions but behaves as a linear viscous fluid where strain rates are small and the ice becomes nearly rigid. The standard viscous–plastic model has poor dynamic response to forcing on a daily timescale. Models do not generally account for high-frequency (sub-daily) inertial and tidal effects on dynamics, though research has shown that such effects can be important in the evolution of the ice cover [26,43]. The thermodynamics and dynamics are coupled through the ice thickness distribution. Essentially, deformation leads to pressure ridging and the formation of open water areas while thermodynamic processes act to ablate ridges and remove open water by ice formation in winter and create thinner ice/open water in summer. Thus, deformation acts to spread out the thickness distribution by promoting thick and thin ice categories while thermodynamic processes work towards a central ice thickness value [28].

Sea ice models typically feature processes of ice thermodynamics and dynamics although the earliest studies essentially used only thermodynamic ice growth and decay. The steady state Stefan relationship is written after Hibler and Flato [29]:

$$\rho_I L \frac{dH}{dt} \approx \frac{k_i}{H} (T_m - T_B),$$

where  $L$  = latent heat of fusion,  $T_m$  = melting point of the ice,  $T_B$  = upper boundary temperature of the ice,  $H$  = ice thickness,  $k_i$  = ice conductivity and  $\rho_I$  = ice density. Ice growth/melt at the underside is a result of the difference between the upward ocean heat flux and the heat conducted away from the ocean/ice interface into the ice. The first 1-D model of sea ice thermodynamics was developed by Maykut and Untersteiner (1971) [51]. A fuller treatment was made by Parkinson and Washington [66]. The model had four layers – ice, snow, ocean, and atmosphere – and 200 km horizontal resolution.

The incorporation of detailed thermodynamic processes includes the presence of snow on the sea ice, leads and polynyas, melt ponds, the effect of internal brine-pocket melting on surface ablation, the storage of sensible and latent heat inside the snow-ice system, and the transformation of snow into slush ice when the snow-ice interface sinks below the waterline due to the weight of snow. Models with enthalpy conservation improve the thermo-

dynamic component of sea ice models [8]. These are starting to be included in larger-scale climate models.

An intermediate one-dimensional thermodynamic sea ice model developed by Ebert and Curry [15] includes leads and a surface albedo parametrization that interacts strongly with the state of the surface, and explicitly includes meltwater ponds (see Fig. 3). Four important positive feedback loops were identified: (1) the surface albedo feedback, (2) the conduction feedback, (3) the lead-solar flux feedback, and (4) the lead fraction feedback. The destabilizing effects of these positive feedbacks were mitigated by two strong negative feedbacks: (1) the outgoing longwave flux feedback, and (2) the turbulent flux feedback. A review of thermodynamic models is given by Steele and Flato [80].

Conservation equations are needed for ice area (concentration) and ice volume (thickness).

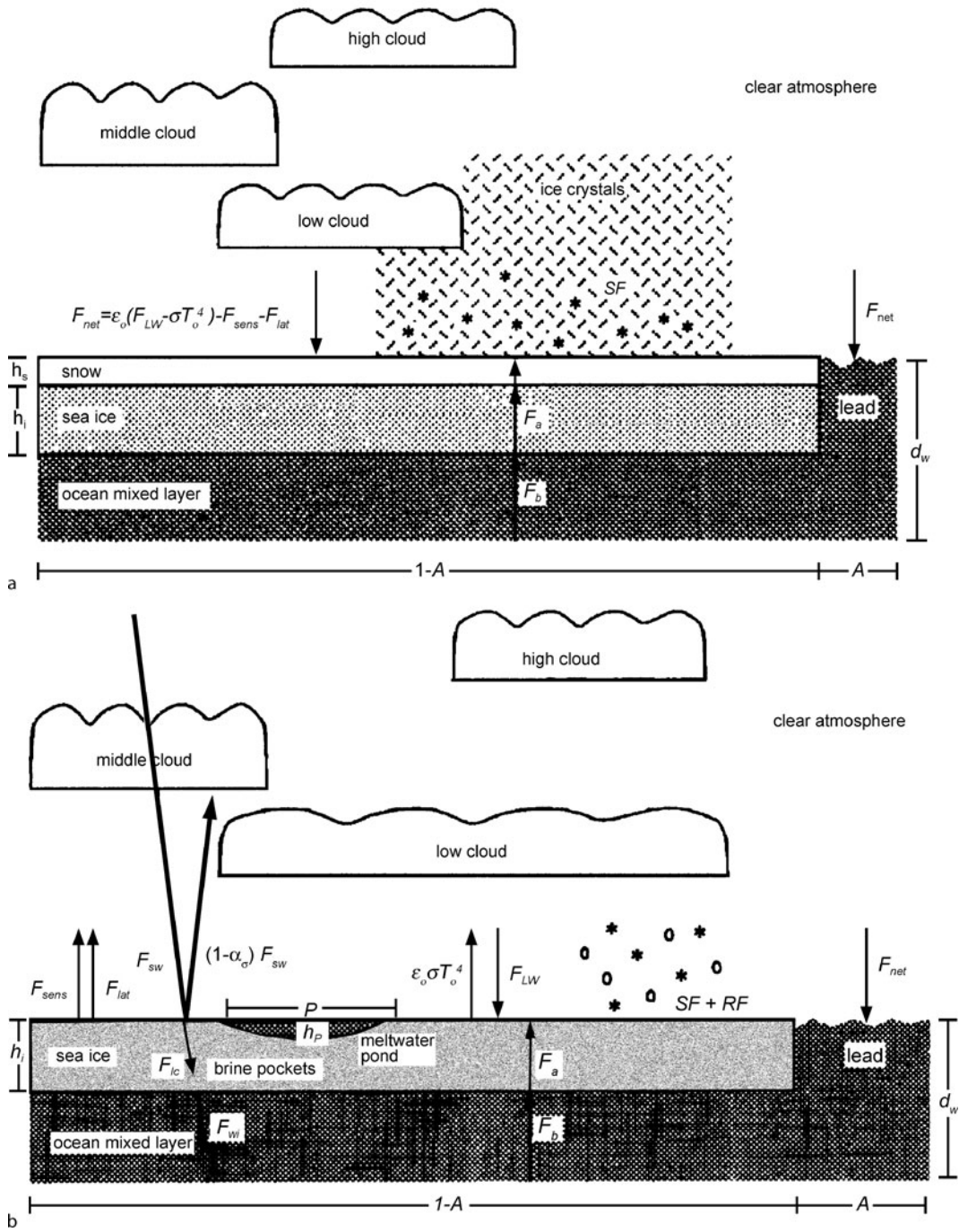
$$\begin{aligned} \partial h / \partial t &= -\nabla \bullet (\mathbf{u}h) + S_h \\ \partial A / \partial t &= -\nabla \bullet (\mathbf{u}A) + S_A, \end{aligned}$$

where  $\mathbf{u}$  is the ice velocity vector and  $S_h$  and  $S_A$  are source terms for mean ice thickness and concentration, respectively. The second equation must also have the constraint that the area  $A \leq 1$ . The ice strength is parametrized only in terms of ice thickness  $h$  and  $A$  [27].

An atmospheric GCM was coupled to a global 1-degree, 20-level ocean GCM with dynamic and thermodynamic sea ice by Washington and Meehl [86] and run with increasing atmospheric  $\text{CO}_2$ . The Coupled Model Intercomparison Project (CMIP) allows a comparison of predicted Arctic sea ice [53]. Of the 12 models, only seven include sea ice motion and only four of these have a prognostic solution to the momentum equation. Apart from errors and approximations in the sea ice representation, the models also suffer from errors in the atmospheric and oceanic forcing fields. While the northern hemisphere ice extent in winter is well simulated overall, the ice thickness does not capture the proper spatial distribution with thicker ice toward North America and Greenland and thinner ice in the Eurasian basin. The simulations for the southern hemisphere show a wider range of extents and thickness. Flato [19] examines the sea ice extent simulated by two GCMs for AD 1900–2100 with the ‘business as usual’ scenario of greenhouse gases and aerosol concentrations. Both show a progressive decrease in ice in both hemispheres although the two models differ significantly in the initial southern hemisphere ice extent.

Martin and Gerdes [50] make a comparison of sea ice drift results from different Arctic Ocean Model Intercomparison Project (AOMIP) sea ice-ocean coupled models

Elbert and Curry: One-dimensional Thermodynamic Sea Ice Model



Cryosphere Models, Figure 3  
 The configuration of a one-dimensional thermodynamic model of sea ice a winter; b summer [15]

and observations for 1979–2001. The models are capable of reproducing realistic drift pattern variability. However, one class of models has a realistic mode at drift speeds around  $3 \text{ cm s}^{-1}$  and a short tail toward higher speeds. Another class shows unrealistically a more even frequency distribution with large probability of drift speeds of 10 to  $20 \text{ cm s}^{-1}$ . Reasons for these differences lie in discrepancies of wind stress forcing as well as sea ice model characteristics and sea ice-ocean coupling. Hunke and Holland [36] underscore the sensitivity of Arctic sea ice and ocean to small changes in forcing parameters. A comparison of three sets of forcing data, all variants of National Centers for Environmental Prediction (NCEP) forcing, give significant differences in ice thickness and ocean circulation using a global, coupled, sea ice-ocean model.

A study of GCMs used for the IPCC Fourth Assessment Report shows that while they produce reasonably similar ice extents in the Arctic, their equilibrium ice thickness values have a wide range due to differences in downwelling infrared radiation [16]. Holland et al. [32] found that in some scenarios of future  $\text{CO}_2$  concentrations the sea ice cover can respond non-linearly with large decreases in extent within only 5–10 years, indicating that the current fitted to observations linear trends may not hold in the future. Stroeve et al. [81] showed that the IPCC models substantially underestimate the observed decline in Arctic sea ice extent compared to observations over the past 50 years. Hence their application in future scenarios is questionable.

Johnson et al. [40] examine the simulated sea ice concentration from nine ice-ocean numerical models in the AOMIP. The models have similar characteristics in winter (100% cover is produced), and most models reproduce an observed minimum in sea ice concentration for September 1990.

An assessment of coupled climate models with respect to the development of Arctic sea ice thickness during the 20th century is examined by Gerdes and Koerberle [22]. Model behavior is compared with results from an ocean–sea ice model using the AOMIP atmospheric forcing for the period 1948–2000. The hindcast exhibits virtually no trend in Arctic ice volume over its integration period 1948–2000. Most of the coupled climate models show a negative trend over the 20th century that accelerates towards the end of that century.

## Glaciers

Glaciers are built up from snow that persists over many years. Initial densification leads to firn (densities of  $400\text{--}830 \text{ kg m}^{-3}$ ) and at some depth, where the air passages be-

tween grains are sealed off ( $\sim 15\text{--}70 \text{ m}$  according to wetness), to glacier ice with a density of  $830\text{--}917 \text{ kg m}^{-3}$  [67]. The glacier has upper accumulation and lower ablation areas, that are annually varying, and the ice slowly flows downhill towards the glacier terminus. Some glaciers occasionally display surges when the ice advances rapidly for a year or two and then stabilizes or retreats.

Glacier models consider either the mass balance and the rate of change of total mass, or the glacier dynamics and interactions between the ice and the bed. The flow velocity is modeled along the centerline of the glacier.

Glacier flow is determined from a relation between the shear strain rate ( $\epsilon_{xy}$ ) and shear stress ( $\tau_{xy}$ ) known as Glen’s flow law [23]:

$$\dot{\epsilon}_{xy} = A \tau_{xy}^n,$$

where  $n \sim 3$ .  $A$  depends on ice temperature, impurities and crystal orientation.

Recommended values of  $A$  decrease from  $6.8 \times 10^{15} \text{ s}^{-1} \text{ kPa}^{-3}$  at  $0^\circ\text{C}$  to  $3.6 \times 10^{-18}$  at  $-50^\circ\text{C}$  (see Table 5.2 in [67]). Stress causes ice to deform by extension/compression, and shear leading to rotation.

Ice flows only by internal deformation when the bed is frozen, but where temperate conditions exist at the base, sliding becomes important. The sliding law relates basal velocity, shear stress, water pressure and the glacier bed characteristics. Weertman’s [88] theory of sliding involved regelation and plastic deformation. Regelation operates over small bumps in the bed ( $< 1 \text{ m}$  dimension). All the ice is at pressure melting point. There is excess pressure on the upstream side of the bump so that the ice there is colder than on the downstream side. This causes heat to flow towards the upstream side through the bump and surrounding ice. The heat transferred melts ice on the upstream side and melt water flows around the bump, refreezing on the downstream side because it is colder than the ice there. Ice also deforms plastically. Near a bump, the longitudinal stress in the ice and, therefore, the strain are above average. The greater the distance over which the stress is enhanced, the greater is the ice velocity. This mechanism works best over larger bumps. Both processes are equally effective at the “controlling obstacle” size, about 1–10 cm.

Paterson [67] shows that the sliding velocity:

$$u = \text{constant} (\tau^{0.5}/R)^4,$$

where  $R$  = roughness and  $\tau$  = basal shear stress.

Water from surface ablation penetrates to the glacier bed and has been shown to lift the ice by as much as 40 cm on the Unteraargletscher, Switzerland (Iken et al. [38]). During rapid uplift events the glacier velocity increased 3–

6 times. When the water pressure at the bed exceeds a certain value (the separation pressure) that depends on the bed roughness, cavities form in the lee of bumps. When the water pressure exceeds a second critical value, sliding becomes unstable.

A numerical ice flow model has recently been used to study the advance of tidewater glaciers into a deep fiord [56]. The results suggest that irrespective of the calving criterion and the accumulation rate in the catchment, the glacier cannot advance into deep water (>300 m) unless sedimentation at the glacier front is included.

Using a first-order theory of glacier dynamics, Oerlemans [61] related changes in glacier length to changes in air temperature. He constructed a temperature history for different parts of the world from 169 records of glacier length. The reconstructed warming in the first half of the 20th century is 0.5°C. The warming signals from glaciers at low and high elevations appear to be very similar.

## Ice Sheets

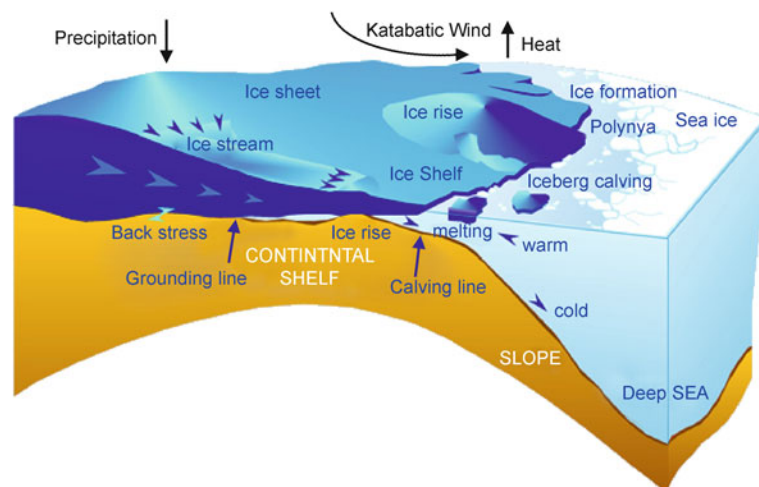
There are currently two major ice sheets in Greenland and Antarctica, but during the Last Glacial Maximum about 20,000 years ago there were massive ice sheets over northern North America and Fennoscandia. The Antarctic ice sheet covers 12.4 million km<sup>2</sup> and reaches thicknesses in excess of 4500 m; it represents a sea level equivalent of 64.3 m. The Greenland ice sheet has an area of 1.8 million km<sup>2</sup> and a maximum thickness of 3200 m; it has a sea level equivalent of 7.5 m. Both ice sheets have a parabolic profile and gentle slopes (1–2°) away from the margins.

Greenland has a suite of snow facies; from the interior outward these are: the dry-snow zone, percolation zone, wet-snow zone, superimposed ice zone, and the ablation area. The equilibrium line is between the ablation area and the higher zones. At its base, Greenland is close to sea level except for fringing coastal mountains, through which the ice reaches the sea in some 20 major outlet glaciers. Antarctica is mostly dry snow and is little affected by melting except at the margins and in the Antarctic Peninsula. In the Antarctic there are major ice shelves (e. g. the Ross Ice Shelf and the Filchner–Ronne Ice Shelf) that buttress large sections of the ice sheet (see Fig. 4) and extend around 44% of the coastline of the continent.

The major problems in ice sheet dynamics, following Paterson (see p. 238 in [66]), are (i) to calculate the distribution of ice thickness and velocity that will maintain a steady state, given the accumulation and ablation rates and surface temperature. The flow parameters, ice thickness at the ice divide, and geothermal heat flux must be specified. Flow lines, the time that ice takes to travel along them, and the age of the ice at different depths, can be calculated from the velocity distribution. (ii) to determine how the system will react to changes in accumulation, ablation or surface temperature.

The earliest ice sheet models assumed that ice deformed as a Newtonian viscous body. Orowan [64] and Nye [59] assumed that ice behaved as a perfectly plastic material, but Glen [23] established the relationship between strain rate and stress in ice as non-linear (see above).

The surface profile of an ice sheet on a horizontal bed of half width  $L$ , thickness  $h$  and thickness at the centre  $H$



Cryosphere Models, Figure 4  
Schematic of an ice sheet and shelf showing the processes at work (after Wikipedia)

is:

$$h^2 = \frac{2\tau_0}{\rho g}(L - x),$$

where  $\rho$  = density,  $g$  = acceleration due to gravity,  $\tau_0$  = the basal shear stress ( $\sim 0$ –100 kPa), and  $(L - x)$  is the distance from the edge measured along a flow line. The equation describes a parabola. The ice thickness at the centre is  $H = (2\tau_0 L / \rho g)^{0.5}$ .

The mass balance ( $B$ ) can be expressed in a mass conservation equation as:

$$B = \frac{\partial q}{\partial x} + \frac{\partial h}{\partial t},$$

where  $h$  = ice thickness,  $t$  = time, and the flux  $q = h \mathbf{u}$ , where  $\mathbf{u}$  is the velocity averaged over the ice thickness (see p. 246 in Paterson [67]).

If the flow term  $\partial q / \partial x$  is small, then the surface elevation will vary in response to the local accumulation/ablation, which will determine the profile. ‘Balance velocities’ are steady state velocities that are calculated from accumulation rate and ice thickness. Paterson [67] shows that for a 1000 km radius circular ice sheet, of perfectly plastic ice with a yield stress of 199 kPa, an accumulation rate of 150 mm of ice/year, and ablation by ice-berg calving at the margin, the ice in the centre would be 4700 m thick. Balance velocities would increase from 1.5 m  $\text{a}^{-1}$  at 100 km from the centre to 45 m  $\text{a}^{-1}$  at 900 km and the travel time for ice to move from the center to the edge would be 150,000 years.

In Greenland and Antarctica much of the ice transport is accomplished by fast flowing ice streams – regions where the ice flow is much faster than on either side. Most occupy deep channels with beds below sea level and terminate either as a floating glacier tongue (an outlet glacier) or become part of an ice shelf. Morgan et al. [54] indicate that while ice streams and outlet glaciers account for only 13% of the coastline of Antarctica, they drain about 90% of the accumulation of the interior.

A hierarchy of land ice models is presented by van der Veen and Payne [85]. The simple lamellar flow model, involves a balance between driving stress and basal drag. The surface and bed topography must be nearly level for lamellar flow, which is a good approximation to conditions in the interior of an ice sheet. In cases where an ice stream is bounded by a rock wall or stagnant ice on one or both sides, lateral drag needs to be incorporated. The proportion of driving stress that is supported by drag at the bed is termed the shape factor [60]; it is less than one for narrow ice streams. An important issue in Antarctica is the inter-

action between the ice sheet and ice shelves. The peripheral ice shelves are thought to exert a back stress that stabilizes the inland ice sheet where it is grounded below sea level, as in most of West Antarctica [82]. Where the bedrock slopes down towards the ice sheet interior, the grounding line is unstable; If the grounding line initially retreats, the ice at the grounding line becomes thicker due to the bedrock slope, and the creep thinning (thinning associated with along-flow gradients in the ice velocity) increases causing the grounding line to retreat further – a positive feedback. The dynamics of ice sheet grounding lines is examined by Schoof [76]. A boundary layer theory for ice flux through the transition zone shows that the flux increases sharply with ice thickness at the grounding line. He finds that marine ice sheets have well-defined, discrete equilibrium profiles, and steady grounding lines cannot be stable on reverse bed slopes. Also, marine ice sheets with over-deepened beds may undergo hysteresis with variations in sea level, accumulation rate, bed slipperiness and ice viscosity.

Two general types of an ice sheet model have been developed. One is prognostic, based on the original work by Budd et al. [11]; the other category is diagnostic, addressing specific aspects of ice sheet processes. Prognostic models involve four sets of equations (van der Veen and Payne [85]). These are: (i) diagnostic equations for the horizontal velocity components as functions of local ice geometry and ice rheology (Glen’s law); (ii) prognostic equations for the evolution of internal ice temperature, given appropriate boundary conditions at the upper and lower ice surfaces; (iii) a diagnostic equation for ice vertical velocity via the divergence of the horizontal velocity; and (iv) a prognostic equation for ice thickness based on the snow accumulation, snow/ice melt and the divergence/convergence of horizontal ice flow. The effects of bedrock depression under the changing weight of the ice load must also be taken into account. Such models have been used to reconstruct ice sheet history over glacial cycles, as well as to assess the responses to future climate change.

Diagnostic models do not address time evolution of the ice sheet and treat the internal stress regime in much greater detail, particularly the contributions of longitudinal and lateral stresses. Recently, models have been developed that do not assume negligible vertical shear. Huybrechts and de Wolde [38] have combined prognostic model elements with a detailed diagnostic model to study the multi-century behaviour of the Antarctic and Greenland ice sheets. A fully dynamic 3-D thermo-mechanical ice sheet model was coupled to a two-dimensional climate model.

A model validation exercise was undertaken by the European Ice Sheet Modeling Initiative (EISMINT). Payne et al. [68] examined the effects of thermo-mechanical coupling while MacAyeal et al. [49] test ice shelf models for the Ross Ice Shelf. Overall, the models agreed in the main features that were simulated.

### Frozen Ground and Permafrost

The surface layers of soil and rock may be seasonally or perennially frozen. Perennially frozen ground or permafrost is frozen for at least two successive summers. The ground need not contain ice but may be rock below 0°C. Ground ice may be segregated, in veins, or massive in occurrence. The spatial extent of permafrost ranges from continuous (>90% of the surface is underlain), to discontinuous (50–90% of the surface), sporadic (10–50%) and isolated (<10%). Continuous permafrost is associated with mean annual temperatures below about -7°C. Its thickness ranges from a few meters up to 1500 m in Yakutia. Subsea permafrost also occurs offshore in the Eurasian shelf seas and in the Beaufort Sea; it is a relic of previous glacial intervals when the sea bed was exposed by sea level lowering of up to 135 m under low temperature conditions.

Ground temperatures are largely determined by heat conduction, although in areas of seasonal freezing and discontinuous permafrost localized circulation of groundwater may need to be considered. The thermal properties of the ground vary with the mineral composition, organic content, moisture content (as vapor, water and ice), and temperature, as well as the overlying vegetation and snow cover. A frozen soil algorithm has been developed [92] to detect the near-surface soil freeze/thaw cycle over snow-free and snow-covered land in the United States (see Fig. 5).

The conductive heat transfer is given [89] as:

$$G = -K(dT/dz)$$

where  $K$  = the thermal conductivity ( $W m^{-1} K^{-1}$ ).

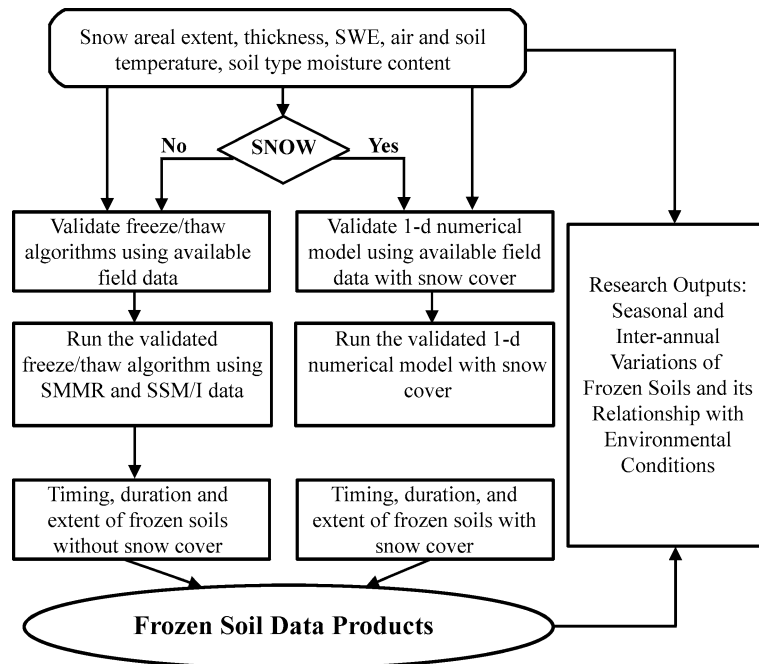
For steady state conditions, the temperature at depth  $T_z$  is written:

$$T_z = T_s + (G/K)z$$

where  $T_s$  = surface temperature.

The heat conduction equation is:

$$\partial T/\partial t = \kappa \partial^2 T/\partial z^2$$



Cryosphere Models, Figure 5

The flowchart of a model processing the effects of snow cover on frozen soil, used to study the timing, duration, number of days, and areal extent of near-surface soil freeze/thaw status [92]

where  $\kappa$  = thermal diffusivity ( $\text{m}^2 \text{s}^{-1}$ ), the coefficient of heat diffusion. To simulate soil freezing and thawing processes, soil water phase change has to be considered (for details, see Lunardini [48]).

Permafrost models can be broadly classified as either equilibrium or process-based transient models according to their underlying methodology. Equilibrium models are based on empirical and semi-empirical relationships between permafrost occurrence and topoclimatic factors (altitude, slope, aspect) and mean air temperature, freezing and thawing indices, and snow cover. They are often used to predict the lateral “boundaries” of permafrost distribution, or to estimate “average” geocryological parameters, thawing indices, snow cover and solar radiation, for example [77]. A ‘frost index’ model was developed by Nelson and Outcalt [55] for the Arctic and has been applied to mountain areas [31]. A wide spectrum of explicit equilibrium permafrost models is available to estimate the thickness of the active layer. The simplest approaches are based on several variations of the analytical Stefan solution to the heat conduction problem with phase change. These methods have been used to estimate regional-scale active layer thickness [56,79]. Kudryavtsev et al. [42] developed a more comprehensive equilibrium permafrost model with analytical solutions that has been adapted and used with Geographical Information System (GIS) technology to estimate active layer thickness at regional [74,78] and circum-arctic [2] scales. The advantages of equilibrium models are their relative simplicity and low data requirements. The major drawback of such models is their inability to resolve seasonal and inter-annual variability [77], which is frequently required for ecological and hydrological studies in the Arctic. These limitations have led to recent spatial adaptations of process-based transient numerical models.

Process-oriented transient models detail one-dimensional heat transfer in soils with phase change driven by either surface temperature [24] or the energy balance components [46]. They account for the major physical processes governing development of the ground thermal regime, simulate soil freezing/thawing processes, and provide insight into the response of soil thermal regime to changes in environmental conditions. These models can provide good results for simulating active layer thickness and permafrost temperatures when driven with known boundary conditions and forcing parameters measured at site-specific locations [73,93]. Recently, such one-dimensional heat transfer models with phase change have been used to simulate regional-scale soil thermal regime [62,63]. However, their adaptation from point-specific to regional-scale is not a straightforward process. It

requires simplification, careful selection of climate forcing data, and treatment of surface and subsurface parameters with largely unknown distributions over the modeled domain.

Most GCMs do not treat permafrost dynamics. However, Nikolsky et al. [58] show that in the Community Land Model (CLM3) GCM improvements can be made to the representation of permafrost dynamics and their climate feedbacks. They do this by increasing the total soil depth by adding new layers, incorporating a surface organic soil layer, and modifying the model’s numerical scheme to include unfrozen water dynamics and more realistic treatment of the model phase changes between ice and water.

The Community Climate System Model (CCSM) has a 5-layer snow model over a 10 layer 3.4 m deep soil model that treats thermal and hydrologic frozen soil processes. A projection made for the 21st century shows severe degradation of the permafrost in the Northern Hemisphere.

Some models address only the active layer that is the top layer of soil that thaws during the summer and freezes again during the autumn. The thaw depth can be analyzed by the Stefan solution for heat transfer in a medium with phase change (Anisimov et al. [3]):

$$z = [(2n\lambda tT)/(\rho wL)]^{0.5}$$

where  $z$  = active layer thickness (m),  $n$  = the ratio of seasonal ground surface and air temperature degree-day sums,  $\lambda$  = thermal conductivity of thawed soil ( $\text{W m}^{-1} \text{K}^{-1}$ ),  $t$  = warm season duration  $>0^\circ\text{C}$  (s),  $T$  = mean warm season temperature ( $^\circ\text{C}$ ),  $\rho$  = soil density ( $\text{kg m}^{-3}$ ),  $w$  = relative water content (decimal proportion) and  $L$  = latent heat of fusion ( $\text{J kg}^{-1}$ ).

A dynamic 3-D terrain model is currently being developed and tested in Svalbard (Humlum [34]). The model takes topographic data, terrain surface characteristics (geomorphology and vegetation) and meteorological variables (air temperature, wind speed and direction, and cloud cover) as input and provides output on phenomena such as terrain surface net radiation balance, snow cover thickness and duration, glacier mass, active layer thickness, stable permafrost thickness and the amount of summer melt water discharge.

A stochastic model was developed by Anisimov et al. [3] and used to calculate the probability density function of active-layer thickness (ALT). Equations for the mean, variance, and higher moments of ALT were derived by applying stochastic averaging to a semi-empirical model of seasonal thawing. The stochastic model was ap-



plied in a case study in the Kuparuk River basin, north-central Alaska.

Shiklomanov et al. [77] compare three models of active layer thickness (ALT) for northern Alaska. One model (NSIDC) is very accurate in the topographically homogeneous Coastal Plain but overestimates (ALT) in the Brooks Range Foothills. The UAF-GIPL 2.0 model reproduced site-specific active layer values well but overestimated ALT on the Coastal Plain. Large differences in ALT fields mainly result from differences in model approaches for characterizing largely unknown spatial distribution of surface (vegetation, snow) and subsurface (soil properties and moisture) conditions.

Data set limitations are a major problem (Anisimov et al. [3]). A permafrost model, forced with available climate data sets, was used to calculate the large-scale characteristics of permafrost in northern Eurasia. Zonal-mean air and ground temperatures, depth of seasonal thawing, and area occupied by near-surface permafrost in Eurasia north of 45° N were analyzed. The 0.5–1.0 °C difference in zonal-mean air temperature between the data sets translates into a 10–20% uncertainty in estimates of near-surface permafrost area, which is comparable to the extent of changes projected for the following several decades.

GCMs have been used to simulate changes in permafrost conditions with global warming. Anisimov and Nelson [1] were the first to study this. Most recently, Saito et al. [74] use a coupled global climate model at high horizontal resolution (0.5° land mesh) with a five-layer, 4.0 m deep soil to evaluate changes in the distribution of frozen ground and subsurface hydrothermal regimes under global warming. Two types of frozen ground were classified according to monthly soil temperatures: “permafrost” for regions with a maximum active layer thickness less than 4 m and “seasonally frozen ground.” Approximately 60% of present-day permafrost would degrade into seasonally frozen ground by 2100 in the circum-Arctic basins.

### Future Directions

In the next 5–10 years we can expect to see more comprehensive treatment of cryospheric processes in climate models. Already, steps are being taken to incorporate ice sheet processes and to enhance the treatment of frozen ground and permafrost. Increased model resolution will place new demands for cryospheric data sets for boundary conditions and as validation data. Mountain glaciers, as well as lake ice, will need to be incorporated especially in regional climate models.

## Bibliography

### Primary Literature

1. Anisimov OA, Nelson FE (1996) Permafrost distribution in the Northern Hemisphere under scenarios of climate change. *Glob Planet Chang* 14:59–72
2. Anisimov OA, Shiklomanov NI, Nelson FE (1997) Global warming and active-layer thickness: results from transient general circulation models. *Glob Planet Chang* 15:61–77
3. Anisimov OA, Shiklomanov NI, Nelson FE (2002) Variability of seasonal thaw depth in permafrost regions: a stochastic modeling approach. *Ecol Model* 153:217–227
4. Armstrong RL, Brodzik MJ, Knowles K, Savoie M (2005) Global monthly EASE-Grid snow water equivalent climatology. Digital media. National Snow and Ice Data Center, Boulder
5. Barry RG (1996) The parameterization of surface albedo for sea ice and its snow cover. *Progr Phys Geog* 20:61–77
6. Barry RG (2002) The role of snow and ice in the global climate system: A review. *Polar Geog* 24:235–246
7. Bartelt P, Lehning M (2002) A physical SNOWPACK model for the Swiss Avalanche Warning Services. Part I: Numerical model. *Cold Reg Sci Technol* 35(3):123–145
8. Bitz CM, Lipscomb WH (1999) An energy-conserving thermodynamic model of sea ice. *J Geophys Res* 105:15669–15677
9. Bovis MJ, Mears AI (1976) Statistical prediction of snow avalanche runout from terrain variables in Colorado. *Arct Alp Res* 8:115–120
10. Brun E, David P, Sudul M, Brunot G (1992) A numerical model to simulate snow cover stratigraphy for operational avalanche forecasting. *J Glaciol* 38:13–22
11. Budd WF, Jenssen D, Radok U (1971) Derived physical characteristics of the Antarctic ice sheet. ANARE Interim Report Series A (IV) *Glaciology Publ*
12. Campbell WJ (1965) The wind-driven circulation of ice and water in a polar ocean. *J Geophys Res* 70:3279–3301
13. Coon MD, Knoke GS, Echert DS, Pritchard RS (1998) The architecture of anisotropic elastic-plastic sea ice mechanics constitutive law. *J Geophys Res* 103(C10):21915–21925
14. Dozier J, Painter TH (2004) Multispectral and hyperspectral remote sensing of alpine snow properties. *Annu Rev Earth Planet Sci* 32:465–494
15. Ebert EE, Curry JA (1993) An intermediate one-dimensional thermodynamic sea ice model for investigating ice-atmosphere interactions. *J Geophys Res* 98(C6):10085–10110
16. Eisenman I, Untersteiner N, Wettlaufer JS (2007) On the reliability of simulated Arctic sea ice in global climate models. *Geophys Res Lett* 34:L10501, doi:10.1029/2007GL029914
17. Essery R, Yang Z-L (2001) An overview of models participating in the snow model intercomparison project (SnowMIP). In: 8th Scientific Assembly of IAMAS, Innsbruck. <http://www.cnrn.meteo.fr/snowmip/>. Accessed 22 Aug 2008
18. Essery R, Long L, Pomeroy JW (1999) A distributed model of blowing snow over complex terrain. *Hydrol Process* 13:2423–2438
19. Flato GM (2004) Sea-ice modelling. In: Bamber JL, Payne AJ (eds) *Mass balance of the cryosphere: Observations and modelling of contemporary and future change*. Cambridge University Press, Cambridge, pp 367–390

20. Frei A, Robinson DA (1995) Evaluation of snow extent and its variability in the Atmospheric Model Intercomparison Project. *J Geophys Res* 103(D8):8859–8871
21. Frei A, Miller JA, Robinson DA (2003) Improved simulations of snow extent in the second phase of the Atmospheric Model Intercomparison Project (AMIP-2). *J Geophys Res* 108(D12):4369, doi:10.1029/2002JD003030
22. Gerdes R, Koeberle C (2007) Comparison of Arctic sea ice thickness variability in IPCC Climate of the 20th Century experiments and in ocean–sea ice hindcasts. *J Geophys Res* 112(C4):C04S13
23. Glen J (1955) The creep of polycrystalline ice. *Proc Roy Soc Lond A* 228:519–538
24. Goodrich LE (1982) The influence of snow cover on the ground thermal regime. *Can Geotech J* 19:421–432
25. Hedstrom N, Pomeroy JW (1998) Measurements and modelling of snow interception in the boreal forest. *Hydrol. Processes* 12:1611–1525
26. Heil P, Hibler WD III (2002) Modeling the high-frequency component of Arctic sea ice drift and deformation. *J Phys Oceanogr* 32:3039–3057
27. Hibler WD III (1979) A dynamic-thermodynamic sea ice model. *J Phys Oceanogr* 9:815–846
28. Hibler WD III (2004) Modelling the dynamic response of sea ice. In: Bamber JL, Payne AJ (eds) *Mass balance of the cryosphere: Observations and modelling of contemporary and future change*. Cambridge University Press, Cambridge, pp 227–334
29. Hibler WD III, Flato GM (1992): Sea ice models. In: Trenberth K (ed) *Climate System Modeling*. Cambridge University Press, New York, pp 413–436
30. Hibler WD III, Schulson EM (2000) On modeling the anisotropic failure and flow of flawed sea ice. *J Geophys Res* 105(C7):17105–17120
31. Hoelzle M, Mittaz C, Eitzelmueller B, Haerberli W (2001) Surface energy fluxes and distribution models of permafrost in European mountain areas: An overview of current developments. *Permafrost Periglacial Process* 12:53–68
32. Holland MM, Bitz CM, Tremblay H (2006) Future abrupt reductions in the summer Arctic sea ice. *Geophys Res Lett* 33:L23503. doi:10.1029/2006GL028024
33. Hopkins MA (1996) On the mesoscale interaction of lead ice and floes. *J Geophys Res* 101:18315–18326
34. Humlum O (2007) Modeling energy balance, surface temperatures, active layer depth and permafrost thickness around Longyear dalen, Svalbard. [http://www.unis.no/research/geology/Geo\\_research/Ole/Modelling.htm](http://www.unis.no/research/geology/Geo_research/Ole/Modelling.htm). Accessed 22 Aug 2008
35. Hunke EC, Dukowicz JK (1997) An elastic–viscous–plastic model for sea ice dynamics. *J Phys Oceanogr* 27:1849–1867
36. Hunke EC, Holland MM (2007) Global atmospheric forcing data for Arctic ice–ocean modeling. *J Geophys Res* 112:C04S14
37. Huybrechts P, de Wolde J (1999) The dynamic response of the Greenland and Antarctic ice sheets to multiple-century climatic warming. *J Climate* 12:2169–2188
38. Iken A, Roethlisberger H, Flotron A, Haerberli W (1983) The uplift of the Unteraargletscher at the beginning of the melt season – a consequence of water storage at the bed. *J Glaciol* 30:15–25
39. Jin J, Gao X, Yang Z-L, Bales RC, Sorooshian S, Dickinson RE, Sun SF, Wu GX (1999) Comparative analyses of physically based snowmelt models for climate simulations. *J Climate* 12:2643–2657
40. Johnson M, Gaffigan S, Hunke E, Gerdes R (2007) A comparison of Arctic Ocean sea ice concentration among the coordinated AOMIP model experiments. *J Geophys Res* 112:C04S11
41. Jordan R (1991) A one-dimensional temperature model for a snow cover. Technical documentation for SNTHERM Special Technical Report 91-16. US Army Cold Regions Research and Engineering Laboratory, Hanover
42. Kudryavtsev VA et al (1974) Fundamentals of frost forecasting in geological engineering investigations. Nauka, Moscow (in Russian). English translation US Armt Cold Regions Res Engr Lan, Hannover, Draft translation 1977
43. Kwok R, Cunningham GF, Hibler III WD (2003) Sub-daily sea ice motion and deformation from RADARSAT observations. *Geophys Res Lett* 30(23):2218 doi:10.1029/2003GL018723
44. Lawrence DM, Slater AG (2007) A projection of severe near-surface permafrost degradation during the 21st century. *Geophys Res Lett* 32:L24401
45. Lindsay RW, Stern HL (2005) A new Lagrangian model of Arctic sea ice. *J Phys Oceanogr* 34:272–283
46. Ling F, Zhang T-J (2004) A numerical model for surface energy balance and thermal regime of the active layer and permafrost containing unfrozen water. *Cold Regions Sci Technol* 38:1–15
47. Liston GE, Hall DK (1995) An energy-balance model of lake-ice evolution. *J Glaciol* 41(138):373–382
48. Lunardini V (1988) Freezing of soil with an unfrozen water content and variable thermal properties. US Army Cold Regions Res Engineering Lab, Hanover, p 31
49. MacAyeal DR et al (1996) An ice-shelf model test based on the Ross ice shelf. *Antarct Ann Glaciol* 23:46–51
50. Martin Y, Gerdes R (2007) Sea ice drift variability in Arctic Ocean Model Intercomparison Project models and observations. *J Geophys Res* 112(C4):C04S10
51. Maykut G, Untersteiner N (1971) Some results from a time-dependent thermodynamic mode; of sea ice. *J Geophys Res* 76:1550–75
52. McClung D, Schaerer P (2006) *The Avalanche Handbook*. The Mountaineers, Seattle
53. Meehl GA, Boer GA, Covey C, Latif M, Stouffer RJ (1997) Intercomparison makes for a better climate model. *EOS* 78:445–446
54. Morgan VI, Jacka TH, Akerman GJ, Clarke AL (1982) Outlet glacier and mass budget studies in Enderby, Kemp and MacRobertson Lands, Antarctica. *Ann Glaciol* 3L:204–210
55. Nelson FE, Outcalt DSI (1987) A computational method for prediction and regionalization of permafrost. *Arct Alp Res* 19:279–88
56. Nelson FE et al (1997) Estimating Active-Layer Thickness over a Large Region: Kuparuk River Basin, Alaska, USA. *Arct Alp Res* 29:367–378
57. Nick EM, van der Veen CJ, Oerlemans J (2007) Controls on advance of tidewater glaciers: Results from numerical modeling applied to Columbia Glaciers. *J Geophys Res* 112:G03S24
58. Nicolsky DJ, Romanovsky VE, Alexeev VA, Lawrence DM (2007) Improved modeling of permafrost dynamics in a GCM Land Surface Scheme. *Geophys Res Lett* 34(8):L08591
59. Nye J (1951) The flow of glaciers and ice sheets as a problem in plasticity. *Proc Roy Soc Lond A* 207:554–572
60. Nye J (1965) The flow of a glacier in a channel of rectangular, elliptic or parabolic cross-section. *J Glaciol* 5:661–690

61. Oerlemans J (2005) Extracting a climate signal from 169 glacier records. *Science* 308:675–677
62. Oelke C et al (2003) Regional-scale modeling of soil freeze/thaw over the Arctic drainage basin. *J Geophys Res* 108(D10):4314
63. Oelke C, Zhang T-J (2004) A model study of circum-Arctic soil temperatures. *Permafrost Periglacial Process* 15:103–121
64. Orowan E (1949) Remarks at the joint meeting of the British Glaciological Society, the British Rheologists Club and the Institute of Metals. *J Glaciol* 1:231–236
65. Overland JE, McNutt SL, Salo S, Groves J, Li S (1998) Arctic sea ice as a granular plastic. *J Geophys Res* 104(C10):21845–21867
66. Parkinson CL, Washington WM (1979) A large-scale numerical model of sea ice. *J Geophys Res* 84:311–337
67. Paterson WSB (1994) *The physics of glaciers*. Pergamon, Elsevier Science, New York, p 480
68. Payne AJ et al (2000) Results from the EISMINT Phase 2 simplified geometry experiments: the effects of thermomechanical coupling. *J Glaciol* 46(153):227–238
69. Perla RI (1980) Avalanche release, motion, and impact. In: Colbeck SC (ed) *Dynamics of snow and ice masses*. Academic Press, New York, pp 397–462
70. Pomeroy JW, Parviainen J, Hedstrom N, Gray DM (1998) Coupled modelling of forest snow interception and sublimation. *Hydrol Process* 12:2317–2337
71. Pritchard RS, Coon M, McPhee MG, Leavitt E (1977) Winter ice dynamics in the nearshore Beaufort Sea. *AIDJEX Bull.* 37, Applied Physics Lab, University of Washington, Seattle, pp 37–93
72. Raymond CF (1980) Temperate valley glaciers. In: Colbeck SC (ed) *Dynamics of snow and ice masses*. New York. Academic Press, pp 79–139
73. Romanovsky VE, Osterkamp TE, Duzbury NS (1997) An evaluation of three numerical models used in simulations of the active layer and permafrost temperature regimes. *Cold Regions Sci Technol* 26:195–201
74. Saito K, Kimoto M, Zhang T, Takata K, Emori S (2007) Evaluating a high-resolution climate model: Simulated hydrothermal regimes in frozen ground regions and their change under the global warming scenario. *J Geophys Res* 112:F02S11
75. Sazonava TS, Romanovsky V (2003) A model for regional-scale estimation of temporal and spatial variability of active layer thickness and mean annual ground temperatures. *Permafrost Periglacial Proc* 14:125–139
76. Schoof C (2007) Ice sheet grounding line dynamics: Steady states, stability, and hysteresis. *J Geophys Res* 112:F03S28
77. Shiklomanov NI et al (2007) Comparison of model-produced active layer fields: Results for northern Alaska. *J Geophys Res* 112(F2):F02S10
78. Shiklomanov NI, Nelson FE (1999) Analytic representation of the active layer thickness field, Kuparuk River Basin, Alaska. *Ecol Model* 123:105–125
79. Shiklomanov NI, Nelson FE (2002) Active-layer mapping at regional scales: a 13-year spatial time series for the Kuparuk region, north-central Alaska. *Permafrost Periglacial Proc* 13:219–230
80. Steele M, Flato GM (2000) Sea ice growth and modeling: A survey. In: Lewis EL et al (eds) *The freshwater budget of the Arctic*. Kluwer, Dordrecht, pp 549–587
81. Stroeve J et al (2007) Arctic sea ice decline: Faster than forecast. *Geophys Res Lett* 34:L09501, doi:10.1029/2007GL029703
82. Thomas RH (1979) The dynamics of marine ice sheets. *J Glaciol* 24:167–177
83. Tremblay L-B, Mysak LA (1997) Modeling sea ice as a granular material, including the dilatancy effect. *J Phys Oceanogr* 27:2342–2360
84. Trujillo E, Ramirez JA, Elder KJ (2007) Topographic, meteorologic and canopy controls on the scaling characteristics of the spatial distribution of snow depth fields. *Water Resour Res* 43:W07409
85. van der Veen CJ, Payne AJ (2004) Modelling land-ice dynamics. In: Bamber JL, Payne AJ (eds) *Mass balance of the cryosphere: Observations and modelling of contemporary and future change*. Cambridge University Press, Cambridge, pp 169–225
86. Washington WM, Meehl GA (1996) High-latitude climate change in a global coupled ocean-atmosphere-sea ice model with increased atmospheric CO<sub>2</sub>. *J Geophys Res* 101(D8):12795–12802
87. Washington WM, Semtner AJ, Parkinson C, Morrison L (1976) On the development of a seasonal change sea-ice model. *J Oceanogr* 6:679–685
88. Weertman J (1957) On the sliding of glaciers. *J Glaciol* 5:287–303
89. Williams PJ, Smith MW (1989) *The frozen earth*. Cambridge University Press, Cambridge, p 306
90. Winstral A, Marks D (2002) Simulating wind fields and snow redistribution using terrain-based parameters to model snow accumulation and melt over a semi-arid mountain catchment. *Hydrol Process* 16:3585–3603
91. World Meteorological Organization (2007) *WMO sea ice nomenclature*, no 269. WMO, Geneva
92. Zhang T-J, Armstrong RL, Smith J (2003) Investigation of the near-surface soil freeze-thaw cycle in the contiguous United States: Algorithm development and validation. *J Geophys Res* 108(D22):8860, GCP 21-1 – 21-14
93. Zhang T-J et al (2005) Spatial and temporal variability in active layer thickness over the Russian Arctic drainage basin. *J Geophys Res* 110:D16101
94. Joint Commission on Oceanography and Marine Meteorology (2007) <http://www.ipy-ice-portal.org/>. Accessed 22 Aug 2008

### Books and Reviews

- Bamber JL, Payne AJ (eds) (2004) *Mass balance of the cryosphere: Observations and modelling of contemporary and future change*. Cambridge University Press, Cambridge, p 644

## Dynamic Games with an Application to Climate Change Models

PRAJIT K. DUTTA

Department of Economics, Columbia University,  
New York, USA

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[The Dynamic – or Stochastic – Game Model](#)

[The Dynamic – or Stochastic – Game: Results](#)

[Global Climate Change – Issues, Models](#)

[Global Climate Change – Results](#)

[Future Directions](#)

[Bibliography](#)

### Glossary

**Players** The agents who take actions. These actions can be – depending on application – the choice of capital stock, greenhouse emissions, level of savings, level of Research & Development expenditures, price level, quality and quantity of effort, etc.

**Strategies** Full contingent plans for the actions that players take. Each strategy incorporates a choice of action not just once but rather a choice of action for every possible decision node for the player concerned.

**Payoffs** The utility or returns to a player from playing a game. These payoffs typically depend on the strategies chosen – and the consequent actions taken – by the player herself as well as those chosen by the other players in the game.

**Game horizon** The length of time over which the game is played, i. e., over which the players take actions. The horizon may be finite – if there are only a finite number of opportunities for decision-making – or infinite – when there are an infinite number of decision-making opportunities.

**Equilibrium** A vector of strategies, one for each player in the game, such that no player can unilaterally improve her payoffs by altering her strategy, if the others' strategies are kept fixed.

**Climate change** The consequence to the earth's atmosphere of economic activities such as the production and consumption of energy that result in a build-up of greenhouse gases such as carbon dioxide.

### Definition of the Subject

The study of dynamic games is an important topic within game theory. Dynamic games involve the study of problems that are a) inherently dynamic in nature (even without a game-theoretic angle) and b) are naturally studied from a strategic perspective. Towards that end the structure generalizes dynamic programming – which is the most popular model within which inherently dynamic but non-strategic problems are studied. It also generalizes the model of repeated games within which strategic interaction is often studied but which structure cannot handle dynamic problems. A large number of economic problems fit these two requirements.

In this paper we examine the dynamic game model. The structure is discussed in detail as well as its principal results. Then the paper introduces a leading important application, the economics of climate change. It is shown that the problem is best studied as a dynamic commons game. Some recent models and associated results are then discussed.

We begin the analysis with a recall of the familiar model of repeated games (whose main results have been presented elsewhere in this volume). That is followed by the generalization of that framework to the model of dynamic – also known as stochastic or Markovian – games. These games may be thought of as “repeated games with a state variable”. The presence of a state variable allows the analysis of situations where there is a fundamental dynamic intrinsic to the problem, a situation – or “state” – that changes over time often on account of the players' past actions. (In contrast to repeated games where an identical stage game is played period after period.) Such a state variable maybe capital stock, level of technology, national or individual wealth or even environmental variables such as the size of natural resources or the stock of greenhouse gases. To provide a concrete illustration of the dynamic game concepts and results, this paper will provide a fairly detailed overview of ongoing research by a number of authors on the very current and important topic of the economics of global climate change.

Section “[The Dynamic – or Stochastic – Game Model](#)” recalls the repeated games structure, introduces the subject of dynamic games and presents the dynamic games model. Section “[The Dynamic – or Stochastic – Game: Results](#)” presents – mostly with proofs – the main results from the theory of dynamic games. Section “[Global Climate Change – Issues, Models](#)” then introduces the problem of climate change, argues why the dynamic game framework is appropriate for studying the problem and presents a family of models that have been recently stud-

ied by Dutta and Radner – and in a variant by Dockner, Long and Sorger. Finally, Sect. “Global Climate Change – Results” presents the main results of these analyzes of the climate change problem. Future directions for research are discussed in Sect. “Future Directions” while references are collected in Sect. “Bibliography”.

## Introduction

In this paper we examine the dynamic game model. The structure is discussed in detail as well as its principal results. Then the paper introduces a leading important application, the economics of climate change. It is shown that the problem is best studied as a dynamic commons game. Some recent models and associated results are then discussed.

## The Dynamic – or Stochastic – Game Model

The most familiar model of dynamic interaction is the Repeated Game model (described elsewhere in this volume). In that set-up players interact every period for many periods – finite or infinite in number. At each period they play exactly the same game, i. e., they pick from exactly the same set of actions and the payoff consequence of any given action vector is identical. Put differently, it is as if there is an intrinsic static set-up, a “state” of the system that never changes. The only thing that changes over time is (potentially) every player’s response to that fixed state, i. e., players (can) treat the game dynamically if they so wish but there is no inherent non-strategic reason to do so. An impartial “referee” choosing on behalf of the players to achieve some optimization aim indeed would pick the same action every period.

Things change in a set-up where the state can change over time. That is the structure to which we now turn. This set-up was introduced by Shapley [26] under the name of Stochastic Games. It has since also been called Markovian Games – on account of the Markovian structure of the intrinsic problem – or Dynamic Games. We will refer to the set-up (for the most part) as Dynamic Games.

## Set-Up

There are  $I$  players, and time is discrete. Each period the players interact by picking an action. Their action interaction take place at a given state which state changes as a consequence of the action interaction. There is a payoff that each player receives in each period based on the action vector that was picked and the state.

The basic variables are:

### Definition

- $t$  Time period  $(0, 1, 2, \dots, T)$ .
- $i$  Players  $(1, \dots, I)$ .
- $s(t)$  State at the beginning of period  $t$ ,  $s(t) \in S$ .
- $a_i(t)$  Action taken by player  $i$  in period,  $a_i(t) \in A_i$
- $a(t)$   $(a_1(t), a_2(t), \dots, a_I(t))$  vector of actions taken in period  $t$ .
- $\pi_i(t)$   $\pi_i(s(t), a(t))$  payoff of player  $i$  in period  $t$ .
- $q(t)$   $q(s(t+1) | s(t), a(t))$  conditional distribution of state at the beginning of period  $t+1$ .
- $\delta$  The discount factor,  $\delta \in [0, 1)$ .

The state variable affects play in two ways as stated above. In any given period, the payoff to a player depends not only on the actions that she and other players take but it also depends on the state in that period. Furthermore, the state casts a shadow on future payoffs in that it evolves in a Markovian fashion with the state in the next period being determined – possibly stochastically – by the state in the current period and the action vector played currently.

The initial value of the state,  $s(0)$ , is exogenous. So is the discount factor  $\delta$  and the game horizon,  $T$ . Note that the horizon can be finite or infinite. All the rest of the variables are endogenous, with each player controlling its own endogenous variable, the actions. Needless to add, both state as well as action variables can be multi-dimensional and when we turn to the climate change application it will be seen to be multi-dimensional in natural ways.

*Example 1*  $S$  infinite – The state space can be countably or uncountably infinite. It will be seen that the infinite case, especially the uncountably infinite one, has embedded within it a number of technical complications and – partly as a consequence – much less is known about this case.

$S$  finite – In this case, imagine that we have a repeated game like situation except that there are a finite number of stage games any one of which gets played at a time.

*Example 2* When the number of players is one, i. e.,  $I = 1$ , then we have a dynamic programming problem. When the number of states is one, i. e.,  $\#(S) = 1$ , then we have a repeated game problem. (Alternatively, repeated games constitute the special case where the conditional distribution brings a state  $s$  always back to itself, regardless of action.) Hence these two very familiar models are embedded within the framework of dynamic games.

## Histories and Strategies

**Preliminaries** – A *history* at time  $t$ ,  $h(t)$ , is a list of prior states and action vectors up to time  $t$  (but not including

$a(t)$ )

$$h(t) = s(0), a(0), s(1), a(1), \dots, s(t).$$

Let the set of histories be denoted  $H(t)$ . A *strategy* for player  $i$  at time  $t$ ,  $\sigma_i(t)$ , is a complete conditional plan that specifies a choice of action for every history. The choice may be probabilistic, i. e., may be an element of  $P(A_i)$ , the set of distributions over  $A_i$ . So a strategy at time  $t$  is

$$\sigma_i(t): H(t) \longrightarrow P(A_i).$$

A strategy for the entire game for player  $i$ ,  $\sigma_i$ , is a list of strategies, one for every period:  $\sigma_i = (\sigma_i(0), \sigma_i(1), \dots, \sigma_i(t), \dots)$ . Let  $\sigma = (\sigma_1, \sigma_2, \dots, \sigma_I)$  denote a vector of strategies, one for each player.

A particular example of a strategy for player  $i$  is a pure strategy  $\sigma_i$  where  $\sigma_i(t)$  is a deterministic choice (from  $A_i$ ). This choice may, of course, be conditional on history, i. e., may be a map from  $H(t)$  to  $A_i$ . Another example of a strategy for player  $i$  is one where the player's choice  $\sigma_i(t)$  may be probabilistic but the conditioning variables are not the entire history but rather only the current state. In other words such a strategy is described by a map from  $S$  to  $P(A_i)$  – and is called a Markovian strategy. Additionally, when the map is independent of time, the strategy is called a stationary Markovian strategy, i. e., a stationary Markovian strategy for player  $i$  is described by a mapping  $f_i: S \longrightarrow P(A_i)$ .

*Example 3* Consider, for starters, a pure strategy vector  $\sigma$ , i. e., a pure strategy choice for every  $i$ . Suppose further that  $q$  the conditional distribution on states is also deterministic. In that case, there is, in a natural way, a unique history that is generated by  $\sigma$ :

$$h(t; \sigma) = s(0), a(0; \sigma), s(1; \sigma), a(1; \sigma), \dots, s(t; \sigma)$$

where  $\mathbf{a}(\tau; \sigma) = \sigma(\tau; h(\tau; \sigma))$  and  $s(\tau + 1; \sigma) = q(s(\tau + 1) | s(\tau; \sigma), \mathbf{a}(\tau; \sigma))$ . This unique history associated with the strategy vector  $\sigma$  is also called the *outcome path* for that strategy. To every such outcome path there is an associated lifetime payoff

$$R_i(\sigma) = \sum_{t=0}^T \delta^t \pi_i(s(t; \sigma), \mathbf{a}(t; \sigma)). \quad (1)$$

If  $\sigma$  is a mixed strategy, or if the conditional distribution  $q$ , is not deterministic, then there will be a joint distribution on the set of histories  $H(t)$  generated by the strategy vector  $\sigma$  and the conditional distribution  $q$  in the obvious way. Moreover, there will be a marginal distribution on the state and action in period  $t$ , and under that

marginal, an expected payoff  $\pi_i(s(\tau; \sigma), \mathbf{a}(\tau; \sigma))$ . Thereafter lifetime payoffs can be written exactly as in Eq. 1.

Consider the game that remains after every history  $h(t)$ . This remainder is called a *subgame*. The restriction of the strategy vector  $\sigma$  to the subgame that starts after history  $h(t)$ , is denoted  $\sigma | h(t)$ .

### Equilibrium

A strategy vector  $\sigma^*$  is said to be a Nash Equilibrium (or NE) of the game if

$$R_i(\sigma^*) \geq R_i(\sigma_i, \sigma_{-i}^*), \quad \text{for all } i, \sigma_i. \quad (2)$$

A strategy vector  $\sigma^*$  is said to be a Subgame Perfect (Nash) Equilibrium of the game – referred to in short as SPE – if not only is Eq. 2 true for  $\sigma^*$  but it is true for every restriction of the strategy vector  $\sigma^*$  to every subgame  $h(t)$ , i. e., is true for  $\sigma^* | h(t)$  as well. In other words,  $\sigma^*$  is a SPE if

$$R_i(\sigma^* | h(t)) \geq R_i(\sigma_i, \sigma_{-i}^* | h(t)), \quad \text{for all } i, \sigma_i, h(t). \quad (3)$$

As is well-known, not all NE satisfy the further requirement of being a SPE. This is because a NE only considers the outcome path associated with that strategy vector  $\sigma^*$  – or, when the outcome path is probabilistic, only considers those outcome paths that have a positive probability of occurrence. That follows from the inequality Eq. 2. However, that does not preclude the possibility that players may have no incentive to follow through with  $\sigma^*$  if some zero probability history associated with that strategy is reached. (Such a history may be reached either by accident or because of deviation/experimentation by some player.) In turn that may have material relevance because how players behave when such a history is reached will have significance for whether or not a player wishes to deviate against  $\sigma^*$ . Eq. 3 ensures that – even after a deviation –  $\sigma^*$  will get played and that deviations are unprofitable.

Recall the definition of a stationary Markovian strategy (SMS) above. Associated with that class of strategies is the following definition of equilibrium. A stationary Markov strategy vector  $f^*$  is a Markov Perfect Equilibrium (MPE) if

$$R_i(f^*) \geq R_i(f_i, f_{-i}^*), \quad \text{for all } i, f_i.$$

Hence, a MPE restricts attention to SMS both on and off the outcome path. Furthermore, it only considers – implicitly – histories that have a positive probability of occurrence under  $f^*$ . Neither “restriction” is a restriction when  $T$  is infinite because when all other players play

a SMS player  $i$  has a stationary dynamic programming problem to solve in finding his most profitable strategy and – as is well-known – he loses no payoff possibilities in restricting himself to SMS as well. And that best strategy is a best strategy on histories that have zero probabilities of occurrence as well as histories that have a positive probability of occurrence. In particular therefore, when  $T$  is infinite, a MPE is also a SPE.

**The Dynamic – or Stochastic – Game: Results**

The main questions that we will now turn to are:

1. Is there always a SPE in a dynamic – or stochastic – game?
2. Is there a characterization for the set of SPE akin to the Bellman optimality equation of dynamic programming? If yes, what properties can be deduced of the SPE payoff set?
3. Is there a Folk Theorem for dynamic games – akin to that in Repeated Games?
4. What are the properties of SPE outcome paths?

The answers to questions 1–3 are very complete for finite dynamic games, i. e., games where the state space  $S$  is finite. The answer is also complete for questions 1 and 2 when  $S$  is countably infinite but when the state space is uncountably infinite, the question is substantively technically difficult and there is reason to believe that there may not always be a SPE. The finite game arguments for question 3 is conceptually applicable when  $S$  is (countably or uncountably) infinite provided some technical difficulties can be overcome. That and extending the first two answers to uncountably infinite  $S$  remain open questions at this point. Not a lot is known about Question 4.

**Existence**

The first result is due to Parthasarathy [23] and applies to the infinite horizon model, i. e., where  $T = \infty$ . When  $T$  is finite, the result and arguments can be modified in a straightforward way as will be indicated in the remarks following the proof.

**Theorem 1** *Suppose that  $S$  is countable,  $A_i$  are finite sets, and the payoff functions  $\pi_i$  are bounded. Suppose furthermore that  $T$  is infinite. Then there is a MPE (and hence a SPE).*

*Proof* The proof will be presented by way of a fixed point argument. The domain for the fixed point will be the set of

stationary Markovian strategies:

$$M_i = \{f_i : S \rightarrow P(A_i), \text{ s.t. for all } s, \sum_{a_i} f_i(a_i; s) = 1, f_i(a_i; s) \geq 0\}.$$

□

**Properties of  $M_i$ :** In the pointwise convergence topology,  $M_i$  is compact. That this is so follows from a standard diagonalization argument by way of which a subsequence can be constructed from any sequence of SMS  $f_i^n$  such that the subsequence, call it  $f_i^{n'}$  has the property that  $f_i^{n'}(s)$  converges to some  $f_i^0(s)$  for every  $s$ . Clearly,  $f_i^0 \in M_i$ . The diagonalization argument requires  $S$  to be countable and  $A_i$  to be finite.

$M_i$  is also clearly convex since its elements are probability distributions on  $A_i$  at every state.

The mapping for which we shall seek a fixed point is the best response mapping:

$$B_i(f) = \{g_i \in M_i : R_i(g_i, f_{-i}) \geq R_i(f_i, f_{-i}), \text{ for all } f_i.$$

Since the best response problem for player  $i$  is a stationary dynamic programming problem, it follows that there is an associated value function for the problem, say  $v_i$ , such that it solves the optimality equation of dynamic programming

$$v_i(s) = \max_{\lambda_i} \left\{ \pi_i(s, \lambda_i, f_{-i}(s)) + \delta \sum_{s'} v_i(s') q(s' | s, \lambda_i, f_{-i}(s)) \right\} \quad (4)$$

where

$$\pi_i(s, \lambda_i, f_{-i}(s)) = \sum_{a_{-i}} \left[ \sum_{a_i} \pi_i(s, a_i, a_{-i}) \lambda_i(a_i) \right] f_i(a_{-i}, s) \quad (5)$$

where  $\lambda_i(a_i)$  is the probability of player  $i$  picking action  $a_i$  whilst  $f_i(a_{-i}, s)$  is the product probability of players other than  $i$  picking the action vector  $a_{-i}$ . Similarly.

$$q(s' | s, \lambda_i, f_{-i}(s)) = \sum_{a_{-i}} \left[ \sum_{a_i} q(s' | s, a_i, a_{-i}) \lambda_i(a_i) \right] f_i(a_{-i}, s). \quad (6)$$

Additionally, it follows that the best best response, i. e.,  $g_i$ , solves the optimality equation, i. e.,

$$v_i(s) = \pi_i(s, g_i, f_{-i}(s)) + \delta \sum_{s'} v_i(s') q(s' | s, g_i, f_{-i}(s)) \quad (7)$$

where  $\pi_i(s, g_i, f_{-i}(s))$  and  $q(s' | s, g_i, f_{-i}(s))$  have the same interpretations as given by Eqs. 5 and 6.

**Properties of  $B_i$ :**  $B_i$  is a correspondence that is convex-valued and upper hemi-continuous. That  $B_i$  is convex-valued follows from the fact that we are operating in the set of mixed strategies, that every convex combination of mixed strategies is itself a mixed strategy and that every convex combination of best responses is also a best response.

To show that  $B_i$  is upper hemi-continuous, consider a sequence of other players' strategies  $f_{-i}^n$ , an associated best response sequence of player  $i$ ,  $g_i^n$  with value function sequence  $v_i^n$ . Note that each of these best responses  $g_i^n$  satisfies the Eqs. 6 and 7 (for the value function  $v_i^n$ ). By diagonalization there exist subsequences and subsequential pointwise convergent limits:  $f_{-i}^n \rightarrow f_{-i}^0$ ,  $g_i^n \rightarrow g_i^0$ , and  $v_i^n \rightarrow v_i^0$ . It suffices to show that

$$v_i^0(s) = \max_{\lambda_i} \left\{ \pi_i(s, \lambda_i, f_{-i}^0(s)) + \delta \sum_{s'} v_i^0(s') q(s' | s, \lambda_i, f_{-i}^0(s)) \right\} \quad (8)$$

and

$$v_i^0(s) = \pi_i(s, g_i^0, f_{-i}^0(s)) + \delta \sum_{s'} v_i^0(s') q(s' | s, g_i^0, f_{-i}^0(s)). \quad (9)$$

Equation. 9 will be proved by using the analog of Eq. 7, i. e.,

$$v_i^n(s) = \pi_i(s, g_i^n, f_{-i}^n(s)) + \delta \sum_{s'} v_i^n(s') q(s' | s, g_i^n, f_{-i}^n(s)). \quad (10)$$

Clearly the left-hand side of Eq. 10 converges to the left-hand side of Eq. 9. Lets check the right-hand side of each equation. Evidently

$$\begin{aligned} & \sum_{a_{-i}} \left[ \sum_{a_i} \pi_i(s, a_i, a_{-i}) g_i^n(a_i) \right] f_i^n(a_{-i}; s) \\ & \rightarrow \sum_{a_{-i}} \left[ \sum_{a_i} \pi_i(s, a_i, a_{-i}) g_i^0(a_i) \right] f_i^0(a_{-i}; s) \end{aligned}$$

since each component of the sum converges and we have a finite sum. Finally,

$$\begin{aligned} & \left| \sum_{s'} v_i^n(s') q(s' | s, g_i^n, f_{-i}^n(s)) \right. \\ & \quad \left. - \sum_{s'} v_i^0(s') q(s' | s, g_i^0, f_{-i}^0(s)) \right| \\ & \leq \left| \sum_{s'} [v_i^n(s') - v_i^0(s')] q(s' | s, g_i^n, f_{-i}^n(s)) \right| \quad (11) \\ & \quad + \left| \sum_{s'} v_i^0(s') [q(s' | s, g_i^n, f_{-i}^n(s)) \right. \\ & \quad \left. - q(s' | s, g_i^0, f_{-i}^0(s))] \right|. \end{aligned}$$

The first term in the right-hand side of the inequality above goes to zero by the dominated convergence theorem. The second term can be re-written as

$$\begin{aligned} & \sum_{s'} v_i^0(s') q(s' | s, a_i, a_{-i}) \\ & \quad \cdot \sum_{a_{-i}} \sum_{a_i} [g_i^n(a_i) f_i^n(a_{-i}; s) - g_i^0(a_i) f_i^0(a_{-i}; s)] \end{aligned}$$

and goes to zero because each of the finite number of terms in the summation over action probabilities goes to zero. Hence the RHS of Eq. 10 converges to the RHS of Eq. 9 and the proposition is proved.

*Remark 1* Note that the finiteness of  $A_i$  is crucial. Else, the very last argument would not go through, i. e., knowing that  $[g_i^n(a_i) f_i^n(a_{-i}; s) - g_i^0(a_i) f_i^0(a_{-i}; s)] \rightarrow 0$  for every action vector  $a$  would not guarantee that the sum would converge to zero as well.

*Remark 2* If the horizon were finite one could use the same argument to prove that there exists a Markovian strategy equilibrium, though not a stationary Markovian equilibrium. That proof would combine the arguments above with backward induction. In other words, one would first use the arguments above to show that there is an equilibrium at every state in the last period  $T$ . Then the value function so generated,  $v_i^T$ , would be used to show that there is an equilibrium in period  $T-1$  using the methods above thereby generating the relevant value function for the last two periods,  $v_i^{T-1}$ . And so on.

The natural question to ask at this point is whether the restriction of countable finiteness of  $S$  can be dropped (and – eventually – the finiteness restriction on  $A_i$ ). The answer, unfortunately, is not easily. The problems are two-fold:



1. *Sequential Compactness of the Domain Problem* – If  $S$  is uncountably infinite, then it is difficult to find a domain  $M_i$  that is sequentially compact. In particular, diagonalization arguments do not work to extract candidate strategy and value function limits.
2. *Integration to the Limit Problem* – Note as the other players change their strategies,  $f_{-i}^n$ , continuation payoffs to player  $i$  change in two ways. They change first because the value function  $v_i^n$  changes, i. e.,  $v_i^n \neq v_i^m$  if  $n \neq m$ . Second, the expected continuation value changes because the measure over which the value function is being integrated,  $q(s' | s, \lambda_i, f_{-i}^n(s))$ , itself changes, i. e.,  $q(s' | s, \lambda_i, f_{-i}^n(s)) \neq q(s' | s, \lambda_i, f_{-i}^m(s))$ . This is the well-known – and difficult – integration to the limit problem: simply knowing that  $v_i^n$  “converges” to  $v_i^0$  in some sense – such as pointwise – and knowing that the integrating measure  $q^n$  “converges” to  $q^0$  in some sense – such as in the weak topology – does not, in general, imply that

$$\int v_i^n dq^n \rightarrow \int v_i^0 dq^0. \tag{12}$$

(Of course in the previous sentence  $q^n$  is a more compact stand-in for  $q(s' | s, \lambda_i, f_{-i}^n(s))$  and  $q^0$  for  $q(s' | s, \lambda_i, f_{-i}^0(s))$ .) There are a limited number of cases where Eq. 12 is known to be true. These results typically require  $q^n$  to converge to  $q^0$  in some strong sense. In the dynamic game context what this means is that very strong convergence restrictions need to be placed on the transition probability  $q$ . This is the underlying logic behind results reported in [7,20,22,24].

Such strong convergence properties are typically not satisfied when  $q$  is deterministic – which case comprises the bulk of the applications of the theory. Indeed simply imposing continuity when  $q$  is deterministic appears not to be enough to generate an existence result. Harris et al. [16] and Dutta and Sundaram [14] contain results that show that there may not be a SPE in finite horizon dynamic games when the transition function  $q$  is continuous. Whether other often used properties of  $q$  and  $\pi_i$  – such as concavity and monotonicity – can be used to rescue the issue remains an open question.

**Characterization**

The Bellman optimality equation has become a workhorse for dynamic programming analysis. It is used to derive properties of the value function and the optimal strategies. Moreover it provides an attractive and conceptually simple way to view a multiple horizon problem as a series of

one-stage programming problems by exploiting the recursive structure of the optimization set-up. A natural question to ask, since dynamic games are really multi-player versions of dynamic programming, is whether there is an analog of the Bellman equation for these games. Abreu, Pearce and Stachetti – APS (1988), in an important and influential paper, showed that this is indeed the case for repeated games. They defined an operator, hereafter the APS operator, whose largest fixed point is the set of SPE payoffs in a repeated game and whose every fixed point is a subset of the set of SPE payoffs. (Thereby providing a necessary and sufficient condition for SPE equilibrium payoffs in much the same way that the unique fixed point of the Bellman operator constitutes the value function for a dynamic programming problem.) As with the Bellman equation, the key idea is to reduce the multiple horizon problem to a (seemingly) static problem.

In going from repeated to dynamic games there are some technical issues that arise. We turn now to that analysis pointing out along the way where the technical pitfalls are. Again we start with the infinite horizon model, i. e., where  $T = \infty$ . When  $T$  is finite, the result and arguments can be modified in a straightforward way as will be indicated in a remark following the proof.

But first, some definitions. Suppose for now that  $S$  is countable.

*APS Operator* – Consider a compact-valued correspondence,  $W$  defined on domain  $S$  which takes values that are subsets of  $\mathbb{R}^I$ . Define the APS operator on  $W$ , call it  $LW$ , as follows:

$$LW(s) = \left\{ \begin{array}{l} v \in \mathbb{R}^I : \exists \widehat{f} \in P(A) \text{ and} \\ w : S \times A \times S \rightarrow W, \\ \text{uniformly bounded, s.t. } v_i = \pi_i(s, \widehat{f}) \\ + \delta \sum_{s'} w_i(s, \widehat{f}, s') q(s' | s, \widehat{f}) \\ \geq \pi_i(s, a_i, \widehat{f}_{-i}) + \delta \sum_{s'} w_i(s, a_i, \widehat{f}_{-i}, s') \\ q(s' | s, a_i, \widehat{f}_{-i}), \text{ for all } a_i, i \end{array} \right\} \tag{13}$$

where, as before,

$$\pi_i(s, \widehat{f}) = \sum_{a_{-i}} \left[ \sum_{a_i} \pi_i(s, a_i, a_{-i}) \widehat{f}_i(a_i) \right] \widehat{f}_{-i}(a_{-i}; s) \tag{14}$$

and

$$q(s' | s, \widehat{f}) = \sum_{a_{-i}} \left[ \sum_{a_i} q(s' | s, a_i, a_{-i}) \widehat{f}_i(a_i) \right] \widehat{f}_{-i}(a_{-i}; s) \tag{15}$$

Finally, let  $V^*$  denote the SPE payoffs correspondence, i. e.,  $V^*(s)$  is the set of SPE payoffs starting from initial state  $s$ . Note that the correspondence is non-empty by virtue of Theorem 1.

**Theorem 2** *Suppose that  $S$  is countable,  $A_i$  are finite sets, and the payoff functions  $\pi_i$  are bounded. Suppose furthermore that  $T$  is infinite. Then a)  $V^*$  is a fixed point of the APS operator, i. e.,  $LV^* = V^*$ . Furthermore, b) consider any other fixed point, i. e., a correspondence  $\tilde{V}$  such that  $L\tilde{V} = \tilde{V}$ . Then it must be the case that  $\tilde{V} \subset V^*$ . Finally, c) there is an algorithm that generates the SPE correspondence,  $V^*$ .*

*Proof of a:* Suppose that  $v^* \in V^*(s)$ , i. e., is a SPE payoff starting from  $s$ . Then, by definition, there is a first-period play,  $f^*$  and a continuation strategy after every one-period history  $h(1)$ ,  $\sigma^*(h(1))$ , such that  $\{f^*, \sigma^*(h(1))\}$  is a SPE. By definition, then, the payoffs associated with each history-dependent strategy,  $\sigma^*(h(1))$ , call them  $w_i(s, f^*, s')$ , satisfy Eq. 13. (Note that  $\pi_i$  bounded implies that the lifetime payoffs  $w_i(s, f^*, s')$  are uniformly bounded. In other words,  $v^* \in LV^*(s)$ . On the other hand, suppose that  $v^* \in LV^*(s)$ . Then, by definition, there is a first-period play,  $f^*$ , and SPE payoffs,  $w(s, f^*, s')$ , that together satisfy Eq. 13. Let the SPE strategy associated with  $w(s, f^*, s')$  be  $\sigma^*(h(1))$ . It is not difficult to see that the concatenated strategy –  $f^*, \sigma^*(h(1))$  – forms a SPE. Since the associated lifetime payoff is  $v^*$ , it follows that  $v^* \in V^*(s)$ .  $\square$

*Proof of b:* Suppose that  $L\tilde{V} = \tilde{V}$ . Let  $v \in L\tilde{V}$ . Then, from Eq. 13, there is  $\tilde{f}$ , and SPE payoffs,  $w(s, \tilde{f}, s')$  that satisfy the equation. In turn, since  $w(s, \tilde{f}, s') \in \tilde{V}(s') = L\tilde{V}(s')$  there is an associated  $\tilde{f}(s')$  and  $w(s', \tilde{f}, s'')$  for which Eq. 13 holds. By repeated application of this idea, we can create a sequence of strategies for periods  $t = 0, 1, 2, \dots$  –  $\tilde{f}, \tilde{f}(s'), \tilde{f}(s, s'), \dots$  such that at each period Eq. 13 holds. Call the strategy so formed,  $\phi$ . This strategy can then not be improved upon by a single-period deviation. A standard argument shows that if a strategy cannot be profitably deviated against in one period then it cannot be profitably deviated against even by deviations in multiple periods. (This idea of “unimprovability” is already present in dynamic programming. Within the context of repeated games, it was articulated by Abreu [1].)  $\square$

*Proof of c:* Note two properties of the APS operator:

**Lemma 1**  *$LW$  is a compact-valued correspondence (whenever  $W$  is compact-valued).*

*Proof* Consider Eq. 13. Suppose that  $v^n \in LW(s)$  for all  $n$ , with associated  $\hat{f}^n$  and  $w^n$ . By diagonalization, there exists a subsequence s.t.  $v^n \rightarrow v^0, \hat{f}^n \rightarrow \hat{f}^0$  and  $w^n \rightarrow w^0$ .

This arguments uses the countability of  $S$  and the finiteness of  $A_i$ . From Eq. 14 evidently  $\pi_i(s, \hat{f}^n) \rightarrow \pi_i(s, \hat{f}^0)$  and similarly from Eq. 15  $\sum_{s'} w_i(s, \hat{f}^n, s')q(s' | s, \hat{f}^n)$  goes to  $\sum_{s'} w_i(s, \hat{f}^0, s')q(s' | s, \hat{f}^0)$ . Hence the inequality in Eq. 13 is preserved and  $v^0 \in LW(s)$ .  $\square$

It is not difficult to see that – on account of the boundedness of  $\pi_i$  – if  $W$  has a uniformly bounded selection, then so does  $LW$ . Note that the operator is also monotone in the set-inclusion sense, i. e., if  $W'(s) \subset W(s)$  for all  $s$  then  $LW' \subset LW$ .

The APS algorithm finds the set of SPE payoffs by starting from a particular starting point, an initial set  $W^0(s)$  that is taken to be the set of all feasible payoffs from initial state  $s$ . (And hence the correspondence  $W^0$  is so defined for every initial state.) Then define,  $W^1 = LW^0$ . More generally,  $W^{n+1} = LW^n, n \geq 0$ . It follows that  $W^1 \subset W^0$ . This is because  $W^1$  requires a payoff that is not only feasible but additionally satisfies the incentive inequality of Eq. 13 as well. From the monotone inclusion property above it then follows that, more generally,  $W^{n+1} \subset W^n, n \geq 0$ . Furthermore,  $W^n(s)$  is a non-empty, compact set for all  $n$  (and  $s$ ). Hence,  $W^\infty(s) = \bigcap_n W^n(s) = \lim_{n \rightarrow \infty} W^n(s)$  is non-empty and compact.

Let us now show that  $W^\infty$  is a fixed point of the APS operator, i. e., that  $LW^\infty = W^\infty$ .

**Lemma 2**  *$LW^\infty = W^\infty$ , or, equivalently,  $L(\lim_{n \rightarrow \infty} W^n) = \lim_{n \rightarrow \infty} LW^n$ .*

*Proof* Clearly, by monotonicity,  $L(\lim_{n \rightarrow \infty} W^n) \subset \lim_{n \rightarrow \infty} LW^n$ . So consider a  $v \in LW^n(s)$ , for all  $n$ . By Eq. 13 there is at each  $n$  an associated first-period play  $f^n$  and a continuation payoff  $w^n(s, f^n, s')$  such that the inequality is satisfied and

$$v_i = \pi_i(s, f^n) + \delta \sum_{s'} w_i^n(s, f^n, s')q(s' | s, f^n).$$

By the diagonalization argument, and using the countability of  $S$ , we can extract a (subsequential) limit  $f^\infty = \lim_{n \rightarrow \infty} f^n$  and  $w^\infty = \lim_{n \rightarrow \infty} w^n$ . Clearly,  $w^\infty \in W^\infty$ . Since equalities and inequalities are maintained in the limit, equally clearly

$$v_i = \pi_i(s, f^\infty) + \delta \sum_{s'} w_i^\infty(s, f^\infty, s')q(s' | s, f^\infty)$$

and

$$v_i \geq \pi_i(s, a_i, f_{-i}^\infty) + \delta \sum_{s'} w_i^\infty(s, a_i, f_{-i}^\infty, s')q(s' | s, a_i, f_{-i}^\infty), \text{ for all } a_i, i$$

thereby proving that  $v \in L(\lim_{n \rightarrow \infty} W^n)(s)$ . The lemma is proved.  $\square$

Since the set of SPE payoffs,  $V^*(s)$ , is a subset of  $W^0(s)$  – and  $LV^*(s) = V^*(s)$  – it further follows  $V^*(s) \subset W^\infty(s)$ , for all  $s$ . From the previous lemma, and part b), it follows that  $V^*(s) \supset W^\infty(s)$ , for all  $s$ . Hence,  $V^* = W^\infty$ . Theorem 2 is proved.  $\square$

A few remarks are in order.

*Remark 1* If the game horizon  $T$  is finite, there is an immediate modification of the above arguments. In the algorithm above, take  $W^0$  to be the set of SPE payoffs in the one-period game (with payoffs  $\pi_i$  for player  $i$ ). Use the APS operator thereafter to define  $W^{n+1} = LW^n$ ,  $n \geq 0$ . It is not too difficult to show that  $W^n$  is the set of SPE payoffs for a game that lasts  $n + 1$  periods (or has  $n$  remaining periods after the first one).

*Remark 2* Of course an immediate corollary of the above theorem is that the set of SPE payoffs  $V^*(s)$  is a compact set for every initial state  $s$ . Indeed one can go further and show that  $V^*$  is in fact an upper hemi-continuous correspondence. The arguments are very similar to those used above – plus the Maximum Theorem.

*Remark 3* Another way to think of Theorem 2 is that it is also an existence theorem. Under the conditions outlined in the result, the SPE equilibrium set has been shown to be non-empty. Of course this is not a generalization of Theorem 1 since Theorem 2 does not assert the existence of a MPE.

*Remark 4* When the state space  $A_i$  is infinite or the state space  $S$  is uncountably infinite we run into technical difficulties. The complications arise from not being able to take limits. Also, as in the discussion of the Integration to the Limit problem, integrals can fail to be continuous thereby rendering void some of the arguments used above.

**Folk Theorem**

The folk theorem for Repeated Games – Fudenberg and Maskin [15] following up on earlier contributions – is very well-known and the most cited result of that theory. It proves that the necessary conditions for a payoff to be a SPE payoff – feasibility and individual rationality – are also (almost) sufficient provided the discount factor  $\delta$  is close enough to 1. This is the result that has become the defining result of Repeated Games. For supporters, the result and its logic of proof are a compelling demonstration of the power of reciprocity, the power of long-term relationships in fostering cooperation through the lurking power of “punishments” when cooperation breaks

down. It is considered equally important and significant that such long-term relationships and behaviors are sustained through implicit promises and threats which therefore do not violate any legal prohibitions against explicit contracts that specify such behavior. For detractors, the “anything goes” implication of the Folk Theorem is a clear sign of its weakness – or the weakness of the SPE concept – in that it robs the theory of all predictive content. Moreover there is a criticism, not entirely correct, that the strategies required to sustain certain behaviors are so complex that no player in a “real-world” setting could be expected to implement them.

Be that as it may, the Folk Theorem question in the context of Dynamic Games then is: is it the case that feasibility and individual rationality are also (almost) enough to guarantee that a payoff is a SPE payoff at high enough  $\delta$ ? Two sets of obstacles arise in settling this question. Both emanate from the same source, the fact that the state does not remain fixed in the play of the games, as it does in the case of Repeated Games. First, one has to think long and hard as to how one should define individual rationality. Relatedly, how does one track feasibility? In both cases, the problem is that what payoff is feasible and individually rational depends on the state and hence changes after every history  $h(t)$ . Moreover, it also changes with the discount factor  $\delta$ . The second set of problems stems from the fact that a deviation play can unalterably change the future in a dynamic game – unlike a repeated game where the basic game environment is identical every period. Consequently one cannot immediately invoke the logic of repeated game folk theorems which basically work because any deviation has only short-term consequences while the punishment of the deviation is long-term. (And so if players are patient they will not deviate.)

Despite all this, there are some positive results that are around. Of these, the most comprehensive is one due to Dutta [9]. To set the stage for that result, we need a few crucial preliminary results. For this sub-section we will assume that  $S$  is finite – in addition to  $A_i$ .

**Feasible Payoffs Role of Markovian Strategies**

– Let  $F(s, \delta)$  denote the set of “average” feasible payoffs from initial state  $s$  and for discount factor  $\delta$ . By that I mean

$$F(s, \delta) = \{v \in \mathbb{R}^I : \exists \text{ strategy } \sigma$$

$$\text{s.t. } v = (1 - \delta) \sum_{t=0}^T \delta^t \pi_i(s(t; \sigma), a(t; \sigma))\}.$$

Let  $\Phi(s, \delta)$  denote the set of “average” feasible payoffs from initial state  $s$  and for discount factor  $\delta$  that are generated by pure stationary Markovian strategies – PSMS. Re-

call that a SMS is given by a map  $f_i$  from  $S$  to the probability distributions over  $A_i$ , so that at state  $s(t)$  player  $i$  chooses the mixed strategy  $f_i(s(t))$ . A pure SMS is one where the map  $f_i$  is from  $S$  to  $A_i$ . In other words,

$$\Phi(s, \delta) = \{v \in \mathbb{R}^I : \exists \text{ PSMS } f \\ \text{s.t. } v = (1 - \delta) \sum_{t=0}^T \delta^t \pi_i(s(t); f), a(t; f)\}.$$

**Lemma 3** *Any feasible payoff in a dynamic game can be generated by averaging over payoffs to stationary Markov strategies, i. e.,  $F(s, \delta) = \text{co}\Phi(s, \delta)$ , for all  $(s, \delta)$ .*

*Proof* Note that  $F(s, \delta) = \text{co}[\text{extreme points } F(s, \delta)]$ . In turn, all extreme points of  $F(s, \delta)$  are generated by an optimization problem of the form:  $\max_{\sigma} \sum_{i=1}^I \alpha_i v_i(s, \delta)$ . That optimization problem is a dynamic programming problem. Standard results in dynamic programming show that the optimum is achieved by some stationary Markovian strategy.  $\square$

Let  $F(s)$  denote the set of feasible payoffs under the long-run average criterion. The next result will show that this is the set to which discounted average payoffs converge:

**Lemma 4**  $F(s, \delta) \rightarrow F(s)$ , as  $\delta \rightarrow 1$ , for all  $s$ .

*Proof* Follows from the fact that a)  $F(s) = \text{co}\Phi(s)$  where  $\Phi(s)$  is the set of feasible long-run average payoffs generated by stationary Markovian strategies, and b)  $\Phi(s, \delta) \rightarrow \Phi(s)$ . Part b) exploits the finiteness of  $S$  (and  $A_i$ ).  $\square$

The lemmas above simplify the answer to the question: What is a feasible payoff in a dynamic game? Note that they also afford a dimensional reduction in the complexity and number of strategies that one needs to keep track of to answer the question. Whilst there are an uncountably infinite number of strategies – even with finite  $S$  and  $A_i$  – including the many that condition on histories in arbitrarily complex ways – the lemmas establish that all we need to track are the finite number of PSMS. Furthermore, whilst payoffs do depend on  $\delta$ , if the discount factor is high enough then the set of feasible payoffs is well-approximated by the set of feasible long-run average payoffs to PSMS.

One further preliminary step is required however. This has to do with the fact that while  $v \in F(s, \delta)$  can be exactly reproduced by a period 0 average over PSMS payoffs, after that period continuation payoffs to the various component strategies may generate payoffs that could be arbitrarily distant from  $v$ . This, in turn, can be problematic since one would need to check for deviations at every one of these (very different) payoffs. The next lemma addresses this problem by showing that there is an averaging

over the component PSMS that is ongoing, i. e., happens periodically and not just at period 0, but which, consequently, generates payoffs that after all histories stays arbitrarily close to  $v$ .

For any two PSMS  $f^1$  and  $f^2$  denote a *time-cycle strategy* as follows: for  $T^1$  periods play proceeds along  $f^1$ , then it moves for  $T^2$  periods to  $f^2$ . After the elapse of the  $T^1 + T^2$  periods play comes back to  $f^1$  for  $T^1$  periods and  $f^2$  for  $T^2$  periods. And so on. Define  $\lambda^1 = T^1/(T^1 + T^2)$ . In the obvious way, denote a general time-cycle strategy to be one that cycles over any finite number of PSMS  $f^k$  where the proportion of time spent at strategy  $f^k$  is  $\lambda^k$  and allows the lengths of time to depend on the initial state at the beginning of the cycle.

**Lemma 5** *Pick any  $v \in \cap_s F(s)$ . Then for all  $\varepsilon > 0$  there is a time cycle strategy such that its long-run average payoff is within  $\varepsilon$  of  $v$  after all histories.*

*Proof* Suppose that  $v = \sum_k \lambda^k(s) v^k(s)$  where  $v^k(s)$  is the long-run average payoff to the  $k$ th PSMS when the initial state is  $s$ . Ensure that  $T^k$  is chosen such that a) the average payoff over those periods under that PSMS –  $1/T^k \sum_{t=0}^{T^k-1} \pi_i(s(t); f^k), a(t; f^k)$  – is within  $\varepsilon$  of  $v^k(s)$  for all  $s$ . And b) that  $T^k(s)/\sum_i T^i(s)$  is arbitrarily close to  $\lambda^k(s)$  for all  $s$ .  $\square$

Since  $\Phi(s, \delta) \rightarrow \Phi(s)$  it further follows that the above result also holds under discounting:

**Lemma 6** *Pick any  $v \in \cap_s F(s)$ . Then for all  $\varepsilon > 0$  there is a time cycle strategy and a discount cut-off  $\delta(\varepsilon) < 1$  such that the discounted average payoffs to that strategy are within  $\varepsilon$  of  $v$  for all  $\delta > \delta(\varepsilon)$  and after all histories.*

*Proof* Follows from the fact that  $(1-\delta)/(1-\delta^T) \sum_{t=0}^{T^k-1} \delta^t \pi_i(s(t); f^k), a(t; f^k)$  goes to  $\frac{1}{T^k} \sum_{t=0}^{T^k-1} \pi_i(s(t); f^k), a(t; f^k)$  as  $\delta \rightarrow 1$ .  $\square$

**Individually Rational Payoffs** Recall that a min-max payoff is a payoff level that a player can guarantee by playing a best response. In a Repeated Game that is defined at the level of the component stage game. Since there is no analog of that in a dynamic game, the min-max needs to be defined over the entire game – and hence is sensitive to initial state and discount factor:

$$m_i(s, \delta) = \min_{\sigma_{-i}} \max_{\sigma_i} R_i(\sigma | s, \delta).$$

Evidently, given  $(s, \delta)$ , in a SPE it cannot be that player  $i$  gets a payoff  $v_i(s, \delta)$  that is less than  $m_i(s, \delta)$ . Indeed that inequality must hold at all states for a strategy to be a SPE, i. e., for all  $s(t)$  it must be the case that

$v_i(s(t), \delta) \geq m_i(s(t), \delta)$ . But, whilst necessary, even that might not be a sufficient condition for the strategy to be a SPE. The reason is that if player  $i$  can deviate and take the game to, say,  $s'$  at  $t + 1$ , rather than  $s(t + 1)$ , he would do so if  $v_i(s(t), \delta) < m_i(s', \delta)$  since continuation payoffs from  $s'$  have to be at least as large as the latter level and this deviation would be worth essentially that continuation when  $\delta$  is close to 1. So sufficiency will require a condition such as  $v_i(s(t), \delta) > \max_s m_i(s, \delta)$  for all  $s(t)$ . Call such a strategy *dynamically Individually Rational*. From the previous lemmas, and the fact that  $m_i(s, \delta) \rightarrow m_i(s)$ , as  $\delta \rightarrow 1$ , where  $m_i(s)$  is the long-run average min-max level for player  $i$  the following result is obvious. The min-max limiting result is due to Mertens and Neyman [21].

**Lemma 7** *Pick any  $v \in \cap_s F(s)$  such that  $v_i > \max_s m_i(s)$  for all  $s$ . Then there is a time-cycle strategy which is dynamically Individually Rational for high  $\delta$ .*

We are now ready to state and prove the main result:

**Theorem 3 (Folk Theorem)** *Suppose that  $S$  and  $A_i$  are finite sets. Suppose furthermore that  $T$  is infinite and that  $\cap_s F(s)$  has dimension  $I$  (where  $I$  is the number of players). Pick any  $v \in \cap_s F(s)$  such that  $v_i > \max_s m_i(s)$  for all  $s$ . Then, for all  $\epsilon > 0$ , there is a discount cut-off  $\delta(\epsilon) < 1$  and a time-cycle strategy that for  $\delta > \delta(\epsilon)$  is a SPE with payoffs that are within  $\epsilon$  of  $v$ .*

*Proof* Without loss of generality, let us set  $\max_s m_i(s) = 0$  for all  $i$ . From the fact that  $\cap_s F(s)$  has dimension  $I$  it follows that we can find  $I$  payoff vectors in that set –  $v^i$ ,  $i = 1, \dots, I$  – such that for all  $i$  a)  $v^i \gg 0$ , b)  $v^j_i > v^i_i$ ,  $j \neq i$ , and c)  $v_i > v^i_i$ . That we can find these vectors such that b) is satisfied follows from the dimensionality of the set. That we can additionally get the vectors to satisfy a) and c) follows from the fact that it is a convex set and hence an appropriate “averaging” with a vector such as  $v$  achieves a) while an “averaging” with  $i$ 's worst payoff achieves c). □

Now consider the following strategy: Norm – Start with a time-cycle strategy that generates payoffs after all histories that are within  $\epsilon$  of  $v$ . Choose a high enough  $\delta$  as required. Continue with that strategy if there are no deviations against it. Punishment – If there is, say if player  $i$  deviates, then min-max  $i$  for  $T$  periods and thereafter proceed to the time-cycle strategy that yields payoffs within  $\epsilon$  of  $v^i$  after all histories. Re-start the punishment whenever there is a deviation.

Choose  $T$  in such a fashion that the payoff to the min-max period plus  $v^i_i$  is strictly less than  $v_i$ . That ensures there is no incentive to deviate against the norm provided the punishment is carried out. That there is incentive for

players  $j \neq i$  to punish player  $i$  follows from the fact that  $v^i_j > v^j_j$  the former payoff being what they get from punishing and the latter from not punishing  $i$ . That there is incentive for player  $i$  not to deviate against his own punishment follows from the fact that re-starting the punishment only lowers his payoffs. The theorem is proved. A few remarks are in order.

*Remark 1* If the game horizon  $T$  is finite, there is likely a Folk Theorem along the lines of the result proved for Repeated Games by Benoit and Krishna [4]. To the best of my knowledge it remains, however, an open question.

*Remark 2* When the state space  $A_i$  is infinite or the state space  $S$  is uncountably infinite we again run into technical difficulties. There is an analog to Lemmas 3 and 4 in this instance and under appropriate richer assumptions the results can be generalized – see Dutta (1993). Lemmas 5–7 and the Folk Theorem itself does use the finiteness of  $S$  to apply uniform bounds to various approximations and those become problematical when the state space is infinite. It is our belief that nevertheless the Folk Theorem can be proved in this setting. It remains, however, to be done.

**Dynamics**

Recall that the fourth question is: what can be said about the dynamics of SPE outcome paths? The analogy that might be made is to the various convergence theorems – sometimes also called “turnpike theorems” – that are known to be true in single-player dynamic programming models. Now even within those models – as has become clear from the literature of the past twenty years in chaos and cycles theory for example – it is not always the case that there are regularities exhibited by the optimal solutions. Matters are worse in dynamic games.

Even within some special models where the single-player optima are well-behaved, the SPE of the corresponding dynamic game need not be. A classic instance is the neo-classical aggregative growth model. In that model, results going back fifty years show that the optimal solutions converge monotonically to a steady-state, the so-called “golden rule”. (For references, see Majumdar, Mitra and Nishimura (2000).) However, examples can be constructed – and may be found in Dutta and Sundaram (1996) and Dockner, Long and Sorger (1998) – where there are SPE in these models that can have arbitrarily complex state dynamics which for some range of discount factor values descend into chaos. And that may happen with Stationary Markov Perfect Equilibrium. (It would be less of a stretch to believe that SPE in general can have

complex dynamics. The Folk Theorem already suggests that it might be so.)

There are, however, many questions that remain including the breadth of SPE that have regular dynamics. One may care less for complex dynamic SPE if it can be shown that the “good ones” have regular dynamics. What also remains to be explored is whether adding some noise in the transition equation can remove most complex dynamics SPE.

## Global Climate Change – Issues, Models

### Issues

The dramatic rise of the world’s population in the last three centuries, coupled with an even more dramatic acceleration of economic development in many parts of the world, has led to a transformation of the natural environment by humans that is unprecedented in scale. In particular, on account of the greenhouse effect, *global warming* has emerged as a central problem, unrivaled in its potential for harm to life as we know it on planet Earth. Seemingly the consequences are everywhere: melting and break-up of the world’s ice-belts whether it be in the Arctic or the Antarctic; heat-waves that set all-time temperature highs whether it be in Western Europe or sub-Saharan Africa; storms increased in frequency and ferocity whether it be Hurricane Katrina or typhoons in Japan or flooding in Mumbai. In addition to Al Gore’s eminently readable book, “An Inconvenient Truth”, two authoritative recent treatments are the Stern Review on the Economics of Climate Change, October, 2006 and the IPCC Synthesis Report, November, 2007. Here are three – additional – facts drawn from the IPCC Report:

1. Eleven of the last twelve years (1995–2006) have been amongst the twelve warmest years in the instrumental record of global surface temperatures (since 1850).
2. If we go on with “Business as Usual”, by 2100 global sea levels will probably have risen by 9 to 88 cm and average temperatures by between 1.5 and 5.5°C.

Various factors contribute to global warming, but the major one is an increase in greenhouse gases (GHGs) – primarily, carbon dioxide – so called because they are transparent to incoming shortwave solar radiation but trap outgoing longwave infrared radiation. Increased carbon emissions due to the burning of fossil fuel is commonly cited as the principal immediate cause of global warming. A third relevant fact is:

3. Before the Industrial Revolution, atmospheric CO<sub>2</sub> concentrations were about 270–280 parts per million

(ppm). They now stand at almost 380 ppm, and have been rising at about 1.5 ppm annually.

The IPCC Synthesis (2007) says “Warming of the climate system is unequivocal, as is now evident from observations of increases in global average air and ocean temperatures, widespread melting of snow and ice, and rising global average sea level.” (IPCC Synthesis Report [17]).

It is clear that addressing the global warming problem will require the coordinated efforts of the world’s nations. In the absence of an international government, that coordination will have to be achieved by way of an international environmental treaty. For a treaty to be implemented, it will have to align the incentives of the signatories by way of rewards for cutting greenhouse emissions and punishments for not doing so. For an adequate analysis of this problem one needs a dynamic and fully strategic approach. A natural methodology for this then is the theory of Subgame Perfect (Nash) equilibria of dynamic games – which we have discussed at some length in the preceding sections.

Although there is considerable uncertainty about the exact costs of global warming, the two principal sources will be a rise in the sea-level and climate changes. The former may wash away low-lying coastal areas such as Bangladesh and the Netherlands. Climate changes are more difficult to predict; tropical countries will become more arid and less productive agriculturally; there will be an increased likelihood of hurricanes, fires and forest loss; and there will be the unpredictable consequences of damage to the natural habitat of many living organisms. On the other hand, emission abatement imposes its own costs. Higher emissions are typically associated with greater GDP and consumer amenities (via increased energy usage). Reducing emissions will require many or all of the following costly activities: cutbacks in energy production, switches to alternative modes of production, investment in more energy-efficient equipment, investment in R&D to generate alternative sources of energy, etc.

The principal features of the global warming problem are:

- *The Global Common* – although the sources of carbon buildup are localized, it is the total stock of GHGs in the global environment that will determine the amount of warming.
- *Near-irreversibility* – since the stock of greenhouse gases depletes slowly, the effect of current emissions can be felt into the distant future.
- *Asymmetry* – some regions will suffer more than others.

- *Nonlinearity* – the costs can be very nonlinear; a rise in one degree may have little effect but a rise in several degrees may be catastrophic.
- *Strategic Setting* – Although the players (countries) are relatively numerous, there are some very large players, and blocks of like-minded countries, like the US, Western Europe, China, and Japan. That warrants a strategic analysis.

The theoretical framework that accommodates all of these features is an *asymmetric dynamic commons* model with the global stock of greenhouse gases as the (common) state variable. The next sub-section will discuss a few models which have most of the above characteristics.

### Models

Before presenting specific models, let us briefly relate the climate change problem to the general dynamic game model that we have seen so far, and provide a historical outline of its study. GHGs form – as we saw above – a global common. The study of global commons is embedded in dynamic commons game (DCG). In such a game the state space  $S$  is a single-dimensional variable with a “commons” structure meaning that each player is able to change the (common) state. In particular, the transition function is of the form

$$s(t+1) = q \left( s(t) - \sum_{i=1}^I a_i(t) \right).$$

The first analysis of a DCG may be found in [18]. That paper considered the particular functional form in which  $q(s(t) - \sum_{i=1}^I a_i(t)) = [s(t) - \sum_{i=1}^I a_i(t)]^\alpha$  for a fixed fraction  $\alpha$ . (And, additionally, Levhari and Mirman assumed the payoffs  $\pi_i$  to be logarithmic.) Consequently, the paper was able to derive in closed form a (linear) MPE and was able to analyze its characteristics.

Subsequently several authors – Sundaram [31], Sobel [27], Benhabib and Radner [3], Rustichini [25], Dutta and Sundaram (1992, [14]), Sorger [28] – studied this model in great generality, without making the specific functional form assumption of Levhari and Mirman, and established several interesting qualitative properties relating to existence of equilibria, welfare consequences and dynamic paths.

More recently in a series of papers by Dutta and Radner on the one hand and Dockner and his co-authors on the other, the DCG model has been directly applied to environmental problems including the problem of global warming. We shall describe the Dutta and Radner work in detail and also discuss some of the Dockner, Long and

Sorger research. In particular, the transition equation is identical in the two models (and described below). What is different is the payoff functions.

We turn now to a simplified climate change model to illustrate the basic strategic ideas. The model is drawn from Dutta and Radner [12]. In the basic model there is no population growth and no possibility of changing the emissions producing technologies in each country. (Population growth is studied in Dutta and Radner [11] while certain kinds of technological changes are allowed in Dutta and Radner [10]. These models will be discussed later.) However, the countries may differ in their “sizes”, their emissions technologies, and their preferences.

There are  $I$  countries. The emission of (a scalar index of) greenhouse gases during period  $t$  by country  $i$  is denoted by  $a_i(t)$ . [Time is discrete, with  $t = 0, 1, 2, \dots$ , ad inf.] Let  $A(t)$  denote the global (total) emission during period  $t$ ;

$$A(t) = \sum_{i=1}^I a_i(t). \quad (16)$$

The total (global) stock of greenhouse gases (GHGs) at the beginning of period  $t$  is denoted by  $g(t)$ . (Note, for mnemonic purposes we are denoting the state variable – the amount of “gas” –  $g$ .) The law of motion – or transition function  $q$  in the notation above – is

$$g(t+1) = A(t) + \sigma g(t), \quad (17)$$

where  $\sigma$  is a given parameter ( $0 < \sigma < 1$ ). We may interpret  $(1 - \sigma)$  as the fraction of the beginning-of-period stock of GHG that is dissipated from the atmosphere during the period. The “surviving” stock,  $\sigma g(t)$ , is augmented by the quantity of global emissions,  $A(t)$ , during the same period.

Suppose that the payoff of country  $i$  in period  $t$  is

$$\pi_i(t) = h_i[a_i(t)] - c_i g(t). \quad (18)$$

The function  $h_i$  represents, for example, what country  $i$ 's gross national product would be at different levels of its own emissions, holding the global level of GHG constant. This function reflects the costs and benefits of producing and using energy as well as the costs and benefits of other activities that have an impact on the emissions of GHGs, e. g., the extent of forestation. It therefore seems natural to assume that  $h_i$  is a strictly concave  $C^2$  function that reaches a maximum and then decreases thereafter.

The parameter  $c_i > 0$  represents the marginal cost to the country of increasing the global stock of GHG. Of course, it is not the stock of GHG itself that is costly, but

the associated climatic conditions. As discussed below, in a more general model, the cost would be nonlinear.

Histories, strategies – Markovian strategies – and outcomes are defined in exactly the same way as in the general theory above – and will, hence, not be repeated. Thus associated with each strategy vector  $\sigma$  is a total discounted payoff for each player

$$v_i(\sigma, g_0) \equiv \sum_{t=0}^{\infty} \delta^t \pi_i(t; \sigma, g_0).$$

Similarly, SPE and MPE can be defined in exactly the same way as in the general theory.

The linearity of the model is undoubtedly restrictive in several ways. It implies that the model is unable to analyze catastrophes or certain kinds of feedback effects running back from climate change to economic costs. It has, however, two advantages: first, its conclusions are simple, can be derived in closed-form and can be numerically calibrated; hence may have a chance of informing policy-makers. Second, there is little consensus on what is the correct form of non-linearity in costs. Partly the problem stems from the fact that some costs are not going to be felt for another fifty to hundred years and forecasting the nature of costs on that horizon length is at best a hazardous exercise. Hence, instead of postulating one of many possible non-linear cost functions, all of which may turn out to be incorrect for the long-run, one can opt instead to work with a cost function which may be thought of as a linear approximation to any number of actual non-linear specifications.

Dockner, Long and Sorger (1998) impose linearity in the emissions payoff function  $h$  (whereas in Dutta and Radner it is assumed to be strictly concave) while their cost to  $g$  is strictly convex (as opposed to the above specification in which it is linear). The consequent differences in results we will discuss later.

## Global Climate Change – Results

In this section we present two sets of results from the Dutta and Radner [12] paper. The first set of results characterize two benchmarks – the global Pareto optima, and a simple MPE, called “Business As Usual” and compares them. The second set of results then characterizes the entire SPE correspondence and – relatedly – the best and worst equilibria. Readers are referred to that paper for further results from this model and for a numerical calibration of the model. Furthermore, for the results that are presented, the proofs are merely sketched.

## Global Pareto Optima

Let  $x = (x_i)$  be a vector of positive numbers, one for each country. A *Global Pareto Optimum (GPO)* corresponding to  $x$  is a profile of strategies that maximizes the weighted sum of country payoffs,

$$v = \sum_i x_i v_i, \quad (19)$$

which we shall call *global welfare*. Without loss of generality, we may take the weights,  $x_i$ , to sum to  $I$ .

**Theorem 4** *Let  $\hat{V}(g)$  be the maximum attainable global welfare starting with an initial GHG stock equal to  $g$ . That function is linear in  $g$ ;*

$$\begin{aligned} \hat{V}(g) &= \hat{u} - wg, \\ w &= \frac{1}{1 - \delta \sigma} \sum_i x_i c_i, \\ \hat{u} &= \frac{\sum_i x_i h_i(\hat{a}_i) - \delta w \hat{A}}{1 - \delta}. \end{aligned} \quad (20)$$

*The optimal strategy is to pick a constant action – emission – every period and after all histories,  $\hat{a}_i$  where its level is determined by*

$$x_i h'_i(\hat{a}_i) = \delta w. \quad (21)$$

*Proof* We shall show by dynamic programming arguments that the Pareto-optimal value function is of the form  $\hat{V} = \sum_{i=1}^I x_i [\hat{u}_i - w_i g]$ . We need to be able to find the constants  $\hat{u}_i$  to satisfy:

$$\begin{aligned} \sum_{i=1}^I x_i [\hat{u}_i - w_i g] &= \max_{a_1, \dots, a_I} \sum_{i=1}^I x_i \\ &\cdot \left[ h_i(a_i) - c_i g + \delta (\hat{u}_i - w_i (\sigma g + \sum_{j=1}^I a_j)) \right]. \end{aligned} \quad (22)$$

Collecting terms that need maximization we can reduce the equation above to

$$\begin{aligned} &\sum_{i=1}^I x_i \hat{u}_i \\ &= \max_{a_1, \dots, a_I} \sum_{i=1}^I x_i \left[ h_i(a_i) - \delta w_i \sum_{j=1}^I a_j \right] + \delta \sum_{i=1}^I x_i \hat{u}_i. \end{aligned} \quad (23)$$

It is clear that the solution to this system is the same for all  $g$ ; call this (first-best) solution  $\hat{a}_i$ . Elementary algebra



reveals that

$$\hat{u}_i = \frac{h_i(\hat{a}_i) - \delta w_i \sum_{j=1}^I \hat{a}_j}{1 - \delta} \quad \text{and} \quad w_i = \frac{c_i}{1 - \delta \sigma}.$$

It is also obvious that  $x_i h'_i(\hat{a}_i) = \delta w_i$ , where  $w = \sum_{i=1}^I x_i w_i$ .  $\square$

Theorem 4 states that, independently of the level of GHG,  $g$ , each country should emit an amount  $\hat{a}_i$ . The fact that the optimal emission is constant follows from the linearity of the model in  $g$ . Notice that on account of the linearity in the gas buildup equation – Eq. 17 – a unit of emission in period  $t$  can be analyzed in isolation as a surviving unit of size  $\sigma$  in period  $t + 1$ ,  $\sigma^2$  in period  $t + 2$ ,  $\sigma^3$  in period  $t + 3$ , and so on. On account of the linearity in cost, these surviving units add  $(\sum_i x_i c_i) \times \delta \sigma$  in period  $t + 1$ ,  $(\sum_i x_i c_i) \times (\delta \sigma)^2$  in period  $t + 2$ , and so on, i. e., the marginal lifetime cost is

$$\frac{1}{1 - \delta \sigma} \sum_i x_i c_i,$$

or  $w$ , and that marginal cost is independent of  $g$ .

#### A Markov–Perfect Equilibrium: “Business as Usual”

This MPE shares the feature that the equilibrium emission rate of each country is constant in time, and it is the unique MPE with this property. We shall call it the “Business-as-Usual” equilibrium. Note that in this equilibrium each country takes account of the incremental damage to itself caused by an incremental increase in its emission rate, but does not take account of the damage caused to other countries.

**Theorem 5 (Business-as-Usual Equilibrium)** *Let  $g$  be the initial stock of GHG. For each country  $i$ , let  $a_i^*$  be determined by*

$$\begin{aligned} h'_i(a_i^*) &= \delta w_i, \\ w_i &= \frac{c_i}{1 - \delta \sigma}, \end{aligned} \quad (24)$$

and let its strategy be to use a constant emission equal to  $a_i^*$  in each period; then this strategy profile is a MPE, and country  $i$ 's corresponding payoff is

$$\begin{aligned} V_i^*(g) &= u_i^* - w_i g, \\ u_i^* &= \frac{h_i(a_i^*) - \delta w_i A^*}{1 - \delta}. \end{aligned} \quad (25)$$

The intuition for the existence of an MPE with constant emissions is similar to the analogous result for the GPO solution. (And indeed for that reason the proof will be omitted.) As long as other countries do not make their emissions

contingent on the level of GHGs, country  $i$  has a constant marginal lifetime cost to emissions. And that marginal cost is independent of  $g$ .

#### Comparison of the GPO and Business as Usual

The preceding results enable us to compare the emissions in the GPO with those in the Business-as-Usual MPE:

$$\text{GPO: } h'_i(\hat{a}_i) = \frac{\delta \sum_j x_j c_j}{x_i(1 - \delta \sigma)}, \quad (26)$$

$$\text{BAU: } h'_i(a_i^*) = \frac{\delta c_i}{1 - \delta \sigma}.$$

Since

$$x_i c_i < \sum_j x_j c_j,$$

it follows that

$$\frac{\delta c_i}{1 - \delta \sigma} < \frac{\delta \sum_j x_j c_j}{x_i(1 - \delta \sigma)}.$$

Since  $h_i$  is concave, it follows that

$$a_i^* > \hat{a}_i. \quad (27)$$

Note that this inequality holds except in the trivial case in which all welfare weights are zero (except one). This result is known as the tragedy of the commons – whenever there is some externality to emissions, countries tend to over-emit in equilibrium. In turn, all this follows from the fact that in the BAU equilibrium each country only considers its own marginal cost and ignores the cost imposed on other countries on account of its emissions; in the GPO solution that additional cost is, of course, accounted for. It follows that the GPO is strictly Pareto superior to the MPE for an open set of welfare weights  $x_i$  (and leads to a strictly lower steady-state GHG level for all welfare weights).

One can contrast these results with those in Dockner, Long and Sorger (1998) that studies a model in which the benefits are linear in emission – i. e.,  $h_i$  is linear – but convex in costs  $c_i(\cdot)$ . The consequence of linearity in the benefit function  $h$  is that the GPO and BAU solutions have a “most rapid approach” (MRAP) property – if  $(1 - \sigma)g$ , the depreciated stock in the next period, is less than a most preferred  $g^*$ , it is optimal to jump the system to  $g^*$ . Else it is optimal to wait for depreciation to bring the stock down to  $g^*$ . In other words, linearity in benefits implies a “one-shot” move to a desired level of gas  $g^*$ , which is thereafter maintained, while linearity in cost (as in the Dutta and Radner model) implies a constant emission rate. What is unclear in the Dockner, Long and Sorger model is why

the multiple players would have the same target steady-state  $g^*$ . It would appear natural that, with asymmetric payoffs, each player would have a different steady-state. The existence of a MRAP equilibrium would appear problematical consequently. The authors impose a condition that implies that there is not too much asymmetry.

### All SPE

We now turn to the second set of results – a full characterization of SPE in Dutta and Radner [12]. We will show that the SPE payoff correspondence has a surprising simplicity; the set of equilibrium payoffs at a level  $g$  is a simple linear translate of the set of equilibrium payoffs from some benchmark level, say,  $g = 0$ . Consequently, it will be seen that the set of emission levels that can arise in equilibrium from level  $g$  is identical to those that can arise from equilibrium play at a GHG level of 0. Note that the fact that the set of equilibrium possibilities is invariant to the level of  $g$  is perfectly consistent with the possibility that, in a particular equilibrium, emission levels vary with  $g$ . However, the invariance property will make for a particularly simple characterization of the best and worst equilibria.

Let  $\mathcal{E}(g)$  denote the set of equilibrium payoff vectors with initial state  $g$ , i. e., each element of  $\mathcal{E}(g)$  is the payoff to some SPE starting from  $g$ .

**Theorem 6** *The equilibrium payoff correspondence  $\mathcal{E}$  is linear; there is a compact set  $U \subset \mathbb{R}^I$  such that for every initial state  $g$*

$$\mathcal{E}(g) = U - \{w_1g, w_2g, \dots, w_Ig\}$$

where  $w_i = c_i/(1 - \sigma\delta)$ ,  $i = 1, \dots, I$ . In particular, consider any SPE, any period  $t$  and any history of play up until  $t$ . Then the payoff vector for the continuation strategies must necessarily be of the form

$$v - (w_1g_t, w_2g_t, \dots, w_Ig_t).$$

The theorem is proved by way of a bootstrap argument. We presume that a (candidate) payoff set has this invariance and show that the linear structure of the model confirms the conjecture. Consequently, we generate another candidate payoff set – which is also state-invariant. Then we look for a fixed point of that operator. In other words, we employ the APS operator to generate the SPE correspondence. Since that has already been discussed in the previous section, it is skipped here.

We will now use the above result to characterize the best – and the worst – equilibria in the global climate change game. Consider the *second-best problem* (from

initial state  $g$  and for a given vector of welfare weights  $x = (x_i; i = 1, \dots, I)$ ), i. e., the problem of maximizing a weighted sum of *equilibrium payoffs*:

$$\max \sum_{i=1}^I x_i V_i(g), \quad V(g) \in \mathcal{E}(g).$$

Note that we consider all possible equilibria, i. e., we consider equilibria that choose to condition on current and past GHG levels as well as equilibria that do not. The result states that the best equilibrium *need not* condition on GHG levels:

**Theorem 7** *There exists a constant emission level  $\bar{a} \equiv \bar{a}_1, \bar{a}_2, \dots, \bar{a}_I$  – such that no matter what the initial level of GHG, the second-best policy is to emit at the constant rate  $\bar{a}$ . In the event of a deviation from this constant emissions policy by country  $i$ , play proceeds to  $i$ 's worst equilibrium. Furthermore, the second-best emission rate is always strictly lower than the BAU rate, i. e.,  $\bar{a} < a^*$ . Above a critical discount factor (less than 1), the second-best rate coincides with the GPO emission rate  $\hat{a}$ .*

The theorem is attractive for three reasons: first, it says that the best possible equilibrium behavior is no more complicated than BAU behavior; so there is no argument for delaying a treaty (to cut emissions) merely because the status quo is simple. Second, the cut required to implement the second-best policy is an across the board cut – independently of anything else, country  $i$  should cut its emissions by the amount  $a_i^* - \bar{a}_i$ . Third, the second-best is exactly realized at high discount factors, rather than asymptotically approached as the discount factor tends to 1.

Sanctions will be required if countries break with the second-best policy and without loss of generality we can restrict attention to the worst such sanction. We turn now to a characterization of this worst equilibrium (for, say, country  $i$ ). One definition will be useful for this purpose:

**Definition 1** An  $i$ -less second-best equilibrium is the solution to a second-best problem in which the welfare weight of  $i$  is set equal to zero, i. e.,  $x_i = 0$ .

By the previous theorem, every such problem has a solution in which on the equilibrium path, emissions are a constant. Denote that emission level  $a(x_{-i})$ :

**Theorem 8** *There exists a “high” emission level  $\bar{a}(i)$  (with  $\sum_{j \neq i} \bar{a}_j(i) > \sum_{j \neq i} a_j^*$ ) and an  $i$ -less second-best equilibrium  $a(x_{-i})$  such that country  $i$ 's worst equilibrium is:*

1. Each country emits at rate  $\bar{a}_j(i)$  for one period (no matter what  $g$  is),  $j = 1, \dots, I$ .

2. From the second period onwards, each country emits at the constant rate  $a_j(x_{-i})$ ,  $j = 1, \dots, I$ .

And if any country  $k$  deviates at either stages 1 or 2, play switches to  $k$ 's worst equilibrium from the very next period after the deviation.

Put another way, for every country  $i$ , a sanction is made up of two emission rates,  $\bar{a}(i)$  and  $a(x_{-i})$ . The former imposes immediate costs on country  $i$ . The way it does so is by increasing the emission levels of countries  $j \neq i$ . The effect of this is a temporary increase in incremental GHG but due to the irreversibility of gas accumulation, a permanent increase in country  $i$ 's costs, enough of an increase to wipe out any immediate gains that the country might have obtained from the deviation. Of course this additional emission also increases country  $j$ 's costs. For the punishing countries, however, this increase is offset by the subsequent permanent change, the switch to the emission vector  $\mathbf{a}(x_{-i})$ , which permanently increases their quota at the expense of country  $i$ 's.

### Generalizations

The models discussed thus far are base-line models and do not deal with two important issues relating to climate change – technological change and capital accumulation. Technological change is important because that opens access to technologies that do not currently exist, technologies that may have considerably lower “emissions to energy” ratios, i. e., cleaner technologies. Capital accumulation is important because an important question is whether or not curbing GHGs is inimical to growth. The position articulated by both developing countries like Indian and China as well as by developed economies like the United States is that it is: placing curbs on emissions would restrict economic activities and hence restrain the competitiveness of the economy.

In Dutta and Radner [10,13] the following modification was made to the model studied in the previous section. It was presumed that the actual emission level associated with energy usage  $e_i$  is  $f_i e_i$  where  $f_i$  is an index of (un)cleanliness – or *emission factor* – higher values implying larger emissions for the same level of energy usage. It was presumed that the emission factor could be changed at cost but driven no lower than some minimum  $\mu_i$ . In other words,

$$0 \leq e_i(t), \quad (28)$$

$$\mu_i \leq f_i(t+1) \leq f_i(t). \quad (29)$$

Capital accumulation and population growth is also allowed in the model but taken to be exogenous. The dy-

namics of those two variables are governed by:

$$g(t) = \sigma g(t-1) + \sum_{i=1}^I f_i(t) e_i(t), \quad (30)$$

$$K_i(t+1) = H[K_i(t)], \quad K_i(t) \nearrow \text{ and unbounded in } t, \quad (31)$$

$$P_i(t+1) = \psi_i P_i(t) + (1 - \psi_i) \Psi, \quad P_i(t) \leq \Psi. \quad (32)$$

The output (gross-domestic product) of country  $i$  in period  $t$  is

$$h_i[K_i(t), P_i(t), e_i(t)],$$

where the function  $h_i$  has all of the standard properties mentioned above. The damage due to the stock of GHG,  $g(t)$ , is assumed to be (in units of GDP):

$$c_i P_i(t) g(t).$$

The cost of reducing the emission factor from  $f_i(t)$  to  $f_i(t+1)$  is assumed to be:

$$\varphi_i [f_i(t) - f_i(t+1)].$$

Immediately it is clear that the state variable now encompasses not just the common stock  $g$  but, additionally, the emission factor profile as well as the sizes of population and capital stock. In other words,  $s = (g, f, K, P)$ . Whilst this significant increase in dimensionality might suggest that it would be difficult to obtain clean characterizations, the papers show that there is some separability. The MPE “Business as Usual” has a separable structure – energy usage  $e_i(t)$  and emission factor choice  $f_i(t+1)$  – depend solely on country  $i$ 's capital stock and population alone. It varies by period – unlike in the base-line model discussed above – as the exogenous variables vary. Furthermore, the emission factor  $f_i(t+1)$  stays unchanged till the population and capital stock cross a threshold level beyond which the cleanest technology  $\mu_i$  gets picked. (This bang-bang character follows from the linearity of the model.)

The Global Pareto Optimal solution has similar features – the energy usage in country  $i$  is directly driven by the capital stock and population of that country. Furthermore the emission factor choice follows the same bang-bang character as for the MPE. However, there is a tragedy of the common in that in the MPE (versus the Pareto optimum) the energy usage is higher – at every state – and the switch to the cleanest technology happens later.

## Future Directions

Within the general theory of dynamic games there are several open questions and possible directions for future research to take. On the existence question, there needs to be a better resolution of the case where the state space  $S$  is uncountably infinite. This is not just a technical curiosity. In applications, typically, in order to apply calculus techniques, we take the state variable to be a subset of some real space. The problem is difficult but one hopes that ancillary assumptions – such as concavity and monotonicity – will be helpful. These assumptions come “cheaply” because they are routinely invoked in economic applications.

The characterization result via APS techniques has a similar technical difficulty blocking its path, as the existence question. The folk theorem needs to be generalized as well to the  $S$  infinite case. Here it is our belief though that the difficulty is not conceptual but rather one where the appropriate result needs to be systematically worked out. As indicated above, the study of the dynamics of SPE paths is in its infancy and much remains to be done here.

Turning to the global climate change application, this is clearly a question of utmost social importance. The subject here is very much in the public consciousness yet academic study especially within economics is only a few years old. Many questions remain: generalizing the models to account for technological change and endogenous capital accumulation, examination of a carbon tax, of cap and trade systems for emission permits, of an international bank that can selectively foster technological change, ... There are – as should be immediately clear – enough interesting important questions to exhaust many dissertations and research projects!

## Bibliography

- Abreu D (1988) On the theory of infinitely repeated games with discounting. *Econometrica* 56:383–396
- Abreu D, Pearce D, Stachetti E (1990) Towards a general theory of discounted repeated games with discounting. *Econometrica* 58:1041–1065
- Benhabib J, Radner R (1992) The joint exploitation of a productive asset: A game-theoretic approach. *Econ Theory* 2:155–190
- Benoit J-P, Krishna V (1987) Finitely repeated games. *Econometrica* 53:905–922
- Dockner E, Long N, Sorger G (1996) Analysis of Nash equilibria in a class of capital accumulation games. *J Econ Dyn Control* 20:1209–1235
- Dockner E, Nishimura K (1999) Transboundary boundary problems in a dynamic game model. *Jpn Econ Rev* 50:443–456
- Duffie D, Geanakoplos J, Mas-Colell A, McLennan A (1994) Stationary Markov equilibria. *Econometrica* 62-4:745–781
- Dutta P (1991) What do discounted optima converge to? A theory of discount rate asymptotics in economic models. *J Econ Theory* 55:64–94
- Dutta P (1995) A folk theorem for stochastic games. *JET* 66:1–32
- Dutta P, Radner R (2004) Self-enforcing climate change treaties. *Proc Nat Acad Sci USA* 101-14:5174–5179
- Dutta P, Radner R (2006) Population growth and technological change in a global warming model. *Econ Theory* 29:251–270
- Dutta P, Radner R (2008) A strategic model of global warming model: Theory and some numbers. *J Econ Behav Organ* (forthcoming)
- Dutta P, Radner R (2008) Choosing cleaner technologies: Global warming and technological change, (in preparation)
- Dutta P, Sundaram R (1993) How different can strategic models be? *J Econ Theory* 60:42–61
- Fudenberg D, Maskin E (1986) The Folk theorem in repeated games with discounting or incomplete information. *Econometrica* 54:533–554
- Harris C, Reny P, Robson A (1995) The existence of subgame perfect equilibrium in continuous games with almost perfect information: A case for extensive-form correlation. *Econometrica* 63:507–544
- Inter-Governmental Panel on Climate Change (2007) Climate Change, the Synthesis Report. IPCC, Geneva
- Levhari D, Mirman L (1980) The great fish war: An example using a dynamic cournot-Nash solution. *Bell J Econ* 11:322–334
- Long N, Sorger G (2006) Insecure property rights and growth: The role of appropriation costs, wealth effects and heterogeneity. *Econ Theory* 28:513–529
- Mertens J-F, Parthasarathy T (1987) Equilibria for Discounted Stochastic Games. Research Paper 8750, CORE. University Catholique de Louvain
- Mertens, Neyman (1983)
- Nowak A (1985) Existence of equilibrium stationary strategies in discounted noncooperative stochastic games with uncountable state space. *J Optim Theory Appl* 45:591–603
- Parthasarathy T (1973) Discounted, positive and non-cooperative stochastic games. *Int J Game Theory* 2–1:
- Rieder U (1979) Equilibrium plans for non-zero sum Markov games. In: Moeschlin O, Pallasche D (ed) *Game theory and related topics*. North-Holland, Amsterdam
- Rustichini A (1992) Second-best equilibria for games of joint exploitation of a productive asset. *Econ Theory* 2:191–196
- Shapley L (1953) Stochastic Games. In: *Proceedings of National Academy of Sciences*, Jan 1953
- Sobel M (1990) Myopic solutions of affine dynamic models. *Oper Res* 38:847–53
- Sorger G (1998) Markov-perfect Nash equilibria in a class of resource games. *Econ Theory* 11:79–100
- Stern N (2006) Review on the economics of climate change. HM Treasury, London. [www.sternreview.org.uk](http://www.sternreview.org.uk)
- Stern Review on the Economics of Climate Change, October, (2006)
- Sundaram R (1989) Perfect equilibrium in a class of symmetric dynamic games. *J Econ Theory* 47:153–177

## Earthquake Clusters over Multi-dimensional Space, Visualization of

DAVID A. YUEN<sup>1</sup>, WITOLD DZWINEL<sup>2</sup>,  
YEHUDA BEN-ZION<sup>3</sup>, BEN KADLEC<sup>4</sup>

<sup>1</sup> Dept. of Geology and Geophysics,  
University of Minnesota, Minneapolis, USA

<sup>2</sup> Dept. of Computer Science, AGH University  
of Sci. and Technol., Kraków, Poland

<sup>3</sup> Department of Earth Sciences,  
University of Southern California, Los Angeles, USA

<sup>4</sup> Department of Computer Science,  
University of Colorado, Boulder, USA

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Earthquakes Clustering](#)

[Multidimensional Feature Space](#)

[The Detection and Visualization of Clusters  
in Multi-Dimensional Feature Space](#)

[Description of the Data](#)

[Earthquake Visualization by Using Clustering  
in Feature Space](#)

[Remote Problem Solving Environment \(PSE\)  
for Analyzing Earthquake Clusters](#)

[Future Directions](#)

[Acknowledgments](#)

[Bibliography](#)

### Glossary

**Grid** Virtual metacomputer, which uses a network of geographically distributed local networks, computers and computational resources and services. *Grid Computing* focuses on distributed computing technologies, which are not in the traditional dedicated clusters. *Data Grids* – represent controlled sharing and management of large amounts of distributed data.

**Problem solving environment (PSE)** A specialized computer software for solving one class of problems. They use the language of the respective field and often employ modern graphical user interfaces. The goal is to make the software easy to use for specialists in fields other than computer science. PSEs are available for generic problems like data visualization or large systems of equations and for narrow fields of science or engineering.

**Global seismographic network (GSN)** The goal of the GSN is to deploy permanent seismic recording stations uniformly over the earth's surface. The GSN stations continuously record seismic data from very broad band seismometers at 20 samples per second, and to provide for high-frequency (40 sps) and strong-motion (1 and 100 sps) sensors where scientifically warranted. It is also the goal of the GSN to provide for real-time access to its data via Internet or satellite. Over 75% of the over 128 GSN stations meet this goal as of 2003.

**WEB-IS** A software tool that allows remote, interactive visualization and analysis of large-scale 3-D earthquake clusters over the Internet through the interaction between client and server.

**Scientific visualization** is branch of computer graphics and user interface design that are dealing with presenting data to users, by means of patterns and images. The goal of scientific visualization is to improve understanding of the data being presented.

**Interactive visualization** is a branch of graphic visualization that studies how humans interact with computers to create graphic illustrations of information and how this process can be made more efficient. **Remote-visualization** – the tools for interactive visualization of high-resolution images on remote client machine, rendered and preprocessed on the server.

**OpenGL** A standard specification defining a cross-language cross-platform API for writing applications that produce 2D and 3D computer graphics.

**Sumatra-Andaman earthquake** An undersea earthquake that occurred at 00:58:53 UTC (07:58:53 local time) December 26, 2004, with an epicenter off the west coast of Sumatra, Indonesia. The earthquake triggered a series of devastating tsunamis along the coasts of most landmasses bordering the Indian Ocean, killing large numbers of people and inundating coastal communities across South and Southeast Asia, including parts of Indonesia, Sri Lanka, India, and Thailand.

**Earthquake catalog** Data set consisting of earthquake hypocenters, origin times, and magnitudes. Additional information may include phase and amplitude readings, as well as first-motion mechanisms and moment tensors.

**Pattern recognition** The methods, algorithms and tools to analyze [data](#) based on either [statistical](#) information or on a [priori](#) knowledge extracted from the patterns. The patterns for classification are groups of observations, measurements, objects, defining feature vectors in an appropriate multidimensional feature [space](#).

**Data mining** Algorithms, tools, methods and systems used in extraction of knowledge hidden in a large amount of data.

**Features** denoted  $f_i$  or  $F_j$  ( $i, j$  – feature indices) – a set of variables which carry discriminating and characterizing information about the objects under consideration. The features can represent raw measurements (data)  $f_i$  or can be generated in a non-linear way from the data  $F_j$  (features).

**Feature space** The multidimensional space in which the  $F_k$  vectors are defined. Data and feature vectors represent vectors in respective spaces.

**Feature vector** A collection of features ordered in some meaningful way into multi-dimensional feature vectors  $F_l$  ( $F_l$  where  $l$  – feature vector index) that represents the signature of the object to be identified represented by the generated features  $F_l$ .

**Feature extraction** The procedure of mapping source feature space into output feature space of lower dimensionality, retaining the minimal value of error cost function.

**Multidimensional scaling** The nonlinear procedure of feature extraction, which minimizes the value of the “stress” being the function of differences of all the distances between feature vectors in the source space and corresponding distances in the resulting space of lower dimensionality.

**Data space** The multi-dimensional space in which the data vectors  $f_k$  exist.

**Data vector** A collection of features ordered in some meaningful way into multi-dimensional vectors  $f_k$  ( $f_k, k$  – data vector index) and  $f_k = [m_k, z_k, x_k, t_k]$  where  $m_k$  is the magnitude and  $x_k, z_k, t_k$  – its epicentral coordinates, depth and the time of occurrence, respectively.

**Cluster** Isolated set of feature (or data) vectors in data and feature spaces.

**Clustering** The computational procedure extracting clusters in multidimensional feature spaces.

**Agglomerative (hierarchical) clustering algorithm** The clustering algorithm in which at the start the feature vectors represent separate clusters and the larger clusters are built-up in a hierarchical way. The procedure repeats the process of gluing-up the closest clusters up to the stage when a desired number of clusters is achieved.

**k-Means clustering** Non-hierarchical clustering algorithm in which the randomly generated centers of clusters are improved iteratively.

**Multi-resolutional clustering analysis** Due to clustering a hierarchy of clusters can be obtained. The analysis

of the results of clustering in various resolution levels allows for extraction of knowledge hidden in both local (small clusters) and global (large clusters) similarity of multidimensional feature vectors.

**N-body solver** The algorithm exploiting the concept of time evolution of an ensemble of mutually interacting particles.

**Non-hierarchical clustering algorithm** The clustering algorithm in which the clusters are searched for by using global optimization algorithms. The most representative algorithms of this type is **k-means** procedure.

### Definition of the Subject

Earthquakes have a direct societal relevance because of their tremendous impact on human community [59]. The genesis of earthquakes is an unsolved problem in the earth sciences, because of the still unknown underlying physical mechanisms. Unlike the weather, which can be predicted for several days in advance by numerically integrating non-linear partial differential equations on massively parallel systems, earthquake forecasting remains an elusive goal, because of the lack of direct observations and the fact that the governing equations are still unknown. Instead one must employ statistical approaches (e. g., [61,72,82]) and data-assimilation techniques (e. g., [6,53,81]). The nature of the spatio-temporal evolution of earthquakes has to be assessed from the observed seismicity and geodetic measurements. Problems of this nature can be analyzed by recognizing non-linear patterns hidden in the vast amount of seemingly unrelated information. With the proliferation of large-scale computations, data mining [77], which is a time-honored and well-understood process, has come into its own for extracting useful patterns from large incoherent data sets found in diverse fields, such as astronomy, medical imaging, combinatorial chemistry, bio-informatics, seismology, remote sensing and stock markets [75]. Recent advances in information technology, high performance computing, and satellite imagery have led to the availability of extremely large data sets, exceeding Terabytes at each turn, that are coming regularly to physical scientists who need to analyze them quickly. These data sets are non-trivial to analyze without the use of new computer science algorithms that find solutions with a minimal computing complexity. With the imminent arrival of petascale computing by 2011 in USA, we can expect some breakthrough results from clustering analysis. Indeed, clustering has become a widely successful approach for revealing features and patterns in the data-mining process. We describe the method of using clustering as a tool for analyzing complex seismic data sets

and the visualization techniques necessary for interpreting the results. Petascale computing will also spur visualization techniques, which are sorely needed to understand the vast amounts of data compressed in many different kinds of spaces, with spatial, temporal and other types of dimensions [78]. Examples of clusters abound in nature include stars in galaxies, hubs in airline routes and centers of various human relationships [5]. Clustering comes from multi-scale, nonlinear interactions due to the rock rheology and earthquakes.

## Introduction

Earthquake clustering is automatically implicated by the classical Gutenberg–Richter relationship [40], which specifies the frequency of earthquakes between some small magnitude cutoff and a certain large magnitude around 8 [39]. This empirical finding with a broad magnitude range implies that the largest seismic events are surrounded by a large number of smaller events. This clustering may have both spatial and temporal dependences. One of the goals of earthquake clustering studies is to find these special points in a high-dimensional space related to the nature of the dimensional space associated with earthquake dynamics [24,25]. One major goal of this chapter is to introduce the reader to the notion of searching for clustering points in dimensional spaces higher than the 3D physical space we are used to. This concept is crucial to our understanding of the clustering points of earthquakes in these higher-dimensional spaces, which may enable progress in forecasting earthquakes. Information in seismicity data sets can be both relevant and irrelevant from the point of view of deterministic earthquake dynamics. It can be also “entangled” and impossible to be interpreted with normal human perception. The role of data mining is to have a mathematically rigorous algorithm for extracting relevant information from this deluge of data, and make it understandable. Clustering techniques, which are commonly used today in many fields, ranging from biology (e. g., [26]) to astrophysics, allows us to produce specially crafted data models that can be employed for predicting the nature of future events. In more complex cases, these special data models can work in concert with formal mathematical and physical paradigms to give us deeper physical insight.

The concept of clustering has been used for many years in pattern recognition [2,50,78]. The clustering can use more (e. g. [54]) or less mathematically rigorous principles (e. g. [33]). Nowadays clustering and other feature extraction algorithms are recognized as important tools for revealing coherent features in the earth sciences [32,65,66,

67], bioinformatics [51] and in data mining [37,43,44,57]. Depending on the data structures and goals of classification, different clustering schemes must be applied [36,55].

In this chapter we emphasize the role of clustering in the understanding of earthquake dynamics and the way to visualize and interpret the computed results from clustering. All the seismic events occurring over a certain region during a given time period can be viewed as a single cluster of correlated events. The strength of mutual correlations between events, such as correlations in spatial and time positions along with magnitude, cause this single cluster to have very complex internal structure. The correlations – the measures of similarity between events – divide the global cluster into variety of small clusters of multi-scale nature, i. e., small clusters may consist of a cascade of smaller ones. Coming down the scale we record clusters of more and more tightly correlated events. Exploring the nature of events belonging to a single cluster, we can extract common features they possess. Having more information about events belonging to the same cluster we can derive hidden dependences between them. Moreover, we can anticipate the type of an unknown event belonging to a certain cluster from the character of the other events of this cluster.

In the following sections we describe the idea of clustering and the new idea of higher dimensions associated with data sets. We also demonstrate the results of clustering analysis of both synthetic and real data. Long synthetic data were derived by using a model for a segmented strike-slip fault zone in a 3D elastic half-space [7]. The real data represent short time (5 years interval) seismic activities of the Changbaishan volcano (the north-east frontier of the North China craton) and the Japanese Archipelago. Lastly, we also highlight the role of visualization of clusters as an important tool for understanding this type of new data arrangement, and we describe the role played by remote visualization environment specially devised for visualization of earthquake clusters.

## Earthquakes Clustering

### Statistical Laws as Elementary Building Bricks of Earthquake Models

The earthquake prediction problem is of fundamental importance to society and also geosciences. Progress in this field is hampered, mainly because many important dynamic variables – such as stress – are not accessible for direct observations. Moreover, instrumental observations of seismicity are possible only for a fraction of a single large earthquake cycle. Overcoming these difficulties will require combining analyses of model and observed data by

using knowledge extraction instruments. The fundamental process of knowledge extraction is finding dependences between data and/or between model parameters. They can be revealed as patterns (clusters) in time, spatial and feature (parameter) space domains. The most elementary dependences can be expressed in the form of semi-empirical functional laws.

There are a few basic statistical laws which represent the basis for earthquake models development. The frequency-size statistics of regular tectonic earthquakes (excluding swarms and deep focus earthquakes) follow the Gutenberg–Richter relation [39,80,84]:

$$\log N(M) = a - bM \quad (1)$$

where  $N$  is the number of events with magnitude larger than  $M$  and  $a$ ,  $b$  are constants giving, respectively, the overall seismicity rate and relative rates of events in different magnitude ranges. Observed  $b$ -values of regional seismicity typically fall in the range 0.7–1.3.

Aftershock decay rates are usually be described by the Omori–Utsu law [71,79]:

$$\Delta N/\Delta t = K(t + c)^{-p} \quad (2)$$

where  $N$  is the cumulative number of events,  $t$  is the time after the mainshock, and  $K$ ,  $c$ , and  $p$  are empirical constants. The epidemic-type aftershock-sequences (ETAS) model combines the Omori–Utsu law with the Gutenberg–Richter frequency-magnitude relation for a history-dependent occurrence rate of a point process in the form (e. g., [61])

$$\lambda(t|H_t) = \mu + \sum_{t_i < t} \frac{K_0 \exp[\alpha(M_i - M_c)]}{(t - t_i + c)^p} \quad (3)$$

where  $\alpha$  is a constant background rate,  $M_i$  is the magnitude of earthquake at time  $t_i$ ,  $M_c$  is a lower magnitude cut-off,  $H_t$  denotes the history, and the productivity factor  $K_0 \exp[\alpha(M_i - M_c)]$  gives the number of events triggered by a parent earthquake with magnitude  $M_i$ . The ETAS model is used widely in analysis of seismic data, owing to its built-in clustering associated with the incorporation of the Gutenberg–Richter and Omori–Utsu laws. Examples of recent applications can be found in [45,62,68].

These results can be used to derive additional properties such as average recurrence times (e. g., [4,18,19,20,68,89]). It is usually defined as the number of years between occurrences of an earthquake of a given magnitude in a particular area. For example, the probability of a devastating earthquake striking the greater San Francisco Bay Region over the following 25 years (2007–2031)

is 0.62 [68]. Corral [18,19,20] proposed the existence of a universal scaling law for the probability density function  $H(\tau)$  of recurrence times (or interevent times)  $\tau$  between earthquakes in a given region:

$$H(\tau) \cong \lambda \times f(\lambda\tau) . \quad (4)$$

The function  $f(x)$  appears to be similar for many different seismic regions, which suggests some universal properties. The average rate  $\lambda$  represent the region specific constant, whose reciprocal is the only relevant characteristic time for the recurrence times. Molchan [58] showed that under general conditions, the only universal distribution of inter-event times in a stationary point process is exponential. Hainzl et al. [42] and Saichev and Sornette [68] discussed relations between statistics of interevent times, the ETAS model of triggered seismicity, and the Corral [18,19] distribution of Eq. (4).

In the context of earthquake prediction it is important to analyze earthquake cycles with repeating sequences of events such as foreshocks, mainshocks and aftershocks (e. g., [9,74,80]). Apart from qualitative tendencies reflected by statistical laws, the earthquakes exhibit various types of more subtle spatio-temporal clustering, i. e., grouping of events of the same type both in time and in spatial coordinates. The recognition of these patterns followed by the analysis of the reasons of their appearance may lead to the development of improved prediction algorithms.

In the following section we present a closer look of clustering as a knowledge extraction technique and a possible way of its application to earthquake data analysis.

### Basic Concepts of Clustering

Clustering analysis is a mathematical concept whose main useful role is to extract the most similar (or dissimilar) separated sets of objects according to a given similarity (or dissimilarity) measure [2]. Clustering is one of the most fundamental processes generated by nature. For example, people gathering in groups, tribes, demonstrations, parties, cities, produce clusters. Similarly, towns and cities are clusters of buildings while galaxies are clusters of stars. The local computer networks and bacterial colonies are also clusters. The objects forming clusters can be the clusters of smaller objects, which in turn, are clusters of even smaller and smaller building bricks. The complexity of cluster structure reflects the complexity of the real world. The clusters of various shapes, densities and sizes, with additional attributes as colors, transparency etc. built up patterns, which are the fingerprints of all multi-scale pro-



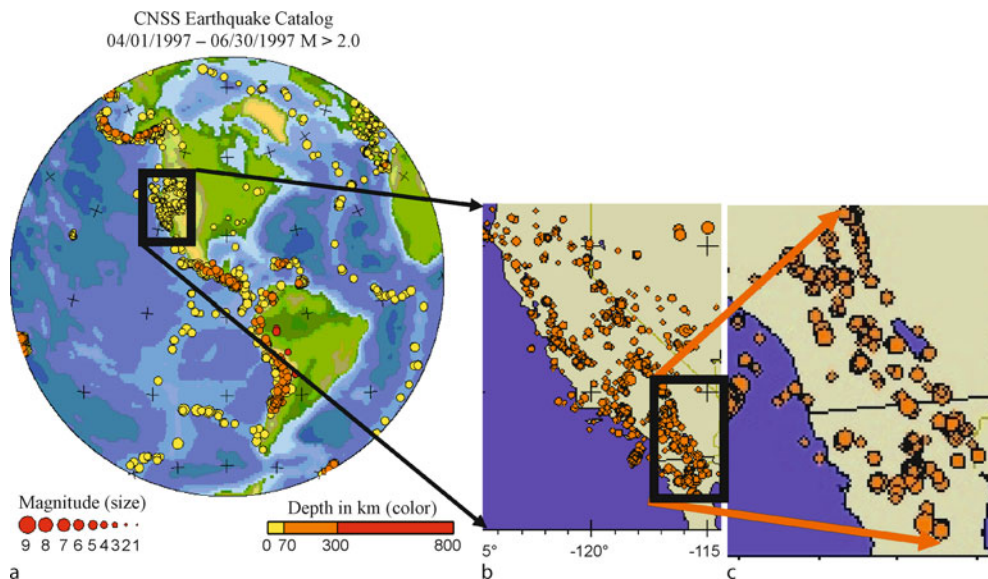
cesses and phenomena. The clusters are the primitives of the patterns.

The same notion of clustering concerns geographical locations and other properties of earthquakes. In Fig. 1a we present a spatial distribution of earthquake epicenters in the western hemisphere of the Earth (data from <http://quake.geo.berkeley.edu/cnss/maps/cnss-map.html>). One can see with the naked eye that their distribution is far from being uniform. We observe both elongated and oblate structures – the earthquake clusters – separated at this resolution by large holes of seismically quiescent area.

Properties of the clusters result from properties of the generating processes. The shape and structure of clusters are visual representation of information on these processes. Therefore, detection of clusters and their analysis is the first step for knowledge extraction from this information. For example, as shown in Fig. 1a, the earthquake clusters on Earth are located in geologically active regions, mainly, on the edges of colliding tectonic plates. The distribution and shape of the earthquake clusters follow the borders between the plates. In Fig. 1b we show the large earthquake cluster from Fig. 1a located at the US western coast. One can distinguish here many smaller clusters of different density separated by geologically inactive area. A similar pattern (see Fig. 1c) is observed by zooming in one of denser clusters from Fig. 1b. This multi-reso-

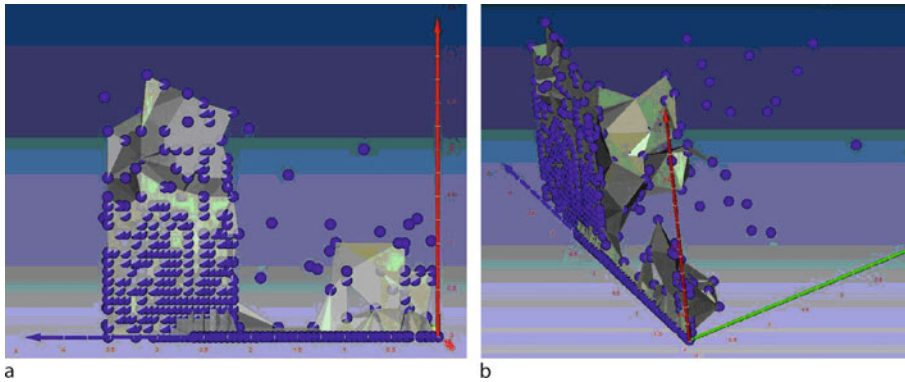
lutional and self-similar system is characteristic for many critical phenomena ▶ *Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of* [3,9,16,74]. The worldwide fault network has a fractal structure (or multifractal) [22,27,79]. Wavelet-based multi-fractal analysis [27] shows clearly several distinct scaling domains in earthquake catalogs revealing rich self-similar multi-scale structure. However, the spatial structure of earthquake clusters alone is inadequate to formulate plausible hypotheses about earthquake dynamics. More information is required.

As shown in Fig. 1, besides the geographical location, earthquakes have additional features such as the time and depth of occurrence and the amount of energy released (proportional to  $10^{\alpha m}$  with  $\alpha \sim 1.5$  and  $m$  the magnitude). These attributes can be used as additional coordinates of, so called, feature space (e. g., [78]). In Fig. 2 we display the earthquake clusters representing the seismic activity nearby the Changbaishan volcano in an abstract 3-D feature space. Apart from geographical location – represented by the distance from the epicenter – other coordinates (features) are employed: the time of occurrence and the magnitude of the earthquake. As shown in Fig. 2, the large cluster of seismic activity is preceded by the small precursory cluster and low activity region. The larger cluster is characterized by the seismic events from broader interval of magnitudes and with satellite earthquakes more



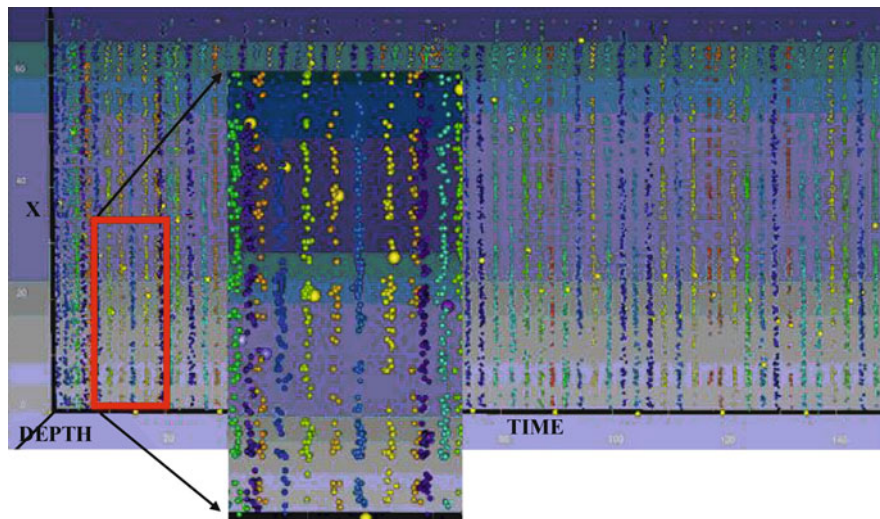
Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 1

Multiscale character of the earthquake clusters. The epicenters of earthquakes of various depth and magnitude are displayed. The data come from the CNSS Earthquake Catalog (<http://quake.geo.berkeley.edu/cnss/maps/cnss-map.html>). a the western hemisphere, b the US western coast c California and Nevada



Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 2

Seismic activity of the Changbaishan volcano during 5 years time span from 07.1999 to 05.2004 (the north-east frontier of the North China craton) [47]. The plates represent the seismic events in 3-D feature space attributed by eruption time (*blue axis*), magnitude (*red axis*) and distance to the epicenter (*green axis*) coordinates. The clusters are rendered using the *wrap point* technique (the Amira visualization package [www.amiravis.com](http://www.amiravis.com)). Two different positions of coordinates are shown. The *large cluster* representing the earthquake swarm is preceded by the *small precursory cluster* of seismic activity and quiescent time period



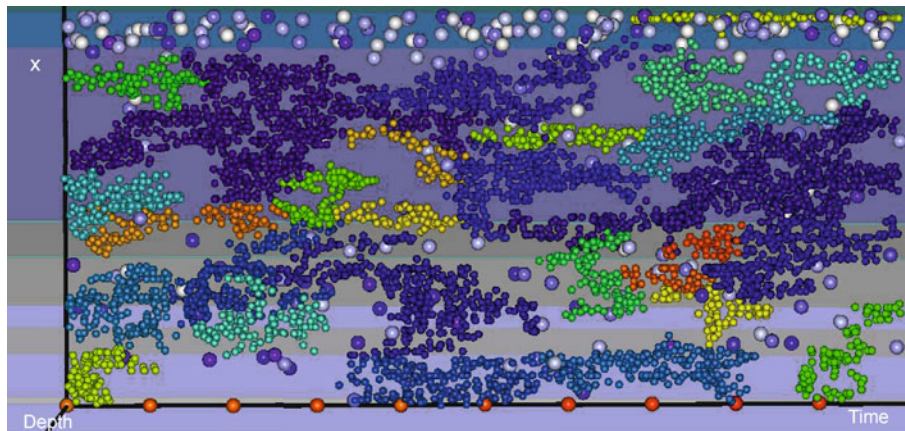
Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 3

The plot reconstructing seismic activity during 150 years from synthetic data [7] (horizontal distance – *X*, depth – *z*; visualized by using the Amira visualization package [1]). Large events (with magnitude  $m > 6$ ) are shown as distinctly larger dots on the background of the lowest magnitude events ( $m < 4$ ). There are visualized patches of low magnitude events preceding larger events [24]. The separate clusters are marked in colors

distant from the epicenter than in the preceding smaller cluster.

The dynamics of the volcanic earthquakes covers only a period of 5 years. The time is too short to conclude about the long-time earthquake dynamics. To obtain data covering much longer time period we used synthetic data generated by numerical simulations of seismicity on a heterogeneous fault governed by 3-D elastic dislocation theory, power-law creep and boundary conditions corresponding

to the central San Andreas Fault [7,28,29]. In Fig. 3 we represent seismic activity during 150 years. This period contains  $M_f \sim 1 - 3 \times 10^4$  events (represented in Fig. 3 by colored dots) in the magnitude interval [3.3–6.8]. Unlike in the Changbaishan case, the seismic events have one more feature – the earthquake depth. Thus the feature space has now four dimensions. In Fig. 3 we display the data distribution in time-depth-position 3-D space. The fourth dimension – the magnitude – is displayed in Fig. 3 by the size



Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 4

The plot reconstructing seismic activity during 1500 years from synthetic data [7]. The largest clusters obtained for events with magnitudes  $4.5 < m < 6$ . Large events ( $m > 6$ ) are shown as distinctly larger plates. The separate clusters are marked in colors

of dot. To make the situation clearer only the large earthquakes with magnitudes  $m > 6$  (large dots) and the smallest ones  $m < 4$  (small dots) are distinguished in Fig. 3. As shown in Fig. 3 and in [24], the synthetic seismic events with magnitudes  $m < 4$  produce stripe-like clusters in the data space. They precede large earthquakes ( $m > 6$ ) and are separated in time by the regions of mixed type of events (i. e., with  $4 < m < 5$ ).

Another system of earthquake clusters are shown in Fig. 4. The synthetic data ( $M_f \sim 10^5$  events) corresponding to the seismic activity during 1500 years were generated by the same model [7] for similar geological and boundary conditions. Only medium size events with  $4.5 < m < 6$  were taken for clustering. In addition to the local strip like clusters of smaller events ( $m < 4$ ) detected for 150-years data, one can observe in Fig. 4a distinct spatio-temporal patchwork structure of clusters of medium sized events ( $4.5 < m < 6$ ). These clusters follow spatio-temporal changes in strength-stress properties of the fault in the region simulated.

In summary, we can highlight very fundamental properties of earthquakes, multi-resolutional clusters are built up by the earthquake epicenters. The clustering is a dynamical process involving many spatio-temporal scales. The dynamic nature of earthquake clusters in a very long time horizon is obvious because everyone can expect that tectonic plates will change dynamically the geo-mechanical properties of the Earth crust. In a long time period covering thousands years, the patterns from Fig. 1 will evolve following the changes in the fault network. More mysterious is the character of earthquake dynamics in spatio-temporal scales allowing for making realistic predictions. We show that in the medium-time period lasting more than

a hundred of years the seismic events may produce periodic system of clusters in approximately equal time intervals with increasing and decreasing seismic activity. The large earthquakes, preceded by the quiescent time periods, appear. The short-time dynamics reveal additionally, that the earthquake swarms are signaled by the smaller precursory cluster of seismic activity. The earthquake attributes such as the magnitude, and the epicenter depth, allow for better interpretation of emerging clusters and exploration of hypotheses space. Therefore, for studying various aspects of earthquake dynamics, including their prediction, we have to analyze the cluster structures in multi-dimensional feature space, to be sure that none of important information will be lost or neglected.

### Multidimensional Feature Space

In Fig. 1 every point  $i$  representing one out of  $M_f$  earthquakes has two dimensions – the geographical coordinates  $\mathbf{x}_i = [x_1, x_2]$  of the epicenter. The point can be treated as 2-D vector  $\mathbf{f}_i$  in the feature space where  $\mathbf{f}_i = \mathbf{x}_i$ . Assuming additional coordinates, at the highest level of resolution, a single seismic event  $i$  can be represented as a five-dimensional data vector  $\mathbf{f}_i = [m_i, z_i, \mathbf{x}_i, t_i]$  where  $m_i$  is the magnitude and  $\mathbf{x}_i, z_i, t_i$  – its epicentral coordinates, depth and the time of occurrence, respectively. The spatio-temporal clusters can be extracted by 3-D visualization similar as those of Figs. 3, 4 distinguishing extra dimension by the size of dots and colors. Only clusters in the three spatially visualized dimensions can be extracted, while the other attributes associated with the earthquake characteristics are used for discriminating among the different types of clusters.

As shown in Figs. 3, 4 and in [24], at the lowest resolution level we can analyze the data locally by looking for clusters with similar events. However, considering a single event on a given area as a feature vector [78] cannot be a good approach from a generalization point of view. The number of events is usually large. There are many noisy background events, which destroy the relevant clusters or produce artificial ones. Moreover, the clustering of raw data neglects the important statistical information, which concerns the entire inspected area. An alternative approach exists in which the entire seismic area can be described as a multidimensional feature vector evolving in time. In the following these features will represent descriptors  $a_k$  (seismicity parameters) corresponding to different statistical properties of all the events measured in a given time interval. The number of descriptors  $N$  defines the dimensionality of the feature vector  $F_i = [a_1, a_2, \dots, a_N]$ ,  $i = 1, 2, \dots, M$ . The vector represents not a single seismic event but it corresponds to seismic situation on the whole controlled area in the subsequent time interval indexed by  $i$ . The number of feature vectors  $M$  is equal to the number of time intervals in which the descriptors are computed. The index  $i$  is a discrete equivalent of time. We expect that the features vectors representing different moments of time also have the tendency to produce clusters in the abstract  $N$ -dimensional feature space. Monitoring changes of these time-series in abstract  $N$ -dimensional space may be used as a proxy for the evolution of stress and a large earthquake cycle on a heterogeneous fault [9].

To explain this approach better, let us assume that we have to analyze the client behaviors in a hypermarket. We can watch every client separately assuming that it can be defined as a feature vector consisting of only two coordinates: the time he entered the shop, money spent. Then we can try to find clusters emerging with time during a shopping day. This cannot be easy due to both a large number of feature vectors (clients) producing statistical noise and lack of correlations between them. Another approach consists in treating as a feature vector not a single client but every subsequent time interval  $t_i = i \times \Delta T$  ( $i = 0, 1, 2, \dots, M_F$ ;  $t_i < t_e$ ;  $M_F = (t_e - t_b)/\Delta T$  and  $t_b$  - beginning of the working day and  $t_e$  - closing time). Let the coordinates of the subsequent feature vector define the following descriptors averaged in  $\Delta t$ : the number of people inside the shop (crowding), the flow, items bought, money spent per person. We note that now the number of feature vectors will be substantially smaller than in the previous approach but the dimensionality of feature space is larger. Let us assume that as a result of clustering we extract two distinct clusters. The first one consists of feature vec-

tors (time intervals) from between 10.00–11.00 and 13.00–14.00. The cluster is characterized by very small values of the first three descriptors (crowding, flow, numbers of items sold) and relatively large expenses. The second cluster consists of time intervals from between 8.00–9.00 and 16.00–18.00 with all descriptors large. We could conclude that the first cluster consists of time intervals from shopping hours that are the favorite for wealthy retired people from the rich village in the neighborhood, while the second cluster is associated with the rush in shopping just before before the beginning and after the end of working hours.

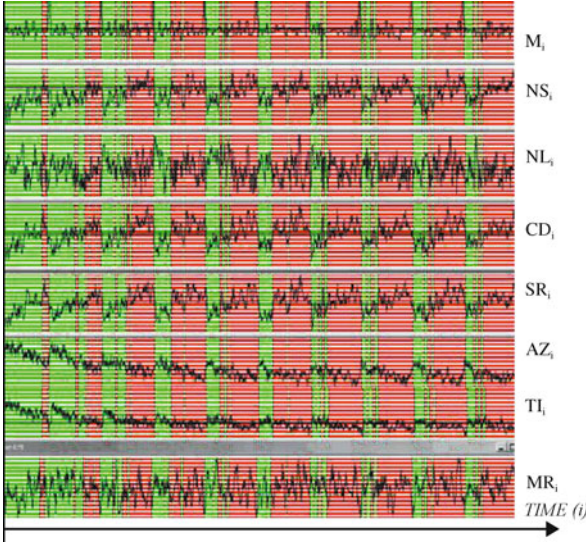
In the same way, the clusters of feature vectors (time intervals) consisting of seismicity parameters should reflect the similarity between seismic activities in various time intervals. As we show before the large seismic events are preceded by precursory events, reflected by an abnormal seismic activity in the whole area. We suppose that these moments of time are similar in the context of the set of seismicity parameters selected. Thus the feature vectors corresponding to precursory events should belong to the same cluster. The idea of predictive system based on clustering consists in detecting the clusters of former foreshocks and signal if the current feature vector – which represents current seismic situation over the area – is or is not the member of this clusters.

The seismicity parameters are computed as time and space averages in a given time and space intervals within a sliding time window with a length  $\Delta T$  and time step  $dt$ . The values of  $a_k$  represents one of the following seismicity parameters: **NS**, **NL**, **CD**, **SR**, **AZ**, **TI**, **MR**. The value of  $dt$  was assumed to be equal to the average time difference between two recorded consecutive events while  $\Delta T$

Earthquake Clusters over Multi-dimensional Space, Visualization of, Table 1

#### Definition of seismicity parameters

<b>NS</b>	Degree of spatial non-randomness at short distances. The differences between distributions of event distances and distances between randomly distributed points.
<b>NL</b>	Degree of spatial non-randomness at short distances.
<b>CD</b>	Spatial correlation dimension calculated on the basis of correlation integrals and on interevent distances.
<b>SR</b>	Degree of spatial repetitiveness represents the tendency of events with similar magnitudes to have nearly the same locations of hypocenters.
<b>AZ</b>	Average depth of the earthquake occurrence.
<b>TI</b>	Inverse of seismicity rate – time interval in which a given (constant) number of events occurs.
<b>MR</b>	Ratio of the numbers of events falling into two different magnitude ranges = $M_r(m \geq M_0)/M_r(m < M_0)$ .



Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 5

The exemplary set of seismicity parameters  $\{M, NS, NL, CD, SR, AZ, TI, MR\}$  in time ( $i$  – subsequent number of the feature vector) for a file from the 1500-years synthetic data catalog (from [25]). The green and red strips show the time moments belonging to the two different clusters. The green cluster corresponds to the time intervals of lower while the red cluster of higher seismic activities. The time series represent about  $M_F = 10^3$  feature vectors  $F_i$

is equal to about 1/10 of the average time distance between two successive large events ( $m > 6$  or  $m > 5$ ). By increasing the values of  $dt$  and  $\Delta T$  one can obtain smoother time series due to better statistics. On the other hand, poorer prediction characteristics can be expected then. We define the seismicity parameters as shown in Table 1 [28,29].

The seismicity parameters produces seven time series and create the abstract 7-dimensional feature space of time events  $F_i = (NS_i, NL_i, CD_i, SR_i, AZ_i, TI_i, MR_i)$  where  $i$  are discretized values of time  $t = t_b + i\Delta T$ . In Fig. 5 we display an example set of seismicity parameters (with average magnitude  $M$ ) for synthetic data [25]. The precise location of the clusters and the visualization of the clustering results are significant challenges in clustering over multi-dimensional space. In the following section we present briefly the basics of clustering and algorithms needed in this venture.

### The Detection and Visualization of Clusters in Multi-Dimensional Feature Space

Our main challenge is to devise a clustering scheme which can divide the  $M$  feature vectors  $x_i$ ,  $i = 1, 2, \dots, M$  into  $k$  separate groups (clusters). More formally, assuming that  $X$

$= \{x_i\}_{i=1, \dots, M}$  and  $x \in \mathbf{R}^N$ ;  $x_i = \{x_{i1}, x_{i2}, \dots, x_{iN}\}$  we define as an  $k$ -clustering of  $X$ , i. e., the partition of  $X$  into  $k$  clusters  $C_1, \dots, C_k$  provided three conditions are met:

- $C_i \neq \emptyset$ ,  $i = 1, \dots, k$  – the clusters are non empty sets,
- $\cup_{i=1, \dots, k} C_i = X$  – the sum of elements inside clusters is equal to the total number of feature vectors,
- $C_i \cap C_j = \emptyset$ ,  $i \neq j$ ,  $j = 1, \dots, k$  – each feature vector belongs to only one cluster.

The computational problem with clustering is that the number of possible clustering of  $M$  vectors into  $k$  groups is given by the Stirling numbers (very large numbers) of the second kind:

$$S(M, k) = \frac{1}{k!} \sum_{i=1}^k (-1)^{k-i} \binom{k}{i} \cdot i^M. \quad (5)$$

Some values of  $S(N, k)$  are:  $S(15, 3) \approx 2 \times 10^6$ ,  $S(20, 4) \approx 45 \times 10^9$ ,  $S(25, 8) \approx 7 \times 10^{17}$ ,  $S(100, 5) \approx 2 \times 10^{68}$ . Knowing that the value of  $N$  in typical clustering problems can be  $10^2$  to  $10^9$  and more we see that the clustering problem is intrinsically hard and exhaustive search – looking through all possible clusterings – cannot be considered. The special clustering schemes based on the proximity measures between feature vectors have to be exploited. The basic steps must be followed in order to develop a clustering task are the following:

1. *Feature selection* – Features must be properly selected to encode as much information as possible. Parsimony and minimum redundancy among the features is a major goal.
2. *Proximity measure* – This is the measure how “similar” (or “dissimilar”) two features vectors are.
3. *Clustering criterion*, which depends on the interpretation of the term “sensible”, depending on the type of clusters expected in the data set e. g., oblate, elongated, “bridged”, circular etc.
4. *Clustering algorithms*. Choose a specific algorithmic scheme that unravels the clustering structure of the data set.
5. *Validation and interpretation of results* are the final processes of clustering.

There are two principal types of clustering algorithms: non-hierarchical and agglomerative schemes [2,50,78].

### Clustering Techniques

The non-hierarchical clustering algorithms are used mainly for extracting compact clusters by using global

knowledge about the data structure. The well known *k-means* based schemes [78], consist in finding the global minimum of the following goal function:

$$J(w, z) = \sum_j \sum_{i \in C_j} |x_i - z_j|^2, \quad (6)$$

where:  $z_j$  is the position of the center of mass of the cluster  $j$ , while  $x_i$  are the feature vectors closest to  $z_j$ . To find a global minimum of function  $J()$ , one repeats many times the clustering procedures for different initial conditions [48]. Each new initial configuration is constructed in a special way from the previous results by using the methods from [48,87]. The cluster structure with the lowest  $J(w, z)$  minimum is selected.

Agglomerative clustering schemes consist in the subsequent merging of smaller clusters into the larger clusters, basing on proximity and clustering criteria. Depending on the definition of these criteria, there exist many agglomerative schemes such as: average link, complete link, centroid, median, minimum variance and nearest neighbor algorithm. The hierarchical schemes are very fast for extracting localized clusters with non-spherical shapes. The proper choice of proximity and clustering criteria depend on many aspects such as dimensionality of data. For example, a smart clustering criterion based on linked-list scheme for finding neighbors used for molecules clustering is completely worthless for clustering  $N$ -dimensional data for which it has extremely high computational complexity. All of agglomerative algorithms suffer from the problem of not having properly defined control parameters, which can be matched for the data of interest and hence can be regarded as invariants for other similar data structures.

Majority of the classical clustering algorithms require knowledge on the number of clusters. However this number is usually unknown a priori. Furthermore, these methods do not perform well in the presence of heavy noise or outliers. Recently, new methods have been proposed that can: deal with noisy data, discover non-spherical clusters and allow for automatic assessment of number of clusters. Some important examples are the Chameleon [55], DBSCAN [70] and CURE [38] algorithms. Unfortunately, these methods are suited only for low dimensional data and are rather inefficient limiting their use for data mining of large-scale sets. For clustering of large data sets of multidimensional data other approaches are in great demand. In the innovative work by Frey and Dueck [33] the authors use the concept of “affinity propagation,” which takes as input measures of similarity between pairs of data points. Real-valued messages are exchanged between data points until a high-quality set of exemplars and corresponding

clusters gradually emerges. Affinity propagation promises to find clusters with much lower error than other methods, and it can do this in less than one-hundredth the amount of time.

Clustering schemes do not produce univocal results. For low dimensional 2-3-D spaces human eye can decide whether the clustering result is optimal or not. However, it becomes hopeless for higher dimensions. There exist many techniques for visualization multidimensional clusters. One of them is the multi-dimensional scaling (MDS) (see overview of mapping techniques in [73]) – the most powerful non-linear mapping technique. This method allows for visualization of the multidimensional data in 2-D or 3-D and for interactive extraction of clusters.

### Multidimensional Scaling

Multi-dimensional scaling or MDS is mathematically a non-linear transformation of  $N$ -dimensional data onto  $n$ -dimensional space, where  $n \geq N$  [23,73,78]. The MDS algorithm bases on the “stress function” criterion. The goal is to maintain all the distances between points  $\mathbf{R}_i \in \omega \subset \mathfrak{R}^N$  in the Euclidean 3-D (or 2-D) space with a minimum error. The “stress function” can be written as follows:

$$E(\omega, \omega') = \sum_{j < i} s_{i,j}^{w \cdot m_i} \cdot (s_{i,j} - s'_{i,j})^{m_i} = \min$$

where:  $s'_{i,j} = (\mathbf{y}_i - \mathbf{y}_j) \cdot (\mathbf{y}_i - \mathbf{y}_j)$ ,  $i, j = 1, \dots, M$ , (7)

and  $D_{i,j}$  – is a squared distance between points  $\mathbf{R}_i, \mathbf{R}_j \in \omega \subset \mathfrak{R}^N$  and  $\mathbf{r}_i, \mathbf{r}_j \in \omega' \subset \mathbb{E}^3$  – coordinates of the respective points in 3-D Euclidean space. The values of  $w$  and  $m_i$  are the parameters of transformation.

The result of mapping depends on the quality of the minimum obtained for the “stress function”. Usually the dimensionality of the “stress function” domain is very high and is equal to  $N \cdot M$ , i. e., thousands, in the smallest and billions in large problems. For more than  $M = 10^3$  feature vectors, the high dimensionality of source space and data complexity may cause the resulting low dimensional patterns to be completely illegible. The application of standard numerical algorithms for finding global minimum of this multimodal, non-linear and complex criterion becomes hopeless. Therefore, for visualization of  $M > 10^3$  multidimensional data samples, more reliable minimization techniques extracting global minimum of the “stress function” are required. In [23] we proposed N-body solver

by ODE's as a heuristic means. The algorithm is as follows:

1. The initial configuration of  $M$  interacting "particles" is generated in  $\mathbb{E}^3$ ,
2. Every "particle" corresponds to the respective  $N$ -dimensional point from  $\mathfrak{R}^N$ ,
3. The "particles" interact with each other with  $\Phi_{i,j}$  particle-particle potential:

$$V_{i,j} = \frac{1}{4} \cdot k \left( r_{i,j}^2 - a_{i,j}^2 \right)^2 \quad (8)$$

( $k$  – is the stiffness factor) and the energy produced is dissipated by the friction force proportional to the velocity of the particles.

4. The system of particles evolves according to the Newtonian equations of motion.

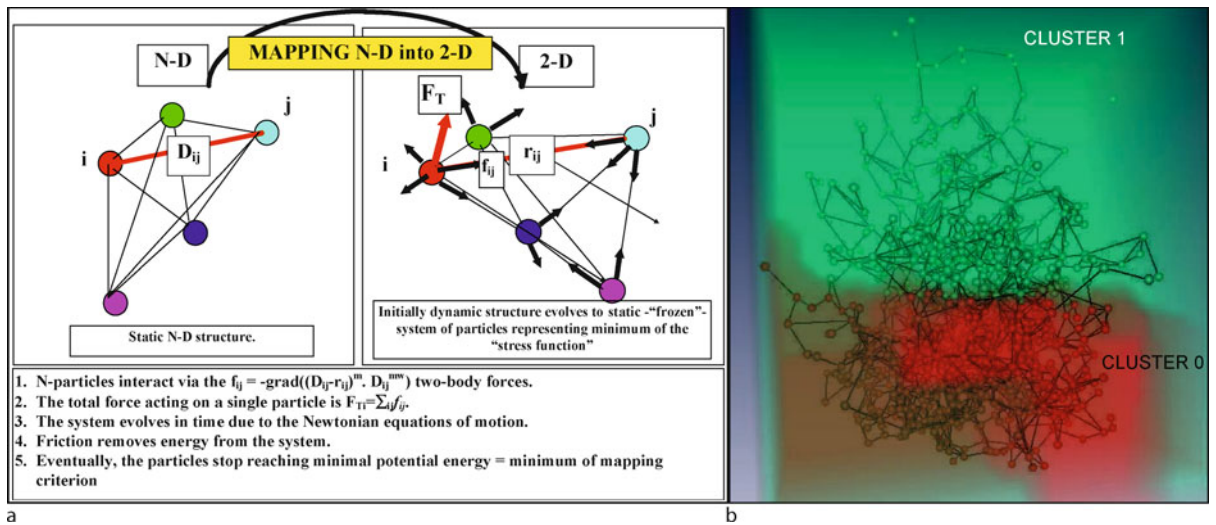
In this way the interactions between each pair of particles are described by various spring like potentials, dependent on the separation distance between particles  $r_{ij}$  and the distance  $D_{ij}$  between respective multidimensional points in  $\mathfrak{R}^N$ . If the distance between particles  $i$  and  $j$  in the output 2(3)-D space is smaller than the distance between respective  $i$  and  $j$  feature vectors in the source  $N$ -D space these points repel one another. Otherwise, i. e. the distance is larger, the particles attract one another. By using the *leap-frog* numerical scheme for time-integration [41] the fol-

lowing formula for velocities and positions of "particles" can be derived from the Newtonian equations:

$$\begin{aligned} \mathbf{v}_i^{n+1/2} &= \frac{(1-\varphi)}{(1+\varphi)} \cdot \mathbf{v}_i^{n-1/2} \\ &+ \frac{\alpha \Delta t}{(1+\varphi)} \left\{ \sum_{j=1}^K (r_{i,j}^n)^2 - a_{i,j}^2 \right\} \mathbf{r}_{i,j}^n + \frac{\mathbf{g}}{\alpha} \mathbf{i}_z \quad (9) \\ \mathbf{r}_i^{n+1} &= \mathbf{r}_i^n + \mathbf{v}_i^{n+1/2} \cdot \Delta t \\ \alpha &= \frac{k}{m}, \quad \varphi = \frac{\lambda}{2m} \cdot \Delta t, \end{aligned}$$

where  $\mathbf{v}_i^n$ , — the particle  $i$ ,  $n$  – the time-step number,  $m = 1$  – particle mass.

As it is common in molecular dynamics [41], the system of "particles" evolves in time until the global (or close to the global) minimum of Eq. (8) (the total potential energy of the particle system) is gained. Two free parameters,  $\lambda$  and  $k$ , have to be fit to obtain the stable state, where the final positions of frozen "particles" reflect the result of  $N$ -D to 3-D mapping. The conceptual scheme of MDS exploiting N-body solver is shown in Fig. 6A. In Fig. 6B we present the feature vectors shown in Fig. 5, using the 7-dimensional feature space which has been transformed by using the MDS procedure and mapping onto the 3-D space. Take a look on the movies (Movie 1 and 2 in Supplementary Materials), which shows how rotation in the 3-D space can help in cluster recognition.



Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 6

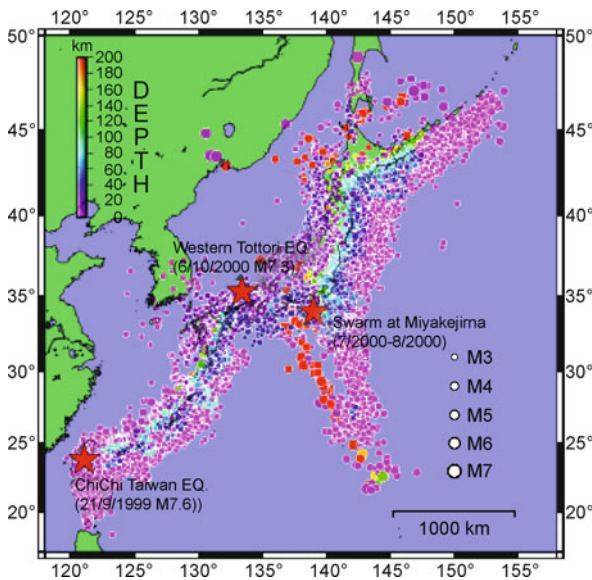
**a** The conceptual diagram of MDS transformation, **b** The clusters from Fig. 6 mapped by using multidimensional scaling into 3D space for synthetic seismic data catalog A covering 1500 years

## Description of the Data

### Natural Datasets

We analyze the observed and synthetic earthquake catalogs for three time intervals of 5, 150 and 1500 years respectively. The observed data (Fig. 7) represents seismic activities of the Japanese islands collected by the Japan Meteorological Agency (JMA).

The JMA Catalog consists of 915,829 events detected in Japan Islands between 1923 and January 31, 2003. The original catalog includes also events with magnitudes less than 1.0. The lowest magnitudes were determined by using a detection level, estimated from the Gutenberg–Richter frequency-size distribution. We have assumed that the cutoff magnitude of earthquake is equal to 3 ( $m > 3$ ). We do not use any cutoff depth of hypocenter events. The seismic events shown in Fig. 7, were recorded during the 5 years time interval from October 1, 1997 to January 31, 2003. The data set processed consists of  $M = 42\,370$  seismic events with magnitudes  $m$ , position in space (latitude  $X$ , longitude  $Y$ , depth  $z$ ) and occurrence time  $t$ .



Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 7

Seismic activities around the Japanese Archipelago with a time period of 5 years. We use the hypocentral data provided by the Japan Meteorological Agency (JMA). The magnitude of the earthquakes (JMA magnitude) and their depths are represented by differences of the radius of the circle and colors, respectively. The red stars symbolize large events such as: Chi-Chi Taiwan earthquake (21/9/1999 M7.6 latitude 23.8 longitude 121.1), Swarm at Miyakejima (7/2000-8/2000 latitude 34.0 longitude 139.0), Western Tottori earthquake (6/10/2000 M7.3 latitude 35.3 longitude 133.4) (from [25])

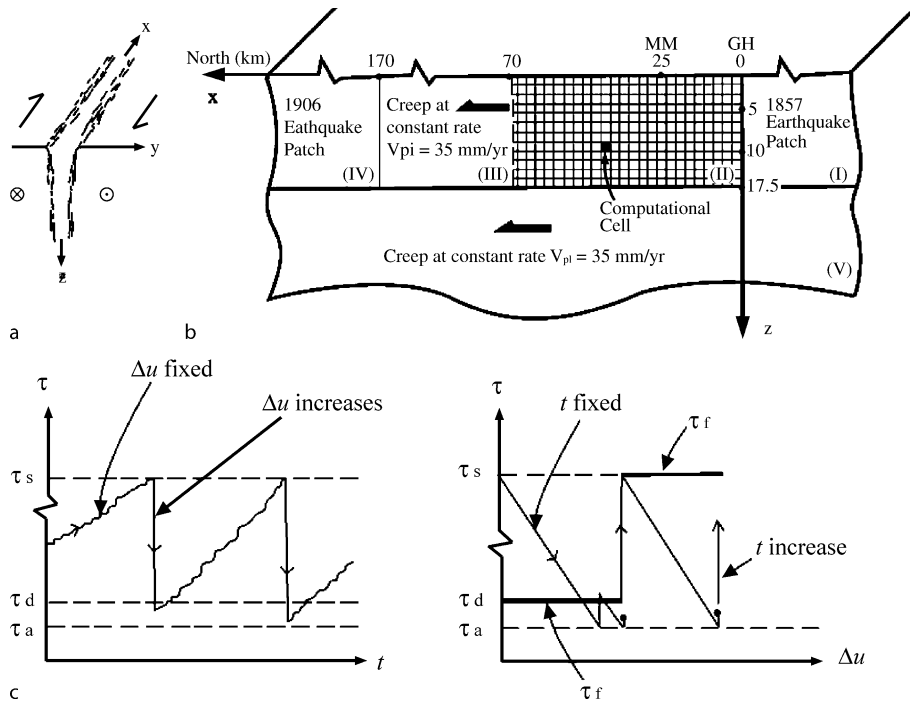
To analyze the seismic activity in longer time periods, we use data from synthetic catalogs generated by numerical earthquake models [7].

### Physical Model of Earthquake Dynamics

The synthetic catalogs are generated by the model of Ben-Zion [7] for a segmented strike-slip fault zone in a 3D elastic half-space, based on earlier developments of Ben-Zion and Rice [10,11]. The model attempts to account for statistical properties of earthquake ruptures on long and narrow fault zones with bends, offsets, etc (Fig. 8a), represented by a cellular structure in a 2D plane with discrete cells and spatial variations of frictional parameters (Fig. 8b). The model contains a computational grid (region II of Fig. 8b) where evolving stress and seismicity are generated in response to ongoing loading imposed as slip boundary conditions on the other fault regions. Regions III and V creep at constant plate velocity of 35 mm/yr, while regions I and IV follow staircase slip histories with recurrence times of 150 yr. The stress transfer due to the imposed boundary conditions and failing grid cells is calculated by using a discretized form of a boundary integral equation and employing the static solution for dislocations in a 3D elastic half-space [10,63].

Deformation at each computational cell is the sum of slip contributions from brittle and creep processes. The brittle process (Fig. 8c) is governed by distributions of static friction  $\tau_s$ , dynamic friction  $\tau_d$ , and arrest stress  $\tau_a$ . The static friction characterizes the brittle strength of a cell until its initial failure in a given model earthquake. When stress  $\tau$  at a cell reaches the static friction, the strength drops to the dynamic friction for the remaining duration of the event. The stress at a failing cell drops to the arrest level  $\tau_a$ , which may be lower than  $\tau_d$  to accommodate dynamic overshoot, producing local slip governed by dislocation theory [17,63]. The static friction, dynamic friction, and arrest stress are connected via a dynamic overshoot coefficient  $D = (\tau_s - \tau_a)/(\tau_s - \tau_d)$ . If the stress transfer from failing regions increases the stress at other cells to their static or dynamic strength thresholds, as appropriate, these cells fail and the event grows. When the stress at all cells is below the brittle failure thresholds, the model earthquake ends and the strength at all failing cells recovers back to  $\tau_s$ . The creep process is governed by a power-law dependence of creep-velocity on the local stress and space-dependent coefficients that increase exponentially with depth and with distance from the southern edge of the computational grid. The chosen parameters produce an overall “pine-tree” stress-depth profile with a “brittle-ductile” transition at a depth of about 12.5





Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 8  
The schematics of the model of Ben-Zion [7] for a segmented strike-slip fault zone in a 3D elastic half-space

km, and variable stress-along-strike profiles with a gradual “brittle-creep” transition near the boundary between regions II and III (see Ben-Zion [7] for additional details). The model generates many realistic features of seismicity compatible with observations, including frequency-size and temporal event statistics, hypocenter distribution with depth and along strike, intermittent criticality, accelerated seismic release, scaling of source time functions and more (e. g., [9,29,56,88]).

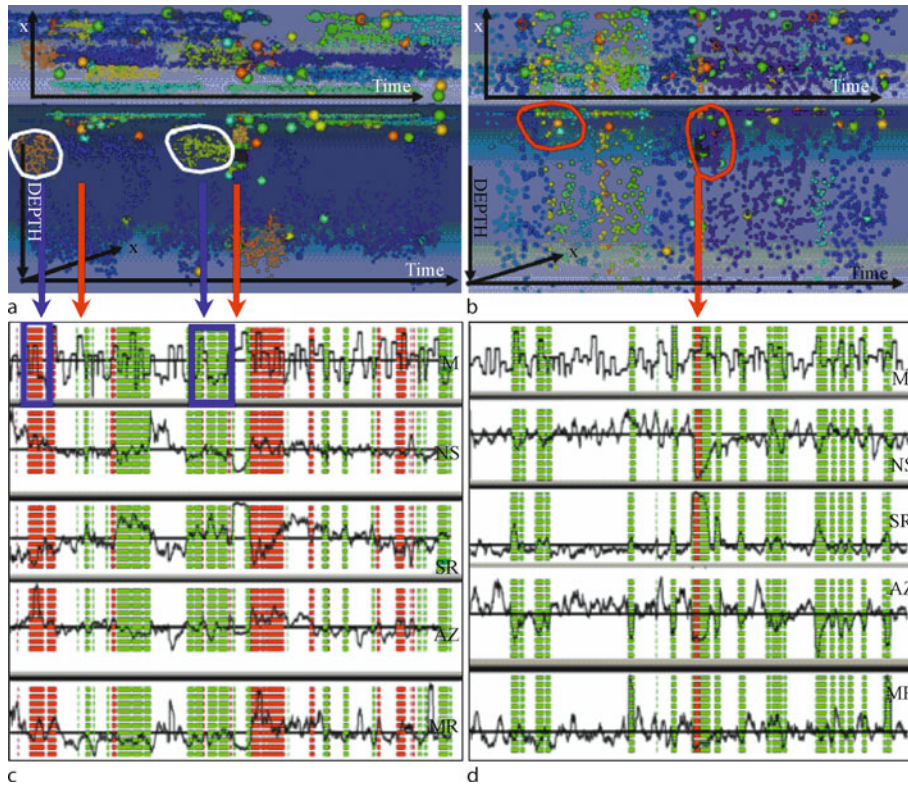
### Synthetic Catalogs

Synthetic data generated by computational models can comprise many events covering large spatial areas and extremely long time spans. Moreover, the synthetic data retain the statistical reliability of the results. The data are free of measurement errors, which occur in estimating earthquake magnitudes and hypocentral locations, and do not suffer from incomplete recording of small events, which exist in natural catalogs. These are significant advantages for our study, which attempts to illustrate clearly the performance of clustering analysis and visualization techniques.

In Sect. “Description of the Data” we analyze synthetic catalogs generated by two model realizations (A and M) of Ben-Zion [7]. The catalogs contain the time, location

and magnitude of earthquakes calculated by the model for 150 and 1500 years. Extensive numerical simulations with several different classes of models, summarized by Ben-Zion [8] and Zöller et al. [9], suggest that the degree of disorder in fault heterogeneities is a tuning parameter of the earthquake dynamics. Catalog A is generated by a model realization tailored to the Parkfield section of the San Andreas fault. Catalog M is generated by a realization of a more-disordered system like the San Jacinto fault or the Eastern California Shear Zone in Southern California. In both data sets the time interval covers all events ( $M \sim 1 - 3 \times 10^4$ ) that have occurred in the last 150 years of simulated fault activity. These simulations were repeated for ten times larger time scale i. e. 1500 year interval (the number of events  $M \sim 10^5$ ) covering hundreds of large earthquakes ( $m > 6$ ) and correspondingly wider time window.

The seismicity parameters were obtained by averaging the data using a sliding time window of constant width  $\Delta T$  and shift  $dt$ . We employ  $\Delta T = 10$  days and  $dt = 2$  days for the Japanese data,  $\Delta T = 10$  months and  $dt = 2$  months for the 150-years synthetic data and  $\Delta T = 30$  months and  $dt = 6$  months for the data covering 1500 years time period. Each parameter in the clustering was normalized with respect to the standard deviation.



Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 9

Real seismic data [49] analyzed by using clustering in both the data **a,b** and the feature **(c,d)** spaces. In panels **a** and **b** one can see the results of clustering in the data space (from two different perspectives,  $X$ -Time and Depth-Time) for small ( $3 < m < 4$ ) and medium magnitude ( $4 < m < 6$ ) events, respectively, represented by small dots. The different colors of the dots denote different clusters. Large events are visualized by the larger spheres. Their colors show the difference in magnitudes  $m$  (red – the largest, green – the smallest). The clusters in panels **a,b** encircled in red display the places with the largest seismic activity, while those in white represent the clusters of small precursory events. The red, white and green stripes in panel **c** and **d** representing 4 (out of 7) seismic parameters and maximum magnitude  $M$  show the clusters of similar time events for situations corresponding to panels **a** and **b**, respectively. The Amira visualization package was used (<http://www.amiravis.com>)

## Earthquake Visualization by Using Clustering in Feature Space

### Short-Time Period

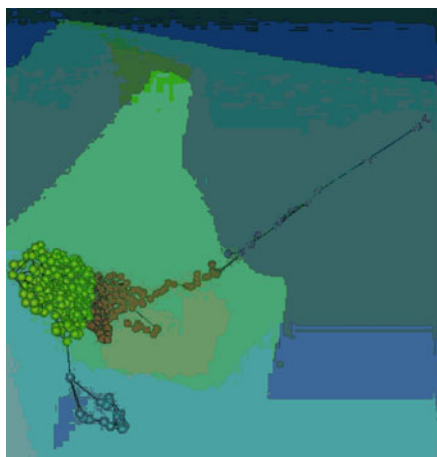
Results of clustering of the observed Japanese seismic catalogs (see Fig. 7) both in raw data and in feature spaces are shown in Fig. 9. At the data resolution level a single seismic event  $i$  can be represented as a multi-dimensional data vector  $f_i = [m_i, z_i, X_i, Y_i, t_i]$  where:  $m_i$  is the magnitude,  $X_i$  – the latitude,  $Y_i$  – the longitude,  $z_i$  and  $t_i$  – the depth and the time of occurrence, respectively. The seismic events are visualized with the Amira package in Fig. 9a,b as irregular clouds of colored dots with  $(z, x, t)$  coordinates.

In accordance with the Gutenberg–Richter relationship, we find that the number of events from various ranges of magnitudes differs considerably, and divide the

entire set of data onto three subsets. The first one comprises the small, the second medium and the last one represents the largest earthquakes displayed in Fig. 9a,b as big dots. The deepest earthquakes  $z > 150$  km are not displayed in the Fig. 9. The various shades represent the magnitudes of earthquakes from  $m = 6$  (green) to  $m = 7$  (red). In Fig. 9 we present the clustering results in both the data  $f_i$  and the feature  $F_i$  spaces. We look for clusters of similar seismic events (data space) and time events (feature space). The dots (data vectors), belonging to the same clusters, have the same color. The Fig. 9a,b is very rich in cluster-like forms, some of them hard to interpret. Correspondence of the cluster structure of data  $f_j (j = 1, \dots, M_f)$  with the clusters of averaged events  $F_i (i = 1, \dots, M_F)$  in the feature space can reveal interesting information. As one can see from the panel C, only three clusters are obtained in the feature space consist-

ing of small data events ( $3 < m < 4$ ). The green cluster corresponds to two relatively large time intervals of small events preceding Miyakejima earthquake and many smaller post shock periods. The time events  $F_i$  from this cluster represent averaged data events  $f_j$ , mainly shallow (AZ) of high degree of spatial repetitiveness SR and small diversity of magnitudes (MR). The red cluster consists of deeper events of smaller repetitiveness, and more diversified in magnitudes. The larger time interval of this type of behavior is recognized just after Miyakejima shock. The white cluster is not interesting in this scale of small events and includes all other events (including the earthquake swarm).

In panel D we display the seismicity parameters, which form three clusters of time events obtained for seismic events of larger magnitude  $4 < m < 6$ . Clusters of these events have different structure than in the previous case. They are parallel to X-depth plane. The borders between clusters roughly correspond to the borders of successive showers of the earthquakes. The red cluster comprises only the earthquakes corresponding to the Miyakejima swarm encircled in red in Fig. 9b. As we can see by the MDS visualization displayed in Fig. 10a, this cluster is made up from a needle of time events sprouting away from the two remaining and oval clusters. The green cluster in Fig. 9d represents the deep events, diversified in magnitude of high repetitiveness and rather high degree of spatial non-randomness at short distances (NS). As shown in Fig. 9, these time events represent mainly the post-swarm series of shocks. The white cluster, as before, includes all the other events.



### Time Period of 150 Years

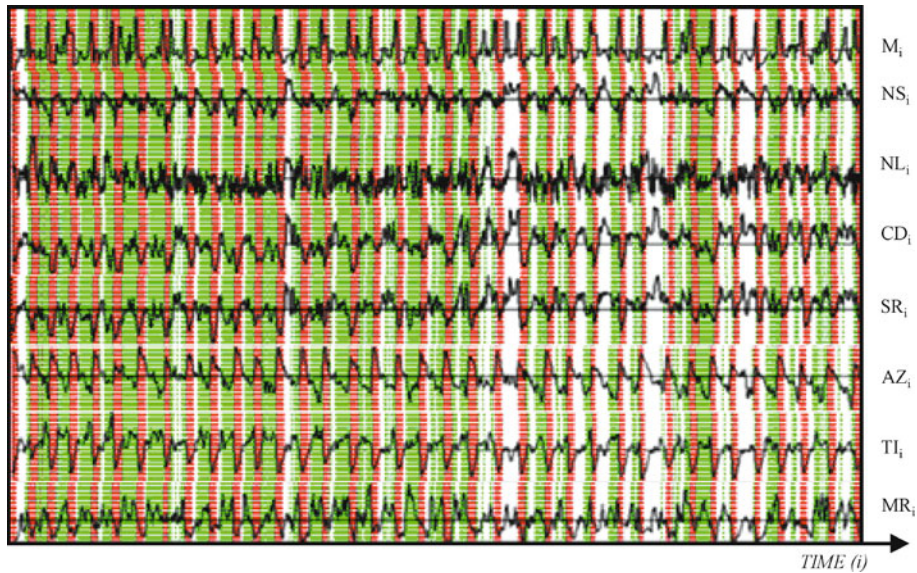
In Fig. 11 we display the time series of seismicity parameters computed for the complete synthetic data catalog A. These time series follow the situation from Fig. 3 where dots represent separate data events. The green, red and white strips in Fig. 11 separate 3 clusters of similar time events represented by 7-dimensional feature vectors. In Fig. 10b these clusters are visualized due to the MDS transformation of 7-dimensional feature space into 3-D space. In Fig. 10b each dot represents a 7-dimensional feature vector mapped into 3-D space by MDS transformation. From the top panel of Fig. 11 displaying the largest events  $M$  in the sliding time window, we may conclude that the white (blue in Fig. 10b) and red clusters from Figs. 10, 11b comprise time events, which correspond to the aftershock effects. The white cluster represents the net aftershock events, while the red one includes the earthquake effects averaged in sliding time window. Conversely, the green cluster (yellow in Fig. 10b) contains the time events preceding the earthquakes.

The selectivity in time of the seismicity parameters depends on the width  $\Delta T$  and shift  $dt$  of the sliding time window. Due to space and time averaging, it is impossible to correlate precisely the appearance of an earthquake with the rest of the seismicity parameters when two earthquakes are too close to each other. Therefore, the sequence of green-red-white cluster events can be broken (Fig. 11) into time domains with many large earthquakes. As shown in Fig. 11 the occurrence of the largest events correlates well with the minima of NS, CD, SR, TI, and maxima of



Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 10

The clusters from feature space mapped into 3D space for a realistic short-time interval seismic data. The small blue cluster at the bottom represents the events at the end of the time interval, which are averaged within a shrinking time window. b The synthetic seismic data catalog A covering 150 years



Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 11  
 The seismicity parameters {*M*,*NS*,*NL*,*CD*,*SR*,*AZ*,*TI*,*MR*} in time for synthetic data catalog A (from [25])

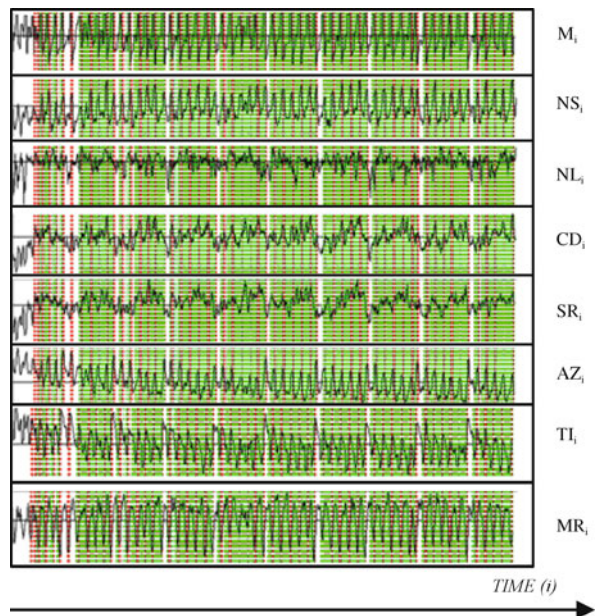
*AZ*, *MR* parameters. This means that the occurrence of large earthquakes is preceded by increasing spatial diffusion of events and increasing seismicity rate. Moreover, the results confirm the some findings from the real data in a shorter time-scale:

1. The events preceding large earthquakes are shallow and have small magnitudes. They have also higher degree of spatial repetitiveness than events from different clusters.
2. The earthquakes accompanying and following the mainshock are rather large in magnitude, deep, have high seismicity rates and low spatial correlation dimension (this drops off rapidly at the onset of large events),

The analysis of synthetic data shows clearly that the clusters in the feature space reflect well the periodicity of increasing and decreasing seismic activity in a given area. For this scale, however, the fine scale characteristics of precursory and after-shock effects become fuzzy.

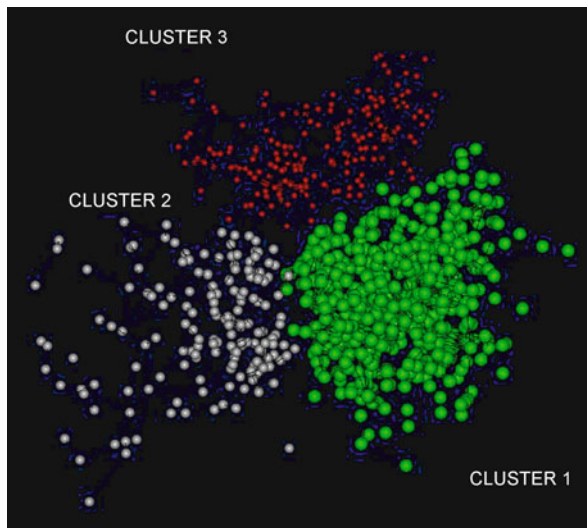
**Time Period of 1500 Years**

In Figs. 4, 5, 6b and Figs. 12, 13 we visualize the feature vectors for data covering 1500 year period for two models: the A model with a Parkfield-type asperity and the M model with multi-size-heterogeneities. In Fig. 4 and Fig. 12 one can recognize two types of clusters with different sizes. The larger cluster comprises feature vectors forming approximately 150-year long periodic time intervals, which



Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 12  
 The seismicity parameters with time for synthetic (catalog M) for seismic data representing time interval of 1500 years. The red and green strips depict the events belonging to red and green clusters from Fig. 7b, respectively

are represented by red strips in Fig. 4 and by green strips in Fig. 12. The second cluster consists of feature vectors from periodic gaps colored in green in Fig. 4 and in white



Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 13

The clusters from Fig. 12 mapped by using multidimensional scaling from a 7-dimensional feature space into 3D space for synthetic seismic data catalog M covering 1500 years

in Fig. 12. This anomalous cluster corresponds to the periodic changes in the character of seismic activities. The third cluster (see Fig. 13), marked in red for M type of data in Fig. 12, consists of periodic and short time intervals representing rapid bursts of seismic activity within every 150-year interval.

In both A and M models the gaps between 150-year long intervals are correlated with decrease of: the correlation dimension (CD), degree of spatial repetitiveness (SR) and seismicity rate. These gaps are preceded by large earthquakes. The simulations used for generating the datasets incorporate imposed large earthquakes on regions (I) and (IV) of Fig. 8b that bound the computational grid (region II), as staircase boundary conditions with a step at every 150 years. The analysis detected the effects of these boundary conditions on the seismicity that is calculated in region II.

There are also evident differences between the A and M data in the time intervals belonging to the second cluster. For A environment the gaps between 150-year intervals are greater and the secondary periodicity within them is less clear. Moreover, within time intervals from the second cluster, the degree of spatial non-randomness decreases at long distances (NL) for M model while for A data it decreases at short distances (NS). In addition, the average depth of earthquakes (AZ) is then clearly larger for A model, while for M data it remains on the average level.

In sum, by means of analyzing earthquake clusters in feature space over long time-scale, we can investigate important characteristics of seismic activity such as:

1. The occurrence of hierarchical time-periodicity in seismic activity caused by increase of short-time correlations and their destruction, respectively. Correlations can be broken both due to short-wave and long-wave resonances of the Poincare type (e. g. during largest earthquakes) [Sornette, 2004].
2. The dependence of seismic activities on the ambient rheological and geological properties of the environment, which strongly modify the cluster structure of the feature vectors.

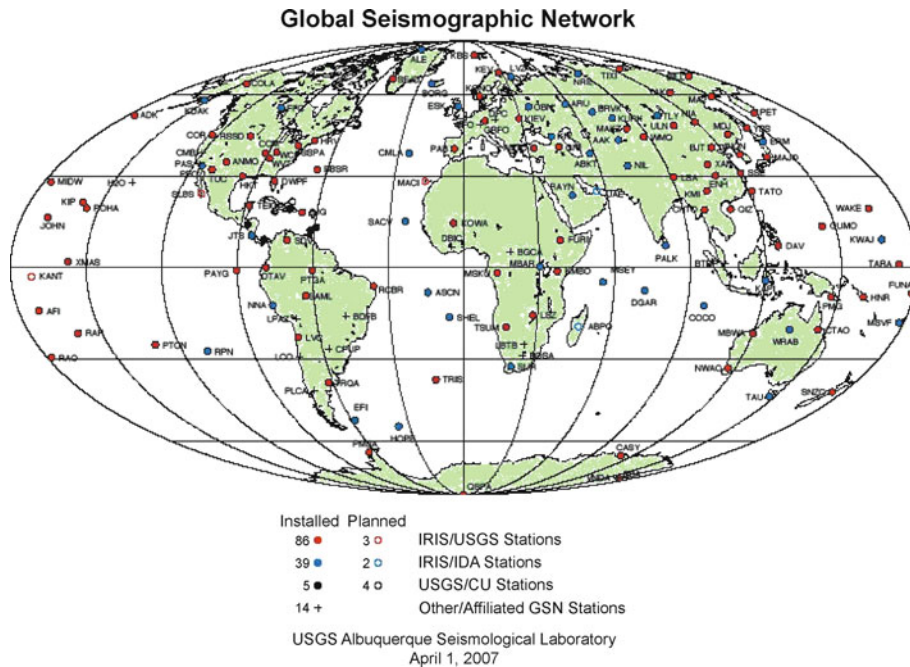
### Remote Problem Solving Environment (PSE) for Analyzing Earthquake Clusters

#### Need for Remote Visualization and Analysis

We need fast access to large databases in order to forecast earthquakes by observation of similarities between thousands and millions of seismic events by visualization of earthquake clusters. The largest earthquake catalogs comprise TBytes of data. Taking into account also the data from tsunami earthquakes and micro-earthquakes in mines, the total amount of data collected by seismic centers spread all over the world is humongous. Moreover, knowledge extraction of earthquake precursors may demand exploration of cross-correlation relationships among many different catalogs. Therefore, both fast communication between data centers and large disk spaces are sorely needed.

As shown in Fig. 14, earthquake seismograph stations, which collect earthquake data from regions with high seismic activity, are distributed worldwide. Therefore, the unprocessed data needs to be stored and then transferred to a dedicated remote server for data processing. After processing, the results must be returned to data acquisition centers and/or other clients. Broadband access to remote facilities dedicated specifically to pattern recognition and visualization allows for scrutinizing local data catalogs by using peer-to-peer connections of data acquisition centers to data preprocessing servers. Clients in the network can automatically compare various types of earthquake catalogs, data measured in distant geological regions, and the results from various theoretical and computational models. By comparing data accessible in the network we have a chance to eliminate the environmental factors and to extract the resultant earthquake precursory effects.

Integration of a variety of hardware, operating systems, and their proper configuration results in many com-



Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 14  
Worldwide distribution of earthquake seismographic stations (© USGS)

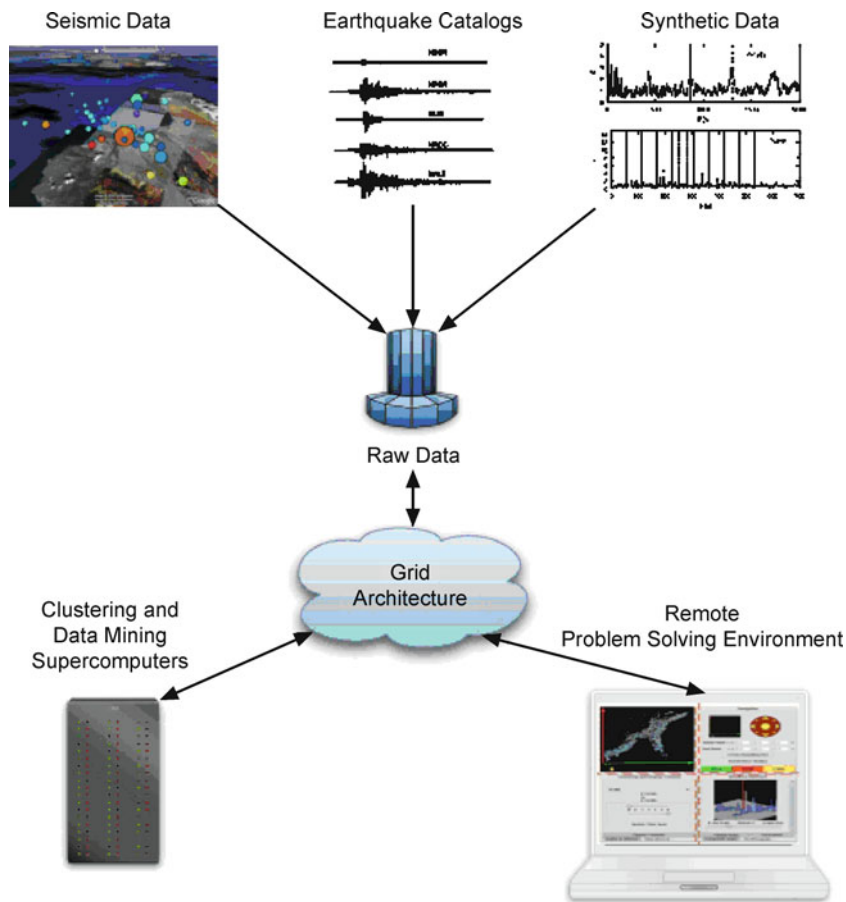
munication problems between data centers. Efficient, reliable, and secure integration of distributed data and software resources, such as pattern recognition and visualization packages, is possible only within the GRID paradigm of computing [13,31]. The GRID mode of computing has flourished rapidly in recent years and has facilitated collaboration and accessibility to many types of resources, such as large data sets, visualization servers and computing engines. Scientific teams have developed easy-to-use, flexible, generic and modular middleware, enabling today's applications to make innovative use of global computing resources. Remote access tools were also produced to visualize huge datasets and monitor performance and data analysis properties, effectively steering the data processing procedures interactively [21]. The TeraGrid project (<http://www.teragrid.org>) is a successful high-performance implementation of such a GRID infrastructure and is being used as an integrated, persistent computational resource at universities and laboratories across the USA. The TeraGrid development impacts also the earthquake science. The National Science Foundation has awarded the Southern California Earthquake Center 15 million service units of computer processing time on supercomputers nationwide [*Grid Today*, August 2007]. These computational resources will be used for simulating thousands of possible earthquakes scenarios in Southern California, includ-

ing the largest breaks on the San Andreas fault ([www.scec.org/cybershake](http://www.scec.org/cybershake)). SCEC will be able to simulate the most disastrous earthquakes ( $M > 7$ ), such as events that could produce Katrina-scale disasters.

We discuss the idea of an integrated problem-solving environment (PSE) intended for the analysis of earthquake clusters for the prediction of earthquakes. A simplified scheme for data acquisition and visualization of earthquake clusters is displayed in Fig. 15. This system promotes portability, dynamic results on-demand, and collaboration among researchers separated by long distances by using a client server paradigm. This is provided through a lightweight front-end interface for users to run locally while the a remote server takes care of intensive processing tasks on large databases, off-screen rendering, and data visualization.

### Grid Environment

In general, large datasets and high-performance computing resources are distributed across the world. When collaboration and sharing of resources are required, a computational GRID infrastructure needs to be in place to connect these servers (see, e.g., [15]). There must exist protocols available to allow clients to tap into these resources and harness their power. The computational grid can be

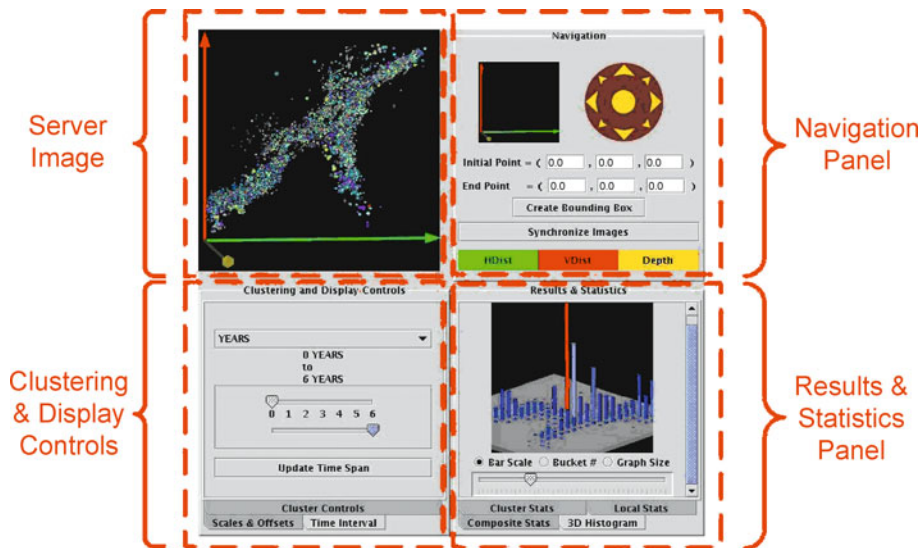


Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 15  
Data acquisition, storage, processing, and remote problem solving environments

seen as a distributed system of “clients”, which consists of either “users” or “resources” and proxies. A GRID can be implemented using an event brokering system designed to run on a large network of brokering nodes. Individually, these brokering nodes are competent servers, but when connected to the brokering system, they are able to share the weight of client requests in a powerful and efficient manner. Examples of this include GRID Resource Brokering [30] and NaradaBrokering.

These GRID architectures are well suited to the functionality of a PSE for earthquake cluster analysis and as an integrated computational environment for data exchange and common ventures. The seismic data centers from the networking point of view represent a complex hierarchical cluster structure. They are located geographically in the regions of high seismic activity within heavily populated areas of economic importance. Therefore, the seismic data centers create distant superclusters of various “density” of computational resources corresponding to the

size and importance of the regions. These superclusters are sparse in the sense of computational resources devoted for earthquake detection and data acquisition. However, these same structures contain important computational, scientific and visualization facilities with strong interest in the analysis of earthquake data and earthquake modeling. The efficient interconnection of these sites is of principal interest. Due to the “small world network” structure of GRID architectures it is possible to select the most efficient routing schemes, considerably shortening the average communication path length between brokers. GRID architectures are appropriate to link the clients, both users and resources, together. Construction of efficient and user friendly Problem Solving Environments requires integration of data analysis and visualization software within the GRID environment, in such a way that it can be easily accessed via the Internet. We created an integrated data interrogation toolkit to act as a PSE for visualization and clustering of seismic data, which we call WEB-IS.



Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 16  
 WEB-IS is an example of a remote earthquake clustering PSE

### Example of Remote PSE

WEB-IS is a software tool that allows remote, interactive visualization and analysis of large-scale 3-D earthquake clusters over the Internet [85] through the interaction between client and server. WEB-IS acts as a PSE through a web portal used to solve problems by visualizing and analyzing geophysical datasets, without requiring a full understanding of the underlying details in software, hardware and communication [34,52]. As shown in Fig. 16, the primary goal of WEB-IS in the geosciences is to provide middleware that sits between the modeling, data analysis tools and the display systems that local or remote users access. In the case of large and physically distributed datasets, it is necessary to perform some preprocessing and then transmit a subset of the data to one or more processes or visualization servers to display. The details of where and how the data migrates should be transparent to the user. WEB-IS makes available to the end users the capability of interactively exploring their data, even though they may not have the necessary resources such as sufficient software, hardware or datasets at their local sites. This method of visualization allows users to navigate through their rendered 3-D data and analyze for statistics or apply earthquake cluster analysis. To the client, the process of accessing and manipulating the data appears simple and robust, while the middleware takes care of the network communication, security and data preparation.

Complete realization of an earthquake clustering PSE consists of:

1. Data analysis tools to implement earthquake clustering techniques;
2. High performance visualization techniques using OpenGL or Amira;
3. The Grid environment;
4. Integration toolkit, such as WEB-IS.

These exist and can work both independently and coupled in a single special purpose system. This system can be developed creating the backbone of the sophisticated computational data acquisition environment, which can be devised specifically for earthquake clustering or for general needs of the geophysical community. Equipped with only PDAs or laptops, and working on location in unreachable desert terrains with remote data acquisition centers or perhaps just analyzing data in one of the many computation facilities located around the globe, geophysicists will be enabled unlimited access to data resources spread all over the world.

We see the principal goal of our work in contributing to the construction of a global warning system, which can be used for prediction of catastrophes such as various types of earthquakes along the circum Pacific belt, where there is a great concentration of people. For example, similar methodology can be used for tsunami earthquake alerting. Theoretical models of faulting and seismic wave propagation used for the computation of radiated seismic energy from broad-band records at teleseismic distances [14] can be adapted to the real-time situation when neither the depth nor the focal geometry of the source



is known accurately. The distance-dependent approximation was used in [60]. By analyzing some singular geophysical parameters such as the energy-to moment ratio  $H$  [60] for regular earthquakes, the results obtained from the theoretical models agree well with values computed from available source parameters (e.g., as published by the National Earthquake Information Center). It appears however that the so called “tsunami earthquakes” – characterized by the significant deficiency of moment release at high frequencies – yield the values of  $H$  considerably different the regular earthquakes. Thus  $H$  value can be used as a suitable criterion for discriminating various types of earthquakes in a short duration of time, like an hour. However, this hypothesis holds only for a few cases. For, so called, “tsunamigenic earthquakes” this difference is not so clear. Moreover, the value of the moment computed on the base of long-period seismic waves can be underestimated. For example, analysis of the longest period normal modes of the Earth, 0S2 and 0S3, excited by the December 26, 2004 Sumatra earthquake [76], yields an earthquake moment of  $1.3 \cdot 10^{30}$  dyn-cm, approximately three times larger than the  $4 \cdot 10^{29}$  dyn-cm measured from long-period surface waves. Therefore, instead of a single-value discrimination we recommend using more parameters (dimensions) for detecting tsunami earthquakes. As shown in [64] and [83], one could employ other T-phase characteristics such as its duration, seismic moment, and spectral strength or even similar features associated with the S-phase. We believe that the lack of success in predicting earthquakes still comes from the lack of communications between researchers and difficulties in free and fast access to the various types of data. Therefore, we hope that globalization of computation, data acquisition and visualization resources, together with fast access through a scale-free network, will provide a triumphant solution to this problem.

### Future Directions

In this chapter we endeavor to bring across the basic concept of clustering and its role in earthquake forecasting. Indeed we find that the clustering of seismic activities reflects both the similarity among them and their correlation properties. As discussed in, e.g., Ben-Zion et al. [9], Saichev and Sornette [68] and Zöller et al. ► [Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space](#), there exists an evolutionary process or memory between successive earthquakes, which impact the distribution of the inter-event times. We believe that by means of earthquake clustering we can capture the essence of this predictive information [27]. Therefore,

in order to carry out real-time earthquake forecasting for short-time scales, it is necessary to derive a thorough understanding of all families of earthquake clusters produced over an earthquake-prone region.

We stress here that in obtaining this type of information one must first be able to detect the precise location of the significant clusters, by filtering out simultaneously the noise and the outliers. While the existence of spatial-temporal clusters is important, they do not reveal the subtle information hidden behind the relations among the data events, such as: spatial-temporal correlation dimensions, correspondence between the numbers of small and large magnitude events, degree of spatial randomness, repetitiveness at different distances and other factors. The features – “descriptors” or seismicity parameters – constructed from the empirical knowledge of the researcher should be largely independent and should represent aptly distinctive features, which are useful for the purpose of pattern recognition. Unlike single events described only by spatio-temporal features (and magnitude), the  $N$ -dimensional feature vectors can represent better the dynamics of the seismically active area in different moments of time. By following the basic rules of learning theory, we may be able to arrive at the number  $N$  and quality of features, which can assure the generalization power of the data and allow us to construct reliable data-models or classifiers.

We have shown that clustering, as a well-honed tool in data mining and pattern recognition, represents the classifier without the teacher, which means that the nature of the clustering is unknown and its exact background must be guessed at from expert knowledge and analysis of the cluster properties. Clustering is a process based on a priori knowledge extraction for constructing the hypothesis space needed for reliable classifiers that can be taught and used for forecasting [25]. However, the quality of these data models depends strongly on the quality of hypothesis space constructed. Consequently, it depends on the quality of clusters extraction. The major problem comes from the lack of a universal clustering scheme, thus making the clustering process somewhat subjective. In this case we must visualize the multidimensional feature space. Visual confirmation gives one a confidence concerning the validity of the clusters and we can then adjust for the optimal clustering procedures by removing the noise and outliers. Among the major goals of earthquake clustering, we can include the following salient points:

- classification of the chaotic properties of seismicity patterns [35], for example to recognize the three main groups of shocks: foreshocks, mainshocks and after-

shocks or to remove the temporary clustering to estimate the background seismicity;

- understanding the correlations between observed properties of earthquakes in different domains (e. g., space, time, number, size);
- understanding the relations between various physical parameters of the models and properties of the generated earthquakes;
- investigating the multi-scale nature of the cluster structure and reconstructing the important and hidden information associated with the stress characteristics.

Classification of type of shocks seems to be an unresolved problem because there are no observable differences between foreshocks, main shocks and aftershocks [68]. Each earthquake is able of triggering other earthquakes according to the basic laws from [46,69]. Despite this difficulty, as shown in [9], it is possible to construct some sort of stochastic classifiers based on theoretical footing. The method proposed here closely related to the epidemic-type aftershock sequence (ETAS) model [61]. It is important that the principal characteristics of ETAS-based models correspond to experimental verifications, i. e., they treat all earthquakes on the same footing and there is not distinction between foreshocks, main shocks and aftershocks. The key points of the method are the probabilities of one event being triggered by a previous event (e. g., [82]). Making use of these probabilities, we can reconstruct the functions associated with the characteristics of earthquake clusters to test a number of plausible hypotheses about the earthquake clustering phenomena.

As shown above by our results on seismicity clustering for the three different time epochs, clustering can be truly regarded as a coarse-graining procedure. We can see details from the smaller scales are erased, thereby exposing the general trends associated with the long correlation length. For large data bases covering long time intervals we can unveil the shorter timescale characteristics by removing the background events, using successive clustering. Eventually, we can build up the strong classifiers. In the case where the long-time data catalogs are missing, we can employ the stochastic classifiers advocated Ben-Zion et al. [9] for prior thresholding of the background data or what is sometimes called “fuzzification” [86]. By this procedure we can construct the hypothesis space for data models by clustering (or fuzzy clustering) procedures.

The results discussed in this paper contribute to the development of improved software infrastructure for analysis of seismicity. A combined clustering analysis of observed and synthetic data, aided by state-of-the-art visualization of multidimensional clusters, undoubtedly lead to

improved earthquake forecasting algorithms with shorter time windows of increased probability of large seismic events.

### Acknowledgments

This research was supported by NSF ITR and Math-Geo grants. WD acknowledges support from the Polish Committee for Scientific Research (KBN) Grant No. 3T11C05926. YBZ acknowledges support from the NSF, USGS and SCEC.

### Bibliography

#### Primary Literature

1. Amira visualization package. <http://www.amiravis.com>
2. Andenberg MR (1973) Clusters analysis for applications. Academic Press, New York
3. Bak P, Tang C (1989) Earthquakes as a self-organized critical phenomena. *J Geophys Res* 94(B11):15635–15637
4. Bak P, Christensen K, Danon L, Scanlon T (2002) Unified scaling law for earthquakes. *Phys Rev Lett* 88:178501
5. Barabasi AL, Jeong H, Neda Z, Ravasz E, Schubert A, Vicsek T (2002) Evolution of the social network of scientific collaborations. *Physica A* 311(3–4):590–614
6. Bennett AF (1992) Inverse methods in physical oceanography. Cambridge University Press, Cambridge, pp 346
7. Ben-Zion Y (1996) Stress, slip and earthquakes in models of complex single-fault systems incorporating brittle and creep deformations. *J Geophys Res* 101:5677–5706
8. Ben-Zion Y (2001) Dynamic rupture in recent models of earthquake faults. *J Mech Phys Solids* 49:2209–2244
9. Ben-Zion Y (2003) Appendix 2, key formulas in earthquake seismology. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International handbook of earthquake and engineering seismology*, Part B. Academic Press, pp 1857–1875
10. Ben-Zion Y, Rice JR (1993) Earthquake failure sequences along a cellular fault zone in a three-dimensional elastic solid containing asperity and nonasperity regions. *J Geophys Res* 98:14109–14131
11. Ben-Zion Y, Rice JR (1995) Slip patterns and earthquake populations along different classes of faults in elastic solids. *J Geophys Res* 100:12959–12983
12. Ben-Zion Y, Eneva M, Liu Y (2003) Large earthquake cycles and intermittent criticality on heterogeneous faults due to evolving stress and seismicity. *J Geophys Res* 108(B6):2307–27
13. Berman F, Fox GC, Hey AJG (2003) Grid computing – making the global infrastructure a reality. *Wiley Series in Communications Networking and Distributed Systems*, pp 1007
14. Boatwright J, Choy GL (1986) Teleseismic estimates of the energy radiated by shallow earthquakes. *J Geophys Res* 91:2095–2112
15. Bollig EF, Jensen PA, Lyness MD, Nacar MA, da Silveira PR, Erlebacher G, Pierce M, Yuen DA (2007) VLAB: Web services, portlets, and workflows for enabling cyber infrastructure in computational mineral physics. *J Phys Earth Planet Inter* 163:333–346

16. Chen C-C, Rundle JB, Li H-C, Holliday JR, Turcotte DL, Tiampo KF (2006) Critical point theory of earthquakes: Observations of correlated and cooperative behavior on earthquake fault systems. *Geophys Res Lett* L18302
17. Chinnery M (1963) The stress changes that accompany strike-slip faulting. *Bull Seismol Soc Am* 53:921–932
18. Corral A (2005) Mixing of rescaled data and Bayesian inference for earthquake recurrence times. *Nonlinear Process Geophys* 12:89–100
19. Corral A (2005) Renormalization-group transformations and correlations of seismicity. *Phys Rev Lett* 95:028501
20. Corral A, Christensen K (2006) Comment on earthquakes descaled: On waiting time distributions and scaling laws. *Phys Rev Lett* 96:109801
21. da Silva CRS, da Silveira PRC, Karki B, Wentzcovitch RM, Jensen PA, Bollig EF, Pierce M, Erlebacher G, Yuen DA (2007) Virtual laboratory for planetary materials: System service architecture overview. *Phys Earth Planet Inter* 163:323–332
22. Davy P, Sornette A, Sornette D (1990) Some consequences of a proposed fractal nature of continental faulting. *Nature* 348:56–58
23. Dzwiniel W, Blasiak J (1999) Method of particles in visual clustering of multi-dimensional and large data sets. *Future Gener Comput Syst* 15:365–379
24. Dzwiniel W, Yuen DA, Kaneko Y, Boryczko K, Ben-Zion Y (2003) Multi-resolution clustering analysis and 3-D visualization of multitudinous synthetic earthquakes. *Vis Geosci* 8:12–25
25. Dzwiniel W, Yuen DA, Boryczko K, Ben-Zion Y, Yoshioka S, Ito T (2005) Nonlinear multidimensional scaling and visualization of earthquake clusters over space, time and feature space. *Nonlinear Process Geophys* 12:117–128
26. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 95:14863–14868
27. Enescu B, Ito K, Struzik ZR (2006) Wavelet-based multiscale analysis of real and simulated time-series of earthquakes. *Geophys J Int* 164:63–74
28. Eneva M, Ben-Zion Y (1997) Techniques and parameters to analyze seismicity patterns associated with large earthquakes. *J Geophys Res* 102(B8):785–795
29. Eneva M, Ben-Zion Y (1997) Application of pattern recognition techniques to earthquake catalogs generated by models of segmented fault systems in three-dimensional elastic solids. *J Geophys Res* 102:24513–24528
30. Ferreira L (2002) Introduction to grid computing with Globus IBM Redbook series. IBM Corporation <http://ibm.com/redbooks>
31. Foster I, Kesselman C (eds) (1998) Building a computational grid: state-of-the art and future directions in high-performance distributed computing. Morgan-Kaufmann, San Francisco
32. Freed AM, Lin J (2001) Delayed triggering of the 1999 Hector Mine earthquake by viscoelastic stress transfer. *Nature* 411:180–183
33. Frey BJ, Dueck D (2007) Clustering by Passing Messages Between Data Points, *Science* 315(5814):972–976
34. Garbow ZA, Erlebacher G, Yuen DA, Sevre EO, Nagle AR, Kaneko Y (2002) Web-based interrogation of large-scale geophysical datasets and clustering analysis of many earthquake events from desktop and handheld devices. American Geophysical Union Fall Meeting, Abstract
35. Goltz C (1997) Fractal and chaotic properties of earthquakes. In: Goltz C (ed) *Lecture notes in earth sciences*, vol. 77. Springer, Berlin, p 3–164
36. Gowda CK, Krishna G (1978) Agglomerative clustering using the concept of nearest neighborhood. *Pattern Recognit* 10:105
37. Grossman RL, Karnath Ch, Kegelmeyer P, Kumar V, Namburu RR (2001) Data mining for scientific and engineering applications. Kluwer, Dordrecht
38. Guha S, Rastogi R, Shim K (1998) CURE: An efficient algorithm for large databases. In: *Proceedings of SIGMOD '98*, Seattle, June 1998. pp 73–84
39. Gutenberg B (1942) Earthquake magnitude, intensity, energy and acceleration. *Bull Seismol Soc Am* 32:163–191
40. Gutenberg B, Richter CF (1954) Seismicity of the earth and associated phenomena. Princeton University Press, Princeton
41. Haile PM (1992) *Molecular Dynamics Simulation*. Wiley, New York
42. Hainzl S, Scherbaum F, Beauval C (2006) Estimating background activity based on interevent-time distribution. *Bull Seismol Soc Am* 96:313–320. doi:10.1785/0120050053
43. Hand D, Mannila H, Smyth P (2001) *Principles of data mining*. MIT Press, Cambridge
44. Hastie T, Tibshirani R, Friedman J (2001) *The elements of statistical learning: Data mining, inference and prediction*. Springer, New York, pp 533
45. Helmstetter A, Sornette D, Grasso J-R (2003) Mainshocks are aftershocks of conditional foreshocks: How do foreshock statistical properties emerge from aftershock laws. *J Geophys Res* 108:2046
46. Helmstetter A, Kagan Y, Jackson D (2005) Importance of small earthquakes for stress transfers and earthquake triggering. *J Geophys Res* 110:B05508
47. Hong H, Kadlec DJ, Yuen DA, Zheng Y, Zhang H, Liu G, Dzwiniel W (2004) Fast timescale phenomena at Changbaisan volcano as inferred from recent seismic activity. *Eos Trans AGU Fall Meet.* 85(47) <http://www.agu.org>
48. Ismail MA, Kamel MS (1989) Multi-dimensional data clustering utilizing hybrid search strategies. *Pattern Recognit* 22(1):77–89
49. Ito T, Yoshioka S (2002) A dike intrusion model in and around Miyakejima, Niijima and Kozushima. *Tectonophysics* 359:171–187
50. Jajuga K, Sokolowski A, Hermann H (eds) (2002) *Classification, clustering and data analysis*. Springer, Berlin, pp 497
51. Jones NC, Pevzner P (2004) *An introduction to bioinformatics algorithms*. MIT Press, Cambridge
52. Kadlec BJ, Yang XL, Wang Y, Bollig EF, Garbow ZA, Yuen DA, Erlebacher G (2003) WEB-IS (Integrated System): An overall view. *Eos Trans AGU* 84(46), Fall Meet. Suppl., Abstract NG11A-0163
53. Kalnay E (2003) *Atmospheric modeling, data assimilation and predictability*. Cambridge University Press, Cambridge, pp 341
54. Karypis G, Han E, Kumar V (1999) Chameleon: A hierarchical clustering algorithms using dynamic modeling. *IEEE Computer* 32(8):68–75
55. Karypis G, Aggarwal R, Kumar V, Shekhar S (1999) Multi-level hypergraph partitioning: applications in VLSI domain. *IEEE Trans on Very Large Scale Systems Integration (VLSI)* 7(1):69–79

56. Mehta AP, Dahmen KA, Ben-Zion Y (2006) Universal mean moment rate profiles of earthquake ruptures. *Phys Rev E* 73:056104
57. Mitra S, Acharya T (2003) *Data mining: multimedia, soft computing and bioinformatics*. Wiley, New Jersey
58. Molchan GM (2005) Interevent time distribution of seismicity: A theoretical approach. *Pure Appl Geophys* 162:1135–1150
59. National Research Council (2003) *Living on an active earth, perspectives on earthquake sciences*. The National Academies Press, Washington DC
60. Newman AV, Okal EA (1998) Teleseismic estimates of radiated seismic energy: the  $S/M0$  discriminant for tsunami earthquakes. *J Geophys Res* 103(B11):23885–23898
61. Ogata Y (1999) Seismicity analysis through point-process modeling: A review. *Pure Appl Geophys* 155:471–507
62. Ogata Y, Zhuang J (2006) Space-time ETAS models and an improved extension. *Tectonophysics* 413:13–23
63. Okada Y (1992) Internal deformation due to shear and tensile faults in a half space. *Bull Seismol Soc Am* 82:1018–1040
64. Okal EA, Alasset P-J, Hyvernaud O, Schindele F (2003) The deficient T waves of tsunami earthquakes. *Geophys J Int* 152:416–432
65. Rundle JB, Gross S, Klein W, Ferguson C, Turcotte DL (1997) The statistical mechanics of earthquakes. *Tectonophysics* 277:147–164
66. Rundle JB, Klein W, Tiampo K, Gross S (2000) Linear pattern dynamics in nonlinear threshold systems. *Phys Rev E* 61(3):2418–2143
67. Rundle JB, Turcotte DL, Klein W (eds) (2000) *GeoComplexity and the physics of earthquakes*. American Geophysical Union, Washington, pp 284
68. Saichev A, Sornette D (2007) Theory of earthquake recurrence times. *J Geophys Res* 112(B4):1–26
69. Saichev A, Helmstetter A, Sornette D (2005) Power law distributions of offspring and generation numbers in branching models of earthquake triggering. *Pure Appl Geophys* 162:1113–1134
70. Sander J, Ester M, Krieger H (1998) Density based clustering in spatial databases, The algorithm DBSCAN and its applications. *Data Min Knowl Discov* 2(2):169–194
71. Shcherbakov R, Turcotte DL (2004) A modified form of Bath's law. *Bull Seismol Soc Am* 94:1968–1975
72. Shcherbakov R, Turcotte DL, Rundle JB (2005) Aftershock statistics. *Pure Appl Geophys* 162(6–7):1051–1076
73. Siedlecki W, Siedlecka K, Sklanski J (1988) An overview of mapping for exploratory pattern analysis. *Pattern Recognit* 21(5):411–430
74. Sornette D (2006) *Critical phenomena in natural sciences*. Springer Series in Synergetics, Berlin, pp 528
75. Sornette D, Johansen A, Bauchaud J-P (1996) Stock market crashes, precursors and replicas. *J Phys Int Finance* 5:167–175
76. Stein S, Okal E (2004) Ultra-long period seismic moment of the great December 26, 2004 Sumatra earthquake and implications for the slip process. <http://www.earth.nwu.edu/people/emile/research/Sumatra.pdf>
77. Tan P-N, Steinbach M, Kumar V (2005) *Introduction to data mining*. Addison Wesley, Boston, pp 769
78. Theodoris S, Koutroumbas K (1998) *Pattern recognition*. Academic Press, San Diego
79. Turcotte DL (1997) *Fractals and chaos in geology and geophysics*, 2nd Edn. Cambridge University Press, New York
80. Utsu T (2002) Statistical Features of Seismicity. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology*, Part A. Academic Press, pp 719–732
81. Van Aalsburg J, Grant LB, Yakovlev G, Rundle PB, Rundle JB, Turcotte DL, Donnellan A (2007) A feasibility study of data assimilation in numerical simulations of earthquake fault systems. *Phys Earth Planet Inter* 163:149–162
82. Vere-Jones D (1976) A branching model for crack propagation. *Pure Appl Geophys* 114(4):711–726
83. Walker DA, Mc Creery CS, Hiyoshi Y (1992) T-phase spectra, seismic moment and tsunamigenesis. *Bull Seismol Soc Am* 82:1275–1305
84. Wesnousky SG (1994) The Gutenberg-Richter or characteristic earthquake distribution, which is it? *Bull Seismol Soc Amer* 84:1940–1959
85. Yuen DA, Garbow ZA, Erlebacher G (2004) Remote data analysis, Visualization and Problem Solving Environment (PSE) based on wavelet analysis in the geosciences. *Vis Geosci* 8:83–92. doi:10.1007/x10069-003-0012-z
86. Zadeh LA (1996) Fuzzy sets, fuzzy logic, and fuzzy systems: selected papers by Lotfi A Zadeh, World Scientific Series In Advances In Fuzzy Systems. World Scientific Publishing, River Edge, pp 826
87. Zhang Q, Boyle R (1991) A new clustering algorithm with multiple runs of iterative procedures. *Pattern Recognit* 24(9):835–848
88. Zöller G, Hainzl S, Ben-Zion Y, Holschneider M (2006) Earthquake activity related to seismic cycles in a model for a heterogeneous strike-slip fault. *Tectonophysics* 423:137–145. doi:10.1016/j.tecto.2006.03.007
89. Zöller G, Ben-Zion Y, Holschneider M (2007) Estimating recurrence times and seismic hazard of large earthquakes on an individual fault. *Geophys J Int* 170:1300–1310. doi:10.1111/j.1365-246X.2007.03480.x

## Books and Reviews

- Ertoz L, Steinbach M, Kumar V (2003) Finding clusters of different size, shapes and densities in noisy, high-dimensional data. Army High Performance Center, technical report, April 2003
- Yuen DA, Kadlec BJ, Bollig EF, Dzwiniel W, Garbow ZA, da Silva C (2005) Clustering and visualization of earthquake data in a grid environment, vol 10/1. *Vis Geosci* <http://www.springerlink.com/content/n60423820556/>

## Earthquake Damage: Detection and Early Warning in Man-Made Structures

MARIA I. TODOROVSKA

Department of Civil Engineering,  
University of Southern California, Los Angeles, USA

### Article Outline

Glossary

Definition of the Subject

Introduction

Literature Review

Damage and Damage-Sensitive Features

Structural Models and Identification

Examples

Future Directions

Bibliography

### Glossary

**Structural health monitoring** Is the process of determining and tracking the structural integrity and assessing the nature of damage in a structure. It is often used interchangeably with *structural damage detection*.

**Inter-story drift** Is the ratio between the relative horizontal displacements at two levels of the structure and the distance between them. It is important to distinguish between drift resulting from *deformation* of the structure, which is directly related to damage, and drift resulting from the deformation of the soil and rocking of the structure as a rigid body. It is also important to estimate reliably the drift due to permanent displacement (its “DC” component), which cannot be done reliably using data from vibrational sensors unless six degrees of freedom of motion (three translations and three rotations) are recorded.

**Soil-structure interaction (SSI)** Is a process occurring during vibration of structures founded on flexible soil, in which the structure and soil interact, and their motions are modified. *Kinematic interaction* refers to the effects of scattering and diffraction of the incident seismic waves from the soil excavation for the foundation. *Dynamic interaction* refers to the effects caused by the inertia forces of the structure and foundation, which lead to deformation of the soil, and results in modification of the *resonant frequencies* and *damping* of the response of the structure, foundation and soil acting as a system.

**Resonant frequencies of vibration** Of a structure on flexible soil are those of the *soil-structure system*, and the energy of the vibrational response is concentrated around these frequencies. They depend on the stiffness of the building and that of the soil. *Fixed-base frequencies of vibration* are the resonant frequencies of the structure on rigid soil, and depend only on the stiffness of the structure. Loss of stiffness of the structure due to damage results in reduction of the fixed-base frequencies, and indirectly of the system frequencies. Monitoring changes in the fixed-base frequencies is most reliable because it eliminates the effects of the soil, which can exhibit (recoverable) nonlinear behavior during strong shaking.

### Definition of the Subject

*Structural health monitoring* and structural damage detection refers to the process of determining and tracking the structural integrity and assessing the nature of damage in a structure. Being able to detect the principal components of damage in structures as they occur during an earthquake or soon after the earthquake, or the absence of it, before physical inspection is possible, is an important and challenging problem. Considering the challenges faced and the potential benefits for safety and for minimizing disruption of productivity, structural health monitoring has the elements of a *grand challenge* problem in civil engineering [12].

*Structural damage* can be described by the following five attributes: existence, location, type, extent, and prognosis for the remaining useful life. Structural damage is a complex state, which can occur on different time scales, suddenly during some catastrophic event such as earthquake or explosion, or gradually over the life of the structure, due to deterioration of the structural materials by aging, service, and exposure to environmental influences. This article is concerned primarily with identification of the most significant components in the space of complex patterns of damage caused by earthquakes. Damage in structures also can be described on different spatial scales, e. g. from small defects and localized damage in a component, to global state of damage of the structural system. Hence the damage detection methods are classified as *local* and *global*. The local methods are those for nondestructive testing (NDT) of materials, which can determine the location of the damage in a structural component. They involve use of actuators (radiating ultrasonic waves into the structural element), and require access to the element. The global methods assess the overall state of damage of the structural system (as it reflects on its overall perfor-

mance during an extreme event). The focus of this review is on the global methods, and intermediate scale methods, which can point to the part of the structure that has been damaged.

Structural damage detection and early warning involve: (1) *recording* some sensory data, (2) *identification* of some structural parameter(s) sensitive to damage (e. g. natural frequencies of vibration, or wave travel times), some characteristic of response (e. g. levels of inter-story drift) that can be correlated with damage, or some other patterns (e. g. abrupt changes in the response detected as novelties), (3) *comparison* of the result of the identification with some knowledge base of correlation of such patterns with levels of damage, and (4) *decision making* (e. g. whether to evacuate or continue occupancy). Because of various uncertainties, the answer can be only expressed probabilistically, and the decision will also depend on the nature of the use of the structure and level of tolerance of the user.

The earliest and most wide-spread methods of structural damage detection are those based on data from vibrational sensors. In fact, the hope to eventually be able to detect hidden damage has been one of the motivations for the development and deployment of seismic sensors in structures. The first strong motion recordings in a building are those during the  $M = 5.4$  Southern California earthquake of October 2, 1933, obtained in the Hollywood Storage Building, the instrumented structure in US with the longest history of recording earthquakes [63]. The earliest identification methods consisted of estimation of the building resonant frequencies and damping, from energy distributions of small amplitude ambient noise and forced vibration tests [3], as well as from earthquake records [63]. These studies identified the resonant frequencies and damping of the soil-structure system, which depend on the properties of the soil, and can change significantly even when there is no damage. Detailed system identification studies from full-scale test vibration data that separate the effects of the soil-structure interaction appeared in the 1970s, following theoretical developments that helped understanding the phenomenon of soil-structure interaction [13,32,33,34,71]. Thirty years later, such studies are still rare, due to a combination of factors, one of which is the inadequate coverage of this topic in the graduate curricula, and the other is the emphasis of earthquake engineering research on laboratory experimentation and numerical simulations, rather than on the full-scale testing of structures [63].

Despite the progress made to date in instrumentation of structures as well as in development of theoretical methods, structural health monitoring systems are deployed in

structures only on an experimental basis. The main obstacles to the routine practical deployment of such systems are: (1) the high cost of sensors and monitoring systems, which limits the number of structures that are instrumented and the detail of the measurements (spatial resolution, e. g.), (2) the low sensitivity and robustness of the methods, and ability to discriminate between changes in the damage sensitive feature caused by damage from changes caused by other factors (e. g. age, level of excitation, and weather), and (3) the paucity of data recorded in damaged structures necessary to calibrate the health monitoring methods. Consequently, the main challenges for future research are: (1) to design low cost but high performance sensors and monitoring systems, making it possible to densely instrument many structures, (2) to develop methods that are robust and sensitive enough to detect also light damage (in particular one that is not visible), and (3) to build a knowledge base that can help reliably relate observed patterns in the data with actual observations of damage.

Recently, structural identification and health monitoring of buildings by detecting changes in wave travel time through the structure has received revived attention and has proven to be very promising [20,24,25,37,39,41,42,53,54,65]. Exploratory applications to data from damaged buildings [53,54] showed that the method (1) is robust when applied to damaging levels of earthquake response data, (2) is not sensitive to the effects of soil-structure interaction, and (3) is local in nature (i. e. gives results consistent with the spatial distribution and degree of the observed damage).

## Introduction

This volume would not be complete without addressing the catastrophic consequences of earthquakes, and damage in soil-structure systems, which is a complex, multidimensional, and highly interrelated set of phenomena.

Since the early days, the mathematical formulation of practical earthquake engineering problems has been dominated by *linear* differential equations [58], which *cannot* lead to chaos. Nevertheless, cost and the increasing needs of society have pushed the design into the nonlinear regimes of large deformations increasing the possibility of encountering chaotic dynamic phenomena in structural response, and have increased the complexity of the possible damage outcomes. However, working with parameters that produce chaotic output reduces the ability to predict the outcome. The chaotic behavior of nonlinear systems does not completely exclude the possibility to predict the response, but introduces an upper bound (prediction

horizons) [30]. Then the remaining question is over what time-scales can the predictions still be reliable. Also, the prediction of response requires a realistic physical *model*, while the practical outcome of most work in engineering remains *empirical*. Consequently, there is a conflict in the classical engineering description of the world. This conflict is in part due to the assumption that nature is moving forward, according to a deterministic law, and in part due to the fact that engineers model the world based on incomplete data, and thus working with unverifiable representation. This leads to the question what models are good for. The problem is further aggravated by the fact that the art of dynamical modeling tends to be neglected in discussions of nonlinear and chaotic systems, in spite of its crucial importance [2]. In the following review of structural health monitoring in earthquake engineering, it is accepted that there is a modeling problem, and the success of a method is gauged by the degree to which its predictions match the observed outcomes.

During the last several decades, stochastic processes have been used to help analyze the irregular behavior of deterministic systems with too many variables to be described in detail. Stochastic processes have been used also to model the deterministic response of structures to earthquake and wind forces, and as an approximate description of deterministic systems sensitive to their initial conditions. In some analyses, random noise is added to the model to account for the differences between the behaviors of model and prototype. This noise represents no more than lack of knowledge of the system structure or inadequacy of the identification procedure [23].

Following a damaging earthquake, buildings, bridges, dams and other structures are physically inspected for damage, and their safety is assessed. To assess the safety of buildings, the city departments of public safety (or their equivalents) dispatch inspectors to the field to “walk through” each building and write a report on the observed damage and safety concerns to its occupants. On the basis of such assessments, a color tag can be assigned to the building: (1) green if the structure is safe, (2) yellow if it has been damaged and needs to be evacuated, but is safe for the occupants to return to retrieve their belongings, and (3) red if it has been damaged to a degree that it is unsafe for the occupants to return to the structure [1]. When the affected area is relatively large, such inspection takes time (several weeks or longer), and the tagging is often first preliminary, to be revised at a later time after a preliminary inspection of all buildings has been completed. Such walk-in inspections can detect only damage that is visible, and there is always considerable subjectivity in the assessments. The major problem with such inspections is

however the timeliness, as aftershocks following the earthquakes can further damage a structure that has survived the main event but is weakened, and endanger the occupants. Another problem is the loss of function of a structure that may be safe, until a more detailed inspection and assessment is possible. This is particularly important for critical facilities, such as hospitals, as well as for major businesses, such as banks, for which interruption of work can cause major financial losses. Without a doubt, the ability to detect damage in structures early, as it occurs or soon after the earthquake, using some structural health monitoring system, and assess the state of safety of the structure before physical inspection is possible, can benefit society immensely. Ideally, based on instrumental data, such systems would be able to detect also hidden damage that is not visible to the naked eye. There would be benefit even when the damage is obvious, if that information is available immediately after the earthquake. To be effective, however, such systems must be sensitive enough to detect at least the significant damage, and also be accurate enough, to avoid false alarms and unnecessary and costly service interruption.

The objective of this article is to review the basic principles on which such systems operate, and to present some illustrative examples of several robust methods applied to full-scale buildings. This is followed by a discussion of remaining critical issues and directions for future research, in the view of the author.

## Literature Review

### Earthquake Damage Detection in Structural Health Monitoring Research

Earthquake damage detection in civil structures, such as buildings and bridges, is closely related to structural health monitoring of structures such as light aerospace structures, rotating machinery and offshore platforms, for example, that are of concern to other disciplines. A review of recent developments in this broader field, as applied to civil and mechanical systems, can be found in Chang et al. [7] and Liu et al. [31]. The earliest, and still the most popular methods for civil structures are those that use data from vibrational sensors, and detect changes in the vibrational characteristics of the structure – frequencies of vibration and mode shapes. Detailed reviews of vibrational methods in the general area of structural health monitoring can be found in a report by Doebling et al. [11], its shorter version as a journal paper [10], and a follow up report by Sohn et al. [40]. Another recent review of the vibrational methods can be found in Carden and Fanning [4].

These detailed reviews conclude that the currently available vibrational methods can determine if the structure has been damaged, but cannot indicate precisely the location of the damage, and are therefore referred to as *global*. Most vibrational methods monitor changes in the *modal* properties of the structures (modal frequencies and mode shapes). The stated difficulties associated with these methods include: (1) the presence of other factors than damage that produce similar effects on the monitored parameters not easy to isolate (e. g. the effects of soil-structure interaction on the measured frequencies of vibration, as well as environmental influences such as temperature and rain; [8,46,47]); (2) the redundancy of the civil engineering structures, which results in low sensitivity of the method (i. e. small change of the overall stiffness and consequently of the measured frequencies) when the damage is localized; and (3) dependence on detailed prior analytical models and/or prior test data for the detection and location of damage (supervised learning), which may not be readily available for a structure, may be outdated, and even when available represent only an idealization of the real structure [7,11]. Further critical issues identified are (4) the scarcity of objective comparisons of different procedures applied to a common data set, and (5) the number and location of sensors (techniques to be seriously considered for implementation in the field should demonstrate that they can perform well for small numbers of measurements). Finally, Doebling et al. [11] conclude that “while sufficient evidence exists to promote the use of measured vibration data for the detection of damage in structures, using both forced-response testing and long-term monitoring of ambient signals, the research needs to be more focused on the specific applications and industries that would benefit from this technology... Additionally, research should be focused more on testing of real structures in their operating environment, rather than laboratory tests of representative structures.”

In the follow up review, Sohn et al. [40] mention as outstanding problems: The reliance on analytical models to obtain the structural parameters from the data, not only in methods involving direct inversion, but also in those that use neural networks; and that the damage sensitive features are also sensitive to changes of the environmental and operational conditions of the structures. They mention as one of the most significant improvements since the previous review [11] the signal processing methods that do not rely on detailed analytic models, such as novelty/outlier analysis, statistical process control charts, and simple hypothesis testing (unsupervised learning), shown to be very effective to identify the onset of damage growth, and the presence of damage but not the damage type. In

this article, one such method – based on detection of novelties using wavelets – is reviewed and illustrated. Another significant advancement is the availability of more affordable MEMS sensors, as well as fiber optics, and piezoceramic sensors, and of wireless data communication technology.

In structural health monitoring literature, the vibrational methods are referred to as *global*, due to the relatively small number of sensors typically installed in structures, and can detect only significant damage [11,40]. The cost of seismic monitoring systems is still high, and trade-offs have to be made between the detail of the instrumentation of a particular structure and the number of structures that are instrumented. The truly *local* methods are those for nondestructive testing (NDT) of materials, which can detect the location of cracks or some other defects in a structural member. These methods typically use: (1) ultrasonic waves, which are attenuated quickly along the wave path, (2) need an actuator to create such waves, and (3) require direct access to the structural member, usually not readily available. Consequently, they are used to detect the location of the damage in a particular structural member, known or suspected to have been damaged, but are too costly and impractical for structural health monitoring of an entire structure [7]. To make a difference for society, structural health monitoring and early warning systems have to be reasonably priced so that they can be installed in many structures.

### Earthquake Damage Detection in Earthquake Engineering Research

In the earthquake engineering research, earthquake damage detection emerges from system identification studies of full-scale structures (typically involving identification of their frequencies of vibration and damping) from ambient and forced vibration test data, or earthquake records. Consequently, it is *data driven*, in contrast to the structural health monitoring research, which focuses on methodologies, validated mostly on “clean” numerically simulated data, and sometimes on laboratory data or small amplitude full-scale data. In the US, the earliest system identification studies from full-scale data follow the first deployment of strong motion instruments in structures [3], and continue through the 1960s [9,19,70]. More sophisticated studies from the system identification point-of-view using earthquake response data appear in the 1970s, following the San Fernando, California earthquake of 1971, which produced strong motion records in many buildings in the Los Angeles metropolitan area [66,67,68,69]. A significant finding of these studies is that the building frequencies of



actual structures vary significantly as a function of the level of the response. The variation is such that the fundamental frequency decreases during the largest shaking, but recovers afterwards during the remaining smaller amplitude shaking, or during subsequent shaking from aftershocks or small amplitude tests. The recovery may be partial or complete, and a large reduction of frequency of vibration during the earthquake is not always associated with visible damage. This is an important fact, as the decrease of the fundamental frequency of vibration is used as one of the global indicators of damage in structural health monitoring research, and also because many sophisticated structural identification methods are based on the assumption of stationarity and time invariance of the response.

Further, system identification studies of structures using earthquake records, considering the effects of the interaction of the structural vibrations with the vibration of the surrounding soil, appear in the 1960 and 1970s. The most detailed such full-scale studies are probably those of the Millikan library in Pasadena [33,34,71]. Understanding and consideration of the effects of soil-structure interaction in system identification and health monitoring of structures is of *fundamental importance* for the development of reliable methodologies, as this phenomenon is an integral part of the seismic response, and affects the estimation of both the frequencies of vibration and the inter-story drift, both used to infer about the state of damage. Nevertheless, these effects are typically ignored in structural health monitoring research. A detailed literature review on full-scale studies of soil-structure interaction can be found in Trifunac et al. [63], and a discussion of critical issues in recording and interpreting earthquake response of full-scale structures can be found in Trifunac and Todorovska [60,61].

### Damage and Damage-Sensitive Features

The damage of a structure can be described by the following five states: (1) no damage, (2) repairable (light and moderate) damage, (3) irreparable damage, (4) extreme damage, and (5) collapse [14].

Damage is associated with large deformations of the structural elements (usually expressed via the inter-story drift), which cause yielding of the structural steel or steel reinforcement and cracking of the structural concrete. Also, damage causes changes of the structural vibrational characteristics (frequencies of vibration), and wave propagation characteristics (wave velocities/travel times). This section presents the rationale for damage detection algorithms based on monitoring such changes. The concepts are illustrated on a simple soil-structure interaction model.

## Structural Models and Identification

### Structure as an Oscillator

From an elementary vibrational viewpoint, a structure responds to earthquake shaking as an oscillator characterized by its frequencies of vibration. The fixed-base frequencies are those of free vibration of the structure on *rigid* ground. They are the eigenvalues of a boundary value problem, and the associated eigenfunctions are referred to as mode shapes in structural engineering. The fixed-base frequencies depend only on the properties of the structure, i. e. on the structural stiffness and mass, while their dependence on the structural damping is small for most structures, which are lightly damped. In the linear range, the response of an n-degree-of-freedom system to earthquake shaking is a superposition of the modal responses. The contribution of the fundamental mode is usually the largest, and in engineering design structures are often represented by an equivalent single degree-of-freedom oscillator. For a single degree-of-freedom oscillator, the natural frequency of vibration is

$$\omega_1 = \sqrt{k/m}, \quad (1)$$

where  $k$  is its stiffness and  $m$  its mass. The frequency of such an oscillator is affected little by typical fluctuations of the mass due to variations in the life load of the structure, and is mostly affected by changes in the stiffness. Damage would cause loss of stiffness, and consequently reduction of the fixed-base frequency of vibration. If  $\omega_{1,\text{ref}}$  is a reference frequency corresponding to reference stiffness  $k_{\text{ref}}$ , then for the damaged structure

$$(\omega_1/\omega_{1,\text{ref}})^2 = k/k_{\text{ref}}. \quad (2)$$

As the fixed-base frequency depends on the *overall* stiffness of the structure, it is by definition a *global* property, and would not change much due to localized damage of civil structures, which are designed to be highly redundant. One advantage of detecting damage by monitoring changes in fixed-base frequency of vibration is that, in the ideal case when the ground is practically rigid (as compared to the structure), and the excitation is relatively broadband, the fixed-base frequency can be determined using only one sensor, on the roof, as the frequency of the peak of the Fourier transform of the roof response. The availability of recorded response at ground level would produce a more accurate estimate, as a transfer-function can be computed between the roof and ground level response motion. Changes in the frequency versus time can be estimated from the Fourier transform in moving windows in time.

Buildings are founded on soil, which is flexible and deforms under the action of forces from the incident waves and from the vibrating structure. Even if rigid, a structure founded on soft soil will vibrate, with the soil acting as a spring. The soil adds both *flexibility* and *dissipation mechanism* to the vibrations of the structure and soil, which act as a coupled system. Two sources of dissipation are (1) scattering of the incident waves from the foundation and (2) radiation of energy into the soil (through vibration of the foundation, which acts as a source of waves radiated in semi-infinite medium). The third source of dissipation is in the structure, and includes a distribution of frictional sources, and hysteretic damping during nonlinear response. The soil-structure system has its own resonant frequencies and “damping”, which is a combination of the contributions from the structure and from the soil. The fundamental frequency of the system is always lower than the fundamental fixed-base frequency of the structure, but the associated system damping can be larger or smaller than the damping of the structure alone, depending on the radiation damping and relative stiffness of the structure with respect to the soil.

In conclusion, the difficulties with Fourier-type analyses for identification of the building frequencies are that these give the resonant frequencies and equivalent damping of the *system*, which depend on the soil, and that they are global properties. Also, there is no knowledge base of changes in such frequencies (for different types of structures and different types of soils) related to different degrees of damage.

### Structure as a Wave Guide

Alternatively, the seismic response can be represented as a superposition of waves that propagate through the structure, reflect from its exterior and interior boundaries and interfere [22,39,41,48,49,50,56,57]. Loss of stiffness due to local damage would cause delays in the wave propagation through the damaged part, which could be detected using seismic response data recorded on each side of the damaged area, along the wave path. A change in wave travel time would depend *only on the changes of the physical properties between the sensors*. Hence, the wave methods are more sensitive to local damage than the modal methods, and should be able to point out the location of damage with a relatively small number of sensors. Additionally, the local changes in travel time are not sensitive to the effects of soil-structure interaction (as demonstrated in [44,45]), which is a major obstacle for the modal methods based on detecting changes in the structural frequencies.

The basic principles of the method are as follows. It is based on D’Alembert’s solution of the wave equation, and representation of the structural response as a superposition of waves traveling through the structure. In contrast, the modal methods are based on representation in the Fourier domain, as superposition of modes of vibration.

The wave travel time between two points

$$\tau = d/V_s, \quad (3)$$

where  $d$  is the distance traveled and  $V_s$  is the equivalent shear wave velocity in the part of the building between the two sensors. The latter is related to the rigidity via

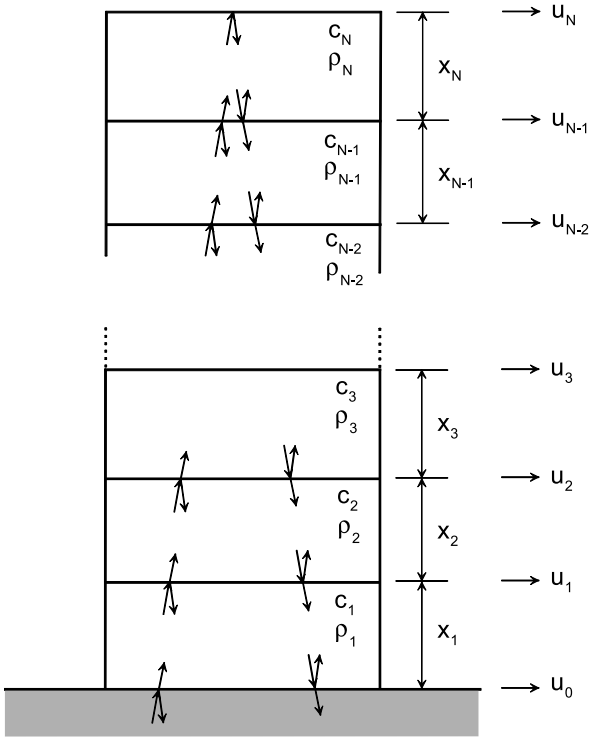
$$V_s = \sqrt{\mu/\rho}, \quad (4)$$

where  $\mu$  is the shear modulus and  $\rho$  is the density. Hence, reduction of rigidity due to damage will produce a reduction of the equivalent shear wave velocity, which will produce an increase in the pulse travel time, relative to the travel time for the undamaged state. Let  $\mu_{\text{ref}}$  be reference rigidity, and  $V_{s,\text{ref}}$  and  $\tau_{\text{ref}}$  be the corresponding shear wave velocity and wave travel time. Then their changes are related as follows

$$\frac{\tau}{\tau_{\text{ref}}} = \frac{1}{V_s/V_{s,\text{ref}}} = \frac{1}{\sqrt{\mu/\mu_{\text{ref}}}}. \quad (5)$$

Global changes can also be detected by monitoring the *total* wave travel time from the base to the roof of a building. Let  $\tau_{\text{tot}}$  be the travel time of seismic waves from the point of fixity (ground level) to the roof. Then the building fundamental fixed-base frequency  $f_1 = 1/(4\tau_{\text{tot}})$  assuming that the building as a whole deforms like a shear beam. Based on this relation,  $f_1$  can be estimated using data from only two horizontal sensors. While the goodness of this approximation of  $f_1$  may vary from one building to another, the changes in  $f_1 = 1/(4\tau_{\text{tot}})$  will still depend *only* on changes in the building itself, and not on changes in the soil, and monitoring of changes in such an estimate of  $f_1$  can be used as a global indicator of damage in a building [44,45].

Figure 1 shows a conceptual model for the analysis, in which the building is a horizontally layered medium, with the interfaces between layers at the floor slabs. For vertically incident waves, or/and a narrow building, the layered medium will be traversed by waves propagating upward and downward. Let the excitation be a pulse. At each interface, an incident pulse will be split into a reflected pulse, and a transmitted pulse, and at the roof total reflection will occur. The transmission and reflection coefficients will depend on the impedance contrast between the layers, in particular on the shear wave velocities, which will change due



Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 1  
Layered building model

to loss of stiffness caused by damage. Because of reflections and material damping, an incident wave pulse will attenuate as it propagates through the structure, and will be also modified in a dispersive medium. This is schematically illustrated in Fig. 1. The total wave motion propagating upward in a layer will be a superposition of all the pulses, those from direct incidence and those from different generations of reflections. The same applies for the pulses propagating downward. The downward propagating pulses that are reflected back into the building, from the interface with the soil, will interfere with the newly incident pulses just transmitted into the building. Eventually, constructive interference will occur, and the standing waves will be formed, which are the fixed-base modes of vibration of the building.

The wave travel times can be detected by tracing the propagation of a pulse. Such a pulse can be created by signal processing of recorded earthquake response data, i. e. by deconvolution of the recorded response, which results in the *system impulse response functions*. These can be obtained by computing the transfer-functions between the motion at a particular level and the reference motion, and then computing inverse Fourier transform. The location of

the virtual source would coincide with the location of the sensor that recorded the reference motion. Let  $u_{ref}(t)$  be the reference motion,  $u_i(t)$  be the motion at level  $i$ . Then the impulse response at that level,  $h_i(t)$ , can be computed as

$$h_i(t) = FT^{-1} \left\{ \frac{\hat{u}_i(\omega)\hat{u}_{ref}^*(\omega)}{|\hat{u}_{ref}(\omega)|^2 + \varepsilon} \right\}, \tag{6}$$

where the hat symbol indicates Fourier transform, the bar indicates complex conjugate, and  $\varepsilon$  is a regularization parameter, used to avoid dividing by a very small number [39]. At the reference level, the transfer-function is unity, and its inverse is a Dirac delta function.

Proof-of-concept applications to two buildings damaged by earthquakes, and to an analytical model of a building-foundation-soil system showed that the method (1) is robust when applied to damaging levels of earthquake response data, (2) is not sensitive to the effects of soil-structure interaction, and (3) is local in nature (i. e. gave results consistent with the spatial distribution and degree of the observed damage) [44,45,53,54]. The damaged buildings are the former Imperial County Services Building – a 6-story RC structure in El Centro, California, damaged by the 1979 Imperial County earthquake and later demolished [52,54], and the 7-story RC building in Van Nuys, damaged by both the 1971 San Fernando and the 1994 Northridge earthquakes [53,62]. Another application is to a building in Banja Luka in former Yugoslavia, using records of 20 earthquakes, one of which led to levels of response that might have caused structural damage, but no damage was reported following a detailed inspection [65]. This study was aimed at learning about the *threshold change* in the building fixed-base frequency, estimated from wave travel time, associated with damage.

While this method is local, its spatial resolution is limited by the number of sensors. A minimum of two sensors (at the base and at the roof) are required to determine if the structure has been damaged, and additional sensors at the intermediate floors would help point out the part of the structure that has been damaged. For example, one additional sensor between these two would help identify if the damage has been in the part of the structure above or beyond that sensor.

There have been only a few publications in the literature on wave propagation methods for structural health monitoring and damage detection in civil structures other than the NDT methods [20,35,37,41,53,54,64]. Similar wave travel time analyses (using deconvolution or the NIOM method) of buildings that have not been damaged include Kawakami and Oyunchimeg [24,25], Snieder and

Şafak [39], Kohler et al. [26], and Todorovska [44,45]. These studies show that the wave travel times reflect well the characteristics of the buildings studied. A recent review can be found in [53,54].

In conclusion, the advantages of this wave method are its local nature achieved with a relatively small number of sensors, its insensitivity to the effects of soil-structure interaction, and the ability to estimate the structural fixed-base frequency using data from only two sensors (one at the base and one at the roof), which will extend the usability of old data. An outstanding issue to its implementation is the lack of a knowledge base relating changes in wave travel times (and fixed-base frequency) with different levels of damage for different types of structures.

### Inter-Story Drift

Structural damage of a building under seismic loads occurs primarily due to large *lateral* deformations of its columns and shear walls, as they are by design much stiffer in the vertical direction to carry the static gravity loads. A measure of the lateral deformations is the inter-story drift. The inter-story drift is also a good indicator of the damage to the architectural (nonstructural) components (partition walls, facade, windows, etc.), which can be costly. As the value of the structure is only about 10–25% of the total construction cost of a building, the damage to the nonstructural components represents a significant portion of the total repair cost following an earthquake. For these reasons, the inter-story drift is one of the performance parameters considered in design. It is important to note that the structural and nonstructural damage are related only to the drift caused by *deformation* of the structure, and not by the drift caused by *rigid body motion*.

The level of structural damage (to a particular element and to the structure as a whole) associated with a particular level of inter-story drift varies depending on the type of structure, height and ductility, among other factors, and is still *not a completely resolved issue in structural engineering* [14]. To illustrate this correlation, Table 1 shows some values of drift associated with different levels of damage (simplified from [14]) for ductile and nonductile moment resisting frames (MRF), and based on experimental data, field observations and measurements and theoretical analyses. (Ductile are those structures that can undergo large nonlinear deformations before failure as opposed to the nonductile ones, which experience quick brittle failure soon after exceeding the linear range of response). It can be seen from Table 1 that, roughly, inter-story drift > 1% for ductile and > 0.5% for nonductile moment resisting frames causes damage beyond repair, and drift > 3% and

Earthquake Damage: Detection and Early Warning in Man-Made Structures, Table 1

Drift ratios (in %) associated with various damage levels (based on [14])

State of damage	Ductile MRF	Nonductile MRF
No damage	< 0.2	< 0.1
Repairable damage		
Light	0.4	0.2
Moderate	< 1.0	< 0.5
Irreparable damage	> 1.0	> 0.5
Severe damage, life safe, partial collapse	1.8	0.8
Collapse	> 3.0	> 1.0

> 1% for the same type of frames is significant for life safety.

Drift-based assessment of the state of damage of a building following an earthquake would require: (1) measurement of the drift during the earthquake shaking, and (2) knowledge base of values of drift associated with different states of damage for the particular structure. The accuracy of the assessment would depend on the accuracy of both the measurements and knowledge base, as discussed in the following.

The drift is commonly estimated from the difference of displacements obtained by double integration of recorded velocities in the structure [28]. While in the past these calculations were performed by specialists, after the data had been manually collected, at present, such calculations can be done in near real time either using telemetry or at the site by “client” software supplied by the instrument manufacturer. Such estimates of drift however are limited by: (1) the inability to estimate reliably the *static* component of the drift associated with permanent deformations (i. e. the drift at  $\omega \rightarrow 0$ ), which is not negligible for structures experiencing large deformations in the nonlinear range of response, when damage occurs, and (2) the inability to separate the drift due to deformation of the structure (which is directly related to damage) from the drift due to rigid body rocking because of inadequate instrumentation.

The inability to estimate reliably the static part of the displacement (and drift) is due to the fact that the traditional (translational) sensors are sensitive also to rotational motions of their support [16,59], which produce errors in the recorded translations and the integrated displacements mimicking permanent displacement [16]. This problem can be solved, by deploying sensors recording all six components of motion (three translations and three rotations) and performing appropriate instrument correction. Such future deployments and their assessment are of

interest to and have been advocated by the International Working Group on Rotational Seismology [29].

The *dynamic* (at  $\omega > 0$ ) drift due to deformation of the structure only is not simple to estimate, especially for structures on soft soil, with significant rocking response of their foundation. The rocking motions of the foundation are due to the wave nature of the incident seismic waves, and also due to feedback forces from the structure acting on the soil. The foundation rocking results in relative horizontal displacement between two floors and is not related to damage. Such excessive relative displacements, can affect the stability of the structure, which may collapse before yielding occurs in its members, but that is out of the scope of this article. The *average dynamic* floor rocking can be calculated from the difference of vertical motions recorded by two sensors on that floor, assuming the floor slab is rigid, but such sensor configurations are not routinely installed even in recent denser deployments in buildings. If the building foundation is fairly rigid, the rigid body rocking of the structure can be estimated from two vertical sensors at foundation level. Unfortunately, even such data is lacking for most of the significant earthquake records in buildings, and even in recent dense deployments (e. g. in [6]). It is noted that vertical sensors are also less sensitive to rotation of their support and to cross-axis motion [15,43].

It should be noted here that permanent displacements can be measured directly using GPS (Global Positioning System), and there have been such deployments in long period structures [5]. While GPS measurements are not contaminated by rotation, they are limited by the fact that what is measured are only the roof *absolute* displacements, which makes it impossible to separate the displacement due to deformation of the structure from the rigid body horizontal translation and rocking. The other two limitations in the presently available systems are the small sampling rate (10–20 Hz) and the limited resolution of GPS for civilian applications ( $\pm 1$  cm horizontally and  $\pm 2$  cm vertically; [5]).

Damage estimation algorithms based on published damage versus drift relationships (e. g. in [1]) started to be implemented by manufacturers of strong motion instruments in structural seismic monitoring systems but there is no data yet of their performance. Despite errors in the assessment resulting from the mentioned difficulties, such algorithms are robust when applied to earthquake data and can be useful within a suite of methods.

Matrices like the one in Table 1 [14] can serve as a knowledge base in assessing the class of damage state for a given maximum drift reached. Such matrices are associated with scatter, due to the variability from one struc-

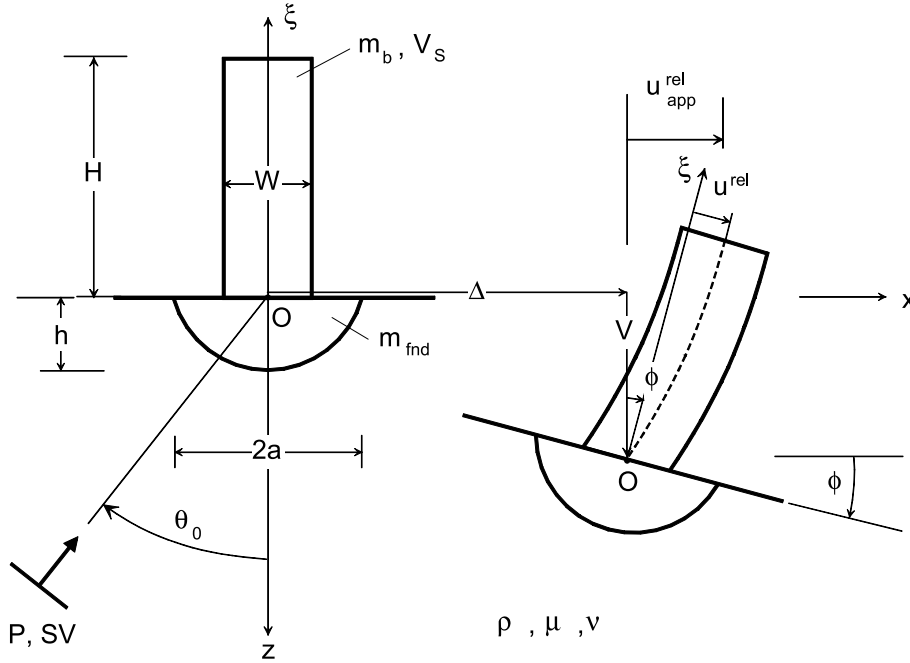
ture to another within the same class. Another source of scatter is the source of the data. Because of the limited amount of full-scale earthquake response data, information for such relationships is complemented by laboratory data (e. g. pushover tests). While the drift in the former is the total drift, which includes the drift due to rigid body motion, the drift in the latter is only due to deformation of the structural elements.

In conclusion, outstanding issues in measuring the drifts are: (1) separation of the drift due to deformation of the structure only, and (2) estimation of the static component of the drift. It may be possible to resolve these issues by deploying six degrees-of-freedom sensors. An outstanding issue in the knowledge base is more accurate drift versus damage state relations for specific buildings.

### System Identification Considering the Effects of Soil-Structure Interaction – Example

As mentioned earlier, both for frequency-based identification and for damage assessment based on drift, the effects of soil-structure interaction have a significant effect on the reliability of the estimation. This section presents a simple soil-structure interaction model, in which the building is represented as a shear beam. It illustrates the different contributions to the inter-story drift, the difference between fixed-base and apparent building frequencies and their relationship, and the relationship between the model fixed-base frequencies and wave travel times. More detailed analysis can be found in [44,45].

The model is shown in Fig. 2. It consists of a shear beam of height  $H$  and fundamental fixed-base frequency of vibration  $f_1$ , representing the building, and a rigid foundation of width  $2a$  embedded in elastic half-space. The excitation, in general, is an incident wave (plane P and SV or a Rayleigh wave). The motion on the surface of the half-space in the absence of any structures and excavations, acting as scatterers, is commonly referred to as “free-field.” The effective motion at the base of the building differs from the free field motion at the half-space surface, because of two phenomena: (1) scattering and diffraction of the incident waves from the excavation for the foundation, (2) differential displacements due to feedback forces from the building and foundation acting on the half-space through the contact with the foundation. The former phenomenon is referred to as kinematic and the latter as dynamic or inertial interaction. For the linear problem and a rigid foundation, the two problems can be solved separately and their effects superimposed. The apparent frequency, which is the one estimated from peaks of energy distributions of the response, is affected mostly by the dy-



Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 2  
Soil-structure interaction model

dynamic interaction, while the apparent drift is also affected by the type of incident waves, and the point rotation they produce on the ground surface, and the size of the foundation relative to the wavelength of the incident waves. To consider only the effects of the dynamic interaction, it suffices to take excitation consisting of only horizontal foundation driving motion.

The building foundation has three degrees of freedom: horizontal translation  $\Delta$ , vertical translation  $V$ , and rocking angle  $\phi$ . In the *linear approximation* of the solution, only the horizontal and rocking motions are coupled, while the vertical motions are independent of the other degrees of freedom. Let the excitation be horizontal driving motion  $\Delta_{inp}$ . Then, the foundation response will be

$$\Delta = \Delta_{inp} + \Delta_{fb} \quad (7)$$

$$\phi = \phi_{fb}, \quad \phi_{inp} = 0, \quad (8)$$

where  $\Delta_{fb}$  and  $\phi_{fb}$  are the feedback horizontal displacement and rocking angle, which depend on the stiffness of the foundation and on the forces with which the structure and foundation interact with the soil, and are the solution of the dynamic equilibrium equations of the foundation. The building horizontal displacement  $u(\xi)$ , as a function of the height  $\xi$  measured from ground level, is a sum of

three terms

$$u(\xi) = \Delta + \phi\xi + u^{rel}(\xi), \quad (9)$$

where the first two terms are from the translation and rotation as a rigid body, and the third term is relative displacement from *deformation* during the vibration. The damage in the building will depend only on  $u^{rel}(\xi)$ . It is noted here that including the coupling between horizontal and vertical motions turns a linear elastic system into a nonlinear elastic system [21].

For the shear beam,  $u(\xi)$  can be computed as a solution of the wave equation for moving boundary conditions. It can be represented as a sum of motions  $u_{\Delta}(\xi)$ , which is due to translation of the base only, and  $u_{\phi}(\xi)$ , which is due to rotation of the base only, where

$$u_{\Delta}(\xi) = \Delta \frac{\cos k_S(H - \xi)}{\cos k_S H} \quad (10)$$

$$u_{\phi}(\xi) = \frac{\phi \sin k_S \xi}{k_S \cos k_S H}, \quad (11)$$

where  $k_S = \omega/V_S$  and  $V_S = \sqrt{\mu_b/\rho_b}$  is the shear wave velocity in the building. Equations (10) and (11), reflecting the interference conditions in the building, imply fundamental fixed-base frequency of the structure  $f_1 = V_S/(4H)$  and overtones at  $f_n = (2n - 1)V_S/(4H)$ ,  $n > 1$ . If  $\tau$  is

the time it takes for a wave to propagate from the base (at  $\xi = 0$ ) to the top (at  $\xi = H$ ), the interference conditions in the shear beam imply

$$f_1 = 1/(4\tau). \quad (12)$$

Let us now consider the frequencies of vibration. If the building did not deform, the foundation and the building would oscillate freely as a rigid body with frequency  $f_{RB}$  such that

$$\frac{1}{f_{RB}^2} = \frac{1}{f_H^2} + \frac{1}{f_R^2}, \quad (13)$$

where  $f_H$  and  $f_R$ , referred to as the horizontal and rocking foundation frequency, depend on the stiffness of the foundation and on the system mass [33]. If the building is flexible and would freely vibrate on a fixed base with fundamental frequency  $f_1$ , on flexible soil it would freely vibrate with fundamental frequency  $f_{sys}$ , which is the soil-structure *system* frequency, and is a result of the coupling between the vibration of the building and the vibrations of the foundation. The following relationship holds approximately

$$\frac{1}{f_{sys}^2} = \frac{1}{f_{RB}^2} + \frac{1}{f_1^2}. \quad (14)$$

This relationship implies that  $f_{sys} < \min(f_1, f_{RB})$ , i. e.  $f_{sys}$  is always lower than both  $f_1$  and  $f_{RB}$ , and that if  $f_1$  and  $f_{RB}$  differ significantly, then  $f_{sys}$  would be closer to the smaller one of them. How much  $f_1$  would differ from  $f_{sys}$  would depend on the *relative* stiffness of the soil compared to the building. The energy of the response of vibrating systems is concentrated around their resonant frequencies, which are measured from the frequency of the peaks of the corresponding transfer-functions. The energy of the building roof response (absolute and relative) will be concentrated around  $f = f_{sys}$ .

Of interest is how to estimate the relevant quantities from recorded response during an earthquake. If the building foundation is fairly rigid, and there are at least two appropriately located vertical sensors to compute the foundation rocking  $\varphi$  (average value), then  $u^{rel}(\xi)$  can be computed. To measure  $f_{sys}$ , the driving motion  $\Delta_{inp}$  is also needed, so that the transfer function between the building response and  $\Delta_{inp}$  can be computed. Motion from a nearby free-field site can be used for that purpose, but such sites are often not available, and truly free-field sites practically do not exist in urban areas. Also, for most instrumented buildings, the foundation rocking cannot be estimated because of the lack of two vertical sensors even under the ideal conditions that the foundation behaves as rigid.

Consequently, in reality, for most instrumented buildings, the true relative roof displacement cannot be estimated from the recorded data, but only the *apparent* relative displacement

$$\begin{aligned} u_{app}^{rel}(H) &= u(H) - \Delta \\ &= u^{rel}(H) + \varphi H \end{aligned} \quad (15)$$

which includes the contribution of the roof displacement due to rigid body rotation, and only the transfer-function  $|u_{app}^{rel}(H)/\Delta|$  can be computed, the peak of which gives the *apparent* building frequency  $f_{app}$ , which is different from both the fixed base frequency and the system frequency.

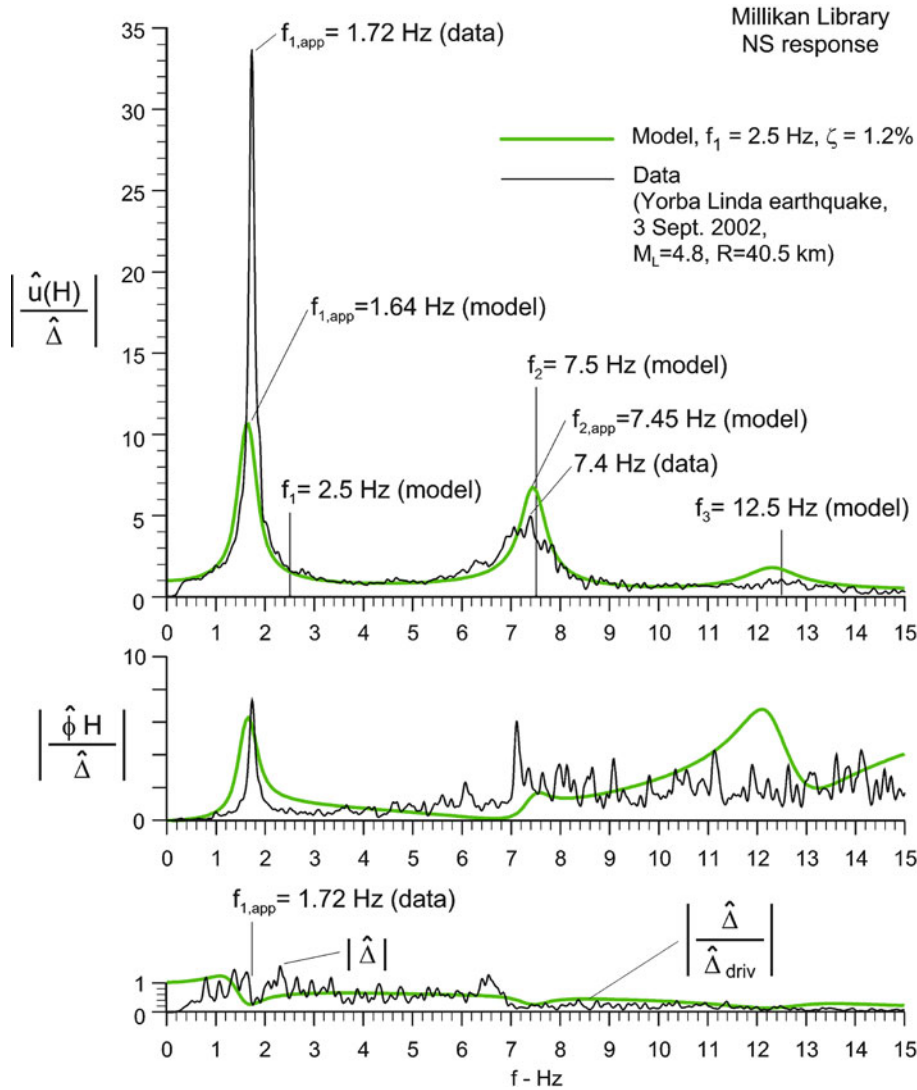
What is of interest for structural health monitoring is that the energy of the roof response will be concentrated around  $f = f_{sys}$ , not around  $f = f_1$ . It is also significant that the damage will depend on  $u^{rel}(\xi)$ , while what is usually measured is  $u^{rel}(\xi) + \varphi H$ .

Figure 3 (redrawn from [45]) shows a comparison of model and measured transfer-functions for a model of the NS response of Millikan library. The model has  $f_1 = 2.5$  Hz, height  $H = 44$  m, shear wave velocity in the soil 300 m/s, and Poisson ratio 0.333, while the data are from the Yorba Linda earthquake of 2002. Figure 4 shows the corresponding impulse response function for a virtual source at the ground floor. It can be seen that there is a very good qualitative agreement despite the model simplicity and roughly chosen parameters.

### Novelty Detection in the Recorded Response

Novelty detection is used in data mining to detect unusual events in data. The unusual events are *outliers* deviating from the *trend*. Within the framework of multi-resolution analysis, the trends and novelties are determined by splitting the signal in two subbands, one smooth (low frequency) and the other one containing the detail (high frequency). By consecutively splitting the smooth subband, trends and detail are obtained at different resolution levels. If  $J$  is the last level, then there will be  $J$  detail subbands  $D_i$ ,  $i = 1, \dots, J$  and one smooth subband  $S_J$ . The last smooth subband can be expanded in a basis of scaling functions  $\varphi_{j,k}(t)$ , and each of the detail subbands – in a basis of wavelet functions  $\psi_{j,k}(t)$ , leading to the representation of a discrete time signal  $s[n]$ ,  $n = 1, \dots, N$

$$\begin{aligned} s[n] &= \sum_{j=1}^J D_j[n] + S_J[n] \\ &= \sum_{j=1}^J \sum_{k=1}^{N/2^j} d_{j,k} \psi_{j,k}[n] + \sum_{k=1}^{N/2^J} s_{J,k} \varphi_{J,k}[n]. \end{aligned} \quad (16)$$



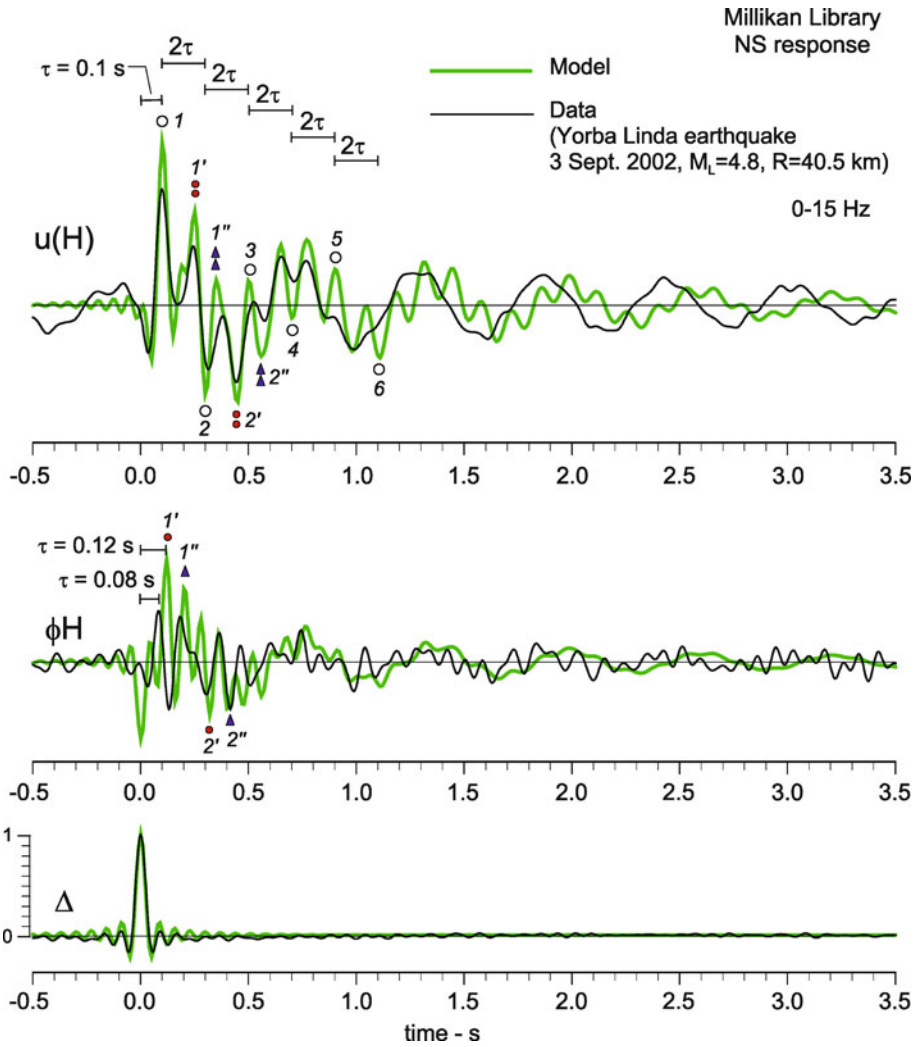
Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 3  
 Model (thick line) versus Yorba Linda, 2002, earthquake (thin line) NS response: transfer-functions of roof response (top), and base rocking response (middle) with respect to horizontal response of ground level. The plot in the bottom shows the model horizontal response at ground level for unit driving motion (thick line), and the Fourier spectrum of the earthquake response at ground level (thin line) on a relative scale

The coefficients of the expansion,  $d_{j,k}$  and  $s_{j,k}$ , can be computed using the fast wavelet transform. The pyramid algorithm on which it is based [36] is shown in Fig. 5.

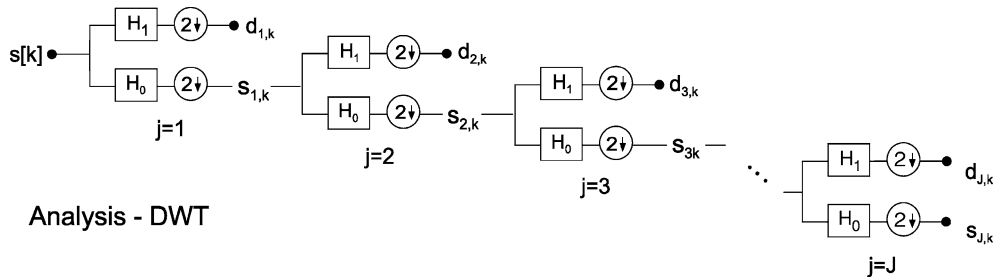
The wavelet functions  $\psi_{j,k}[n]$ , where  $j$  is a level and  $k$  is the time shift, are localized both in frequency and in time, and each wavelet is a projection of the signal onto the corresponding tile of the phase plane. For a wavelet basis that is orthonormal, the square of a wavelet coefficient represents the energy of the signal in the corresponding tile of the phase plane.

The damage detection method is based on the assumption that, when damage occurs and there is a sudden loss of stiffness, there will be some abrupt change in the response that would produce novelties. These would be seen as spikes in the time series of the square of the detail coefficients (e.g.  $d_{1,k}^2$ ,  $k = 1, \dots, N/2$  for the highest detail coefficients) plotted versus the central time of the corresponding wavelet. These spikes indicate high frequency energy in the response. For data with Nyquist frequency 25 Hz, the novelties can be best seen in the highest detail





Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 4  
 Model (*thick line*) versus Yorba Linda, 2002, earthquake (*thin line*) NS response: impulse responses of roof (*top*), and base rocking (*middle*) to input impulse at ground level (*bottom*), i. e. roof horizontal motion, foundation rocking and ground level horizontal motions deconvolved with the resultant horizontal motion at ground level



Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 5  
 The pyramid algorithm for the fast wavelet transform



a



b



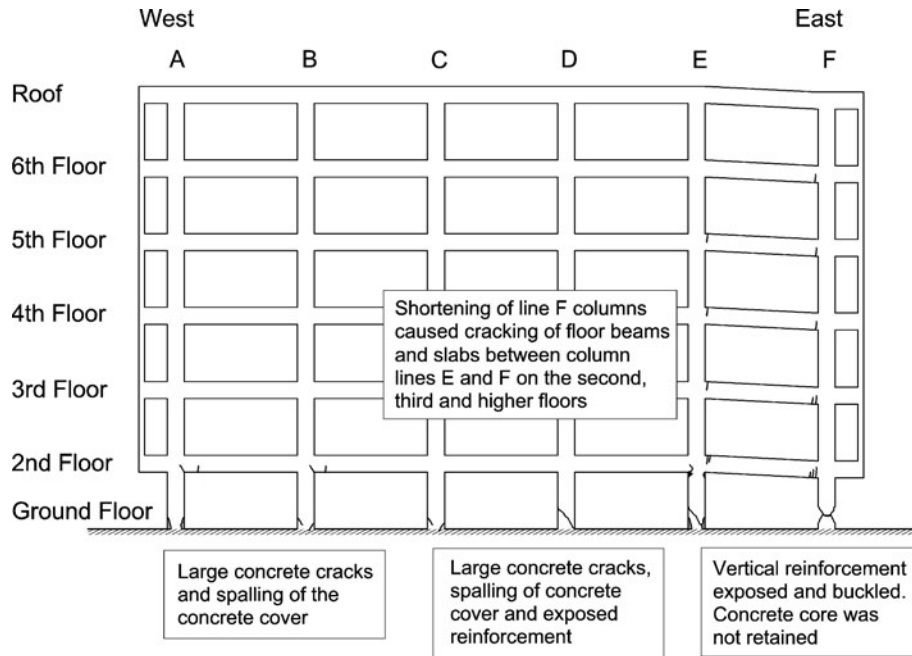
c

Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 6  
Imperial County Services (ICS) building: a view (towards north); b photographs of damage: columns F1 and F2 at the ground floor; and c column F1

subband (12.5 to 25 Hz), which is away from the frequency of the first few modes of typical buildings, where the response is amplified by the structure.

Applications to numerically simulated response of simple models with postulated damage [17,18] have shown that this method *can point out very precisely the time of damage*, but the changes are detectable only if the spikes in the wavelet coefficients are above the noise. Further, the magnitude of the novelties is larger if the sensor is closer

to the location of the damaged member, and may be difficult to detect if the sensor is far from the location of damage. There have been only few applications to earthquake response records in buildings. Rezaei et al. [38] and Hou et al. [18] have shown that there *are* novelties (spikes) in earthquake records of damaged buildings, but have not discussed and extracted other possible causes. Todorovska and Trifunac [51,55] presented a detailed analysis of the correspondence between the spatial distribution and am-



Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 7  
 ICS building: schematic representation of the damage following the 1979 Imperial Valley earthquake (reproduced from [27])

plitudes of the detected novelties and the observed damage for the Imperial County Services building (see illustrations in Sect. “Examples”), and also analyzed the “noise.” Their study shows that: (1) the spatial distribution and magnitudes of the novelties were generally consistent with the spatial distribution and degree of the observed damage, (2) the timing of those suggesting major damage agreed with the time of significant drops in frequency and of large inter-story drifts, and (3) were much larger in the transverse response, in which the building was stiffer.

In summary, the method of novelties is very effective in determining the time of occurrence of damage, and can reveal the spatial distribution and degree of damage if there is sufficiently dense instrumentation. Unresolved issues are how to distinguish novelties that are not caused by damage, and small novelties due to larger damage far from the sensor from those due to small damage close to the sensor.

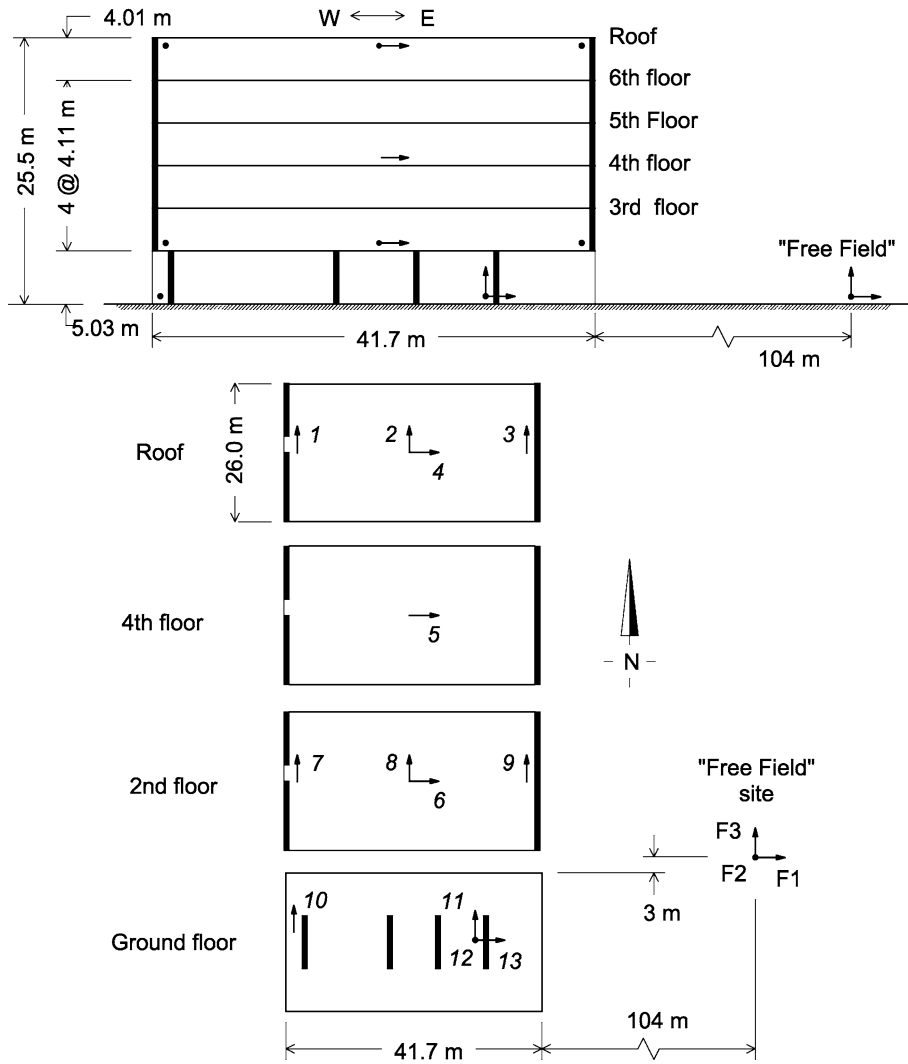
## Examples

In this section, the methods previously described are illustrated for the former Imperial County Services (ICS) building – a rare example of an instrumented building damaged by an earthquake, for which description of damage and the strong motion data are available. The build-

ing is first described and the strong motion data of the Imperial Valley earthquake, which severely damaged the building.

The ICS building was a 6-story reinforced concrete structure located in El Centro, California (Fig. 6a). It was designed in compliance with the 1967 Uniform Building Code, and its construction was completed in 1969. It had plan dimensions  $41.70 \times 26.02$  m, height 25.48 m, and pile foundation. Up to depth of 9 m, the underlying soil consisted of soft to medium-stiff damp sandy clay with organic materials, with inter-layers of medium dense moist sand, and beneath 9 m it consisted of stiff, moist sandy clay and silty clay [27].

The building was severely damaged by the Imperial Valley earthquake of October 15, 1979 ( $M = 6.6$ ), and was later demolished (Fig. 6b,c). Figure 7 shows a schematic representation of the observed damage. The major failure occurred in the columns of frame F (at the east end of the building) at the ground floor. The vertical reinforcement was exposed and buckled, and the core concrete could not be contained, resulting in sudden failure and shortening of the columns subjected to excessive axial loads. This in turn caused an incipient vertical fall of the eastern end of the building, causing cracking of the floor beams and slabs near column line F on the second, third and higher floors. Columns in lines A, B, D, and E



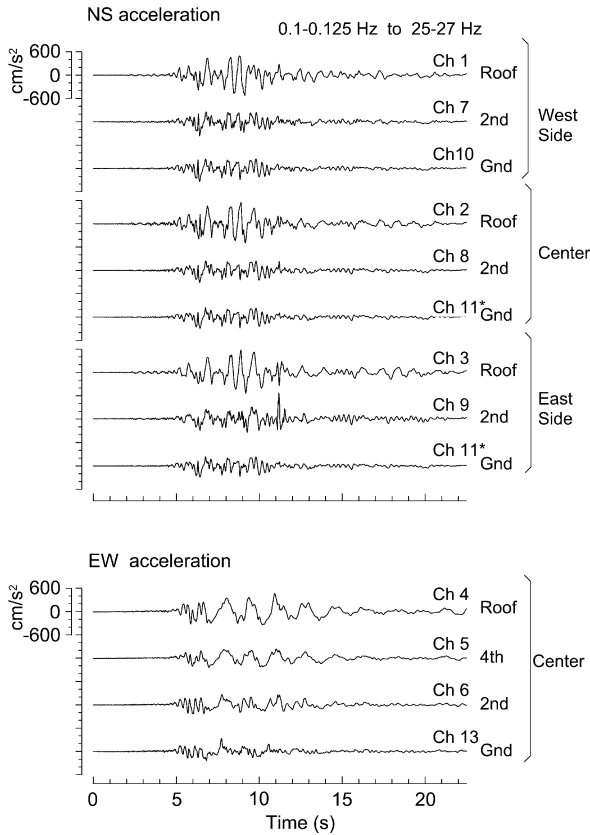
Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 8  
 ICS building: layout of the seismic monitoring array

also suffered damage. Columns in frames A and E did not suffer as extensive damage as shortening and buckling of the reinforcement in line F at the east side, but large concrete cracks and exposed reinforcement could be seen near the base. In the columns in interior frames B through E, visible cracks and spalling of the concrete cover were also observed [27].

The building was instrumented by a 16-channel seismic monitoring array (installed by the California Geological Survey, formerly the California Division of Mines and Geology) consisting of a 13-channel structural array of force balance accelerometers (FBA-1), with a central analog recording system, and a tri-axial SMA-1 accelerome-

ter in the “free field,” approximately 104 m east from the northeast corner of the building (Fig. 8). Figure 9 shows the accelerations (corrected) during the Imperial Valley earthquake. The peak accelerations at the roof and ground floor were 571  $\text{cm/s}^2$  and 339  $\text{cm/s}^2$  in the NS direction and 461  $\text{cm/s}^2$  and 331  $\text{cm/s}^2$  in the EW direction.

Figure 10 shows the NS (top) and EW (bottom) inter-story drifts computed from band-pass filtered displacements (between 0.1–0.125 Hz and 25–27 Hz) (redrawn from [52]). Hence, they represent only a limited view of the actual drifts – through a tapered window in the frequency domain, and a combination of the drift due to *rigid body rocking* (one of the effects of soil-structure interac-



Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 9

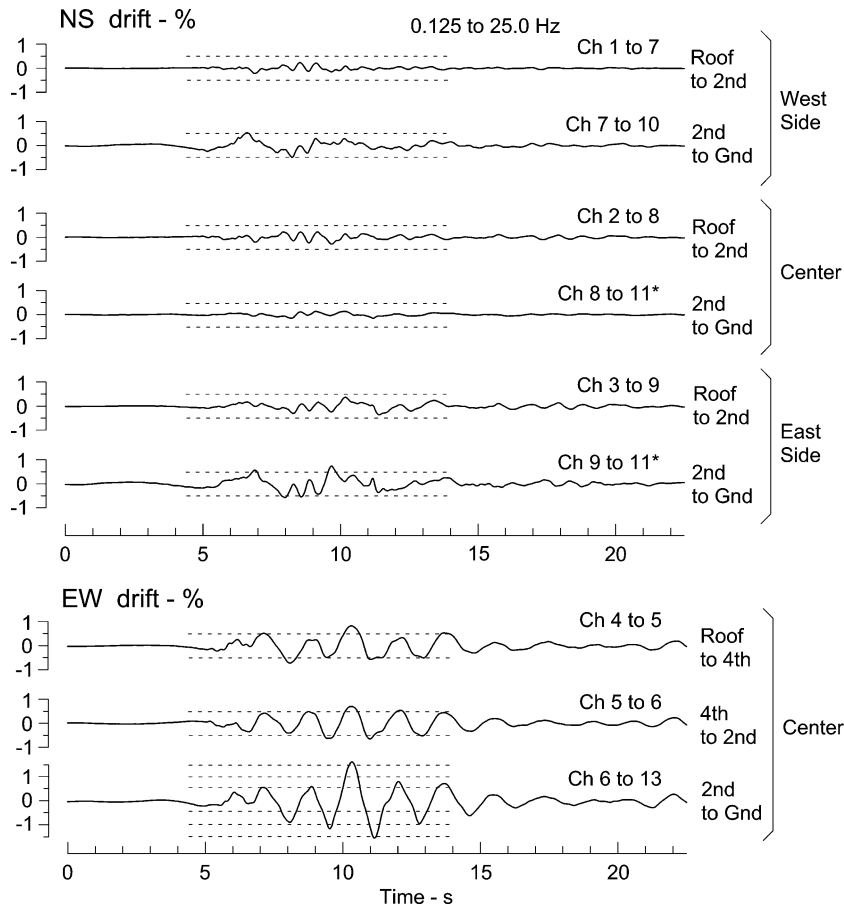
ICS building: accelerations (NS and EW components) recorded during the 1979 Imperial Valley earthquake

tion) and drift due to relative *deformation* of the building. The horizontal lines show 0.5%, 1%, and 1.5% drift levels. The plotted drifts suggest: (1) “soft” first story in both NS and EW directions, (2) larger flexibility in the EW direction, and (3) significant torsional response, probably amplified by the wave passage, and by the asymmetric distribution of stiffness in the NS direction at the soft first story (see Fig. 8). It can be seen that during the most severe shaking, the inter-story drifts exceeded 0.5% for NS and 1.5% for EW motions, consistent with irreparable to severe damage (Table 1).

Figure 11 shows results of time frequency analysis (using Gabor transform) for the EW response (redrawn from [52]). Parts a and b show the ground floor accelerations, and the roof relative displacements (at the center of the building), both included as background information. Part c shows the skeleton (the thicker line), which is a smoothed estimate of the amplitude envelope of the estimated signal, which is the relative roof response near the

first system frequency. The thin line is the actual amplitude envelope (that for the broad-band signal), determined by Hilbert transform. This plot is included to help monitor rapid changes in the amplitude of the signal and artifacts in the estimate of instantaneous frequency caused by violations of the asymptoticity condition. Part d shows the Fourier spectra of the relative roof displacement (the solid line), and of the ground floor acceleration (the dashed line, on a relative scale), both included as background information. Part e shows the variations of the system frequency as a function of amplitude of response (estimated from the ridge and skeleton of the Gabor transform), with the arrows indicating the direction of increasing time. Part f shows the variations of the system frequency versus time, estimated from the ridge of the Gabor transform. The missing segments and the dashed lines in parts e and f correspond to time intervals where the estimates cannot be obtained or are not believed to be reliable, due to rapid variations of the envelope of the amplitude, and/or very weak “signal.” The rectangle in part f with sides  $2\sigma_t = 1.42$  s and  $2\sigma_\nu = 0.22$  Hz illustrates the theoretical uncertainty of the estimates due to the finite resolution of the Gabor transform. In practice, the uncertainty is larger due to violations of the asymptoticity assumption. Finally, the numbered open dots (occurring at different times in parts b, c, e, and f) correspond to some characteristic points in time associated with changes in amplitude or frequency, as well as a few other points in-between. It can be seen that the EW frequency dropped rapidly from  $\nu \approx 0.88$  Hz at  $t \approx 3.5$  s to  $\nu \approx 0.67$  Hz at  $t \approx 7$  s ( $\Delta\nu \approx 0.21$  Hz  $> \sigma_\nu$ ;  $\Delta\nu/\nu \approx 24\%$ ), and then continued to drop gradually to  $\nu \approx 0.53$  Hz at  $t \approx 17$  s ( $\Delta\nu \approx 0.14$  Hz  $\approx \sigma_\nu$ ;  $\Delta\nu/\nu \approx 20.9\%$ ).

Figure 12 shows results of impulse response analysis for the EW response (redrawn from [53]). The different types of lines correspond to different time intervals of the recorded motion, before, during, and after the major damage occurred:  $t < 7$  s,  $7 < t < 13$  s, and  $t > 13$  s (based on novelty analysis, discussed below). The plots on the left correspond to an input impulse at the ground floor, and those on the right – to an input impulse at the top. The latter plots show two waves propagating downwards, one acausal (in negative time, representing the wave going up) and one causal (in positive time). The delays in the pulse arrival during the second and third time interval are obvious, and are consistent with the occurrence of damage, as determined using other methods. The wave travel times suggest, for EW motions initial wave velocities of 201 m/s through the first floor, 183 m/s between the 2nd and 4th floors, and 111 m/s between the 4th floor and roof. The velocity of an equivalent uniform shear beam is 142 m/s.



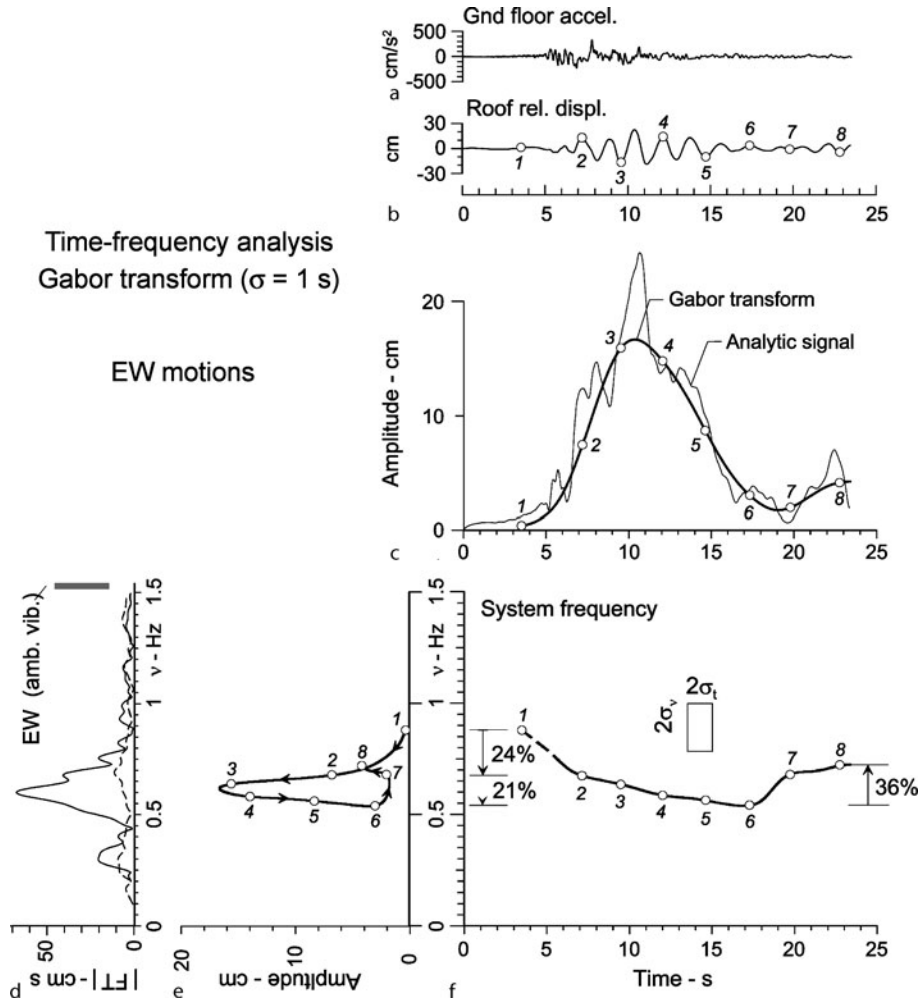
Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 10  
 ICS building: inter-story drifts during the Imperial Valley earthquake

Figure 13 shows the corresponding reduction of stiffness. It can be seen that, for EW motions, the reduction was the largest in the first story (80% during the second time window), but was also large in the upper stories (72% between the 2nd and 4th floors, and 60% between the 4th floor and roof). This is consistent with the spatial distribution of the observed damage (Fig. 6), which was the largest in the first story.

Figure 14 shows the results of novelty analysis, for the EW accelerations (part a) and for the NS accelerations at the east side of the building, where the most severe damage occurred (part b) (redrawn from [55]). The inter-story drifts (in %) between the corresponding stories are also shown, by a solid line for NS and by a dashed line for EW motions. Selected novelties are identified by letters. Novelties T1–T3 are believed to be caused by damage, and are seen in all channels. Novelties G1–G3 and g1–g4 originate in the ground motion, and L1–L6 are possibly caused by

local damage close to the sensor, or by other causes. By far the largest novelty is T3, which has amplitude more than an order of magnitude larger than all other novelties in the NS acceleration at the 2nd floor at the east side of the building, where the most severe damage (failure of the first story columns of frame F) occurred. The timing of T3 suggests that the collapse of the columns of the first story occurred at about 11.2 s after trigger. The other two large novelties consistent with the observed damage, T1 and T2, occurring at about 8.2 s and 9.2 s after trigger, indicating damage that weakened the structure, before the collapse of the first story columns.

Figure 15 (redrawn from [54]) shows a comparison of different values of frequency for EW motions:  $f_1$  from wave travel times (the gray line), system frequency  $f_{sys}$  estimated from time-frequency analysis (the red line; [52], and  $f_1$  using ETABS models [27]. T1, T2, and T3 mark the times of occurrence of major damage, as indicated by nov-

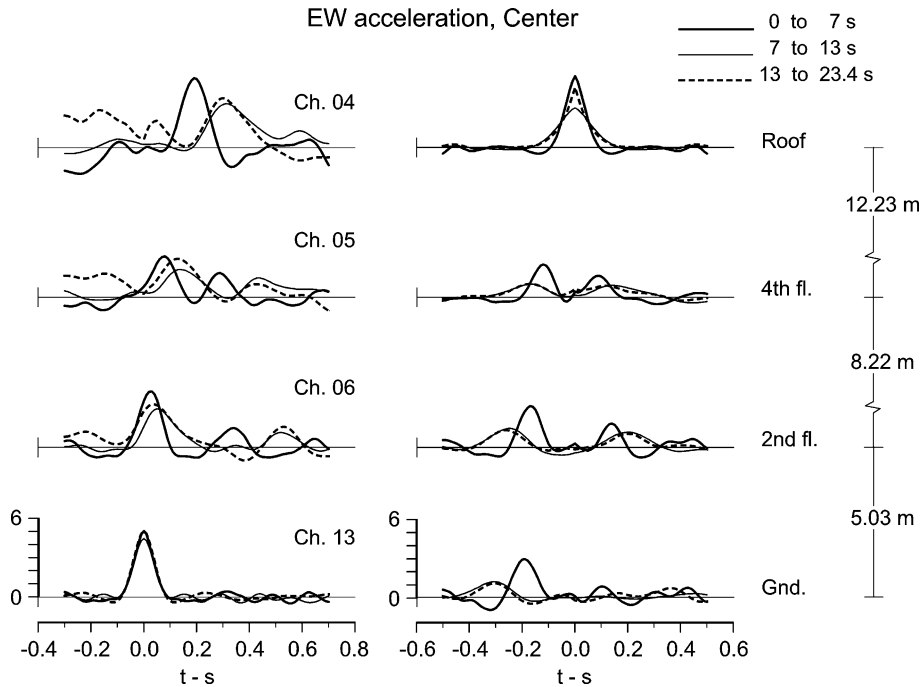


Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 11  
 ICS building: time frequency analysis for EW response

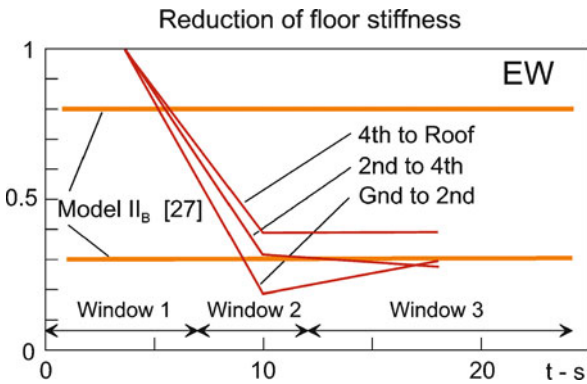
elties in the response [51,51]. It can be seen that  $f_1$  from wave travel times is consistent with the results of other independent studies.

Finally, Fig. 16 shows results for another building, the Van Nuys 7-story hotel, which has been damaged by earthquakes [53]. It shows a comparison of fixed-base frequency  $f_1$  during 11 earthquakes estimated from wave travel times, and system frequency  $f_{sys}$  during the same earthquakes estimated by time frequency analysis (Gabor transform), as well as estimates of  $f_{sys}$  during ambient vibration tests. The analysis shows that, during the San Fernando earthquake,  $f_1$  decreased by about 40% (relative to its value within the first 5 s from trigger), which corresponds to a decrease in the global rigidity of about 63%. During the Northridge earthquake,  $f_1$  decreased by about

22% (relative to its value within the first 3 s from trigger), which corresponds to a decrease in the global rigidity of about 40%. The analysis also showed that, although  $f_{sys}$  was always smaller than  $f_1$ , their difference varied, contrary to what one could expect from a linear soil-structure interaction model. It also showed that while  $f_{sys}$  was significantly lower during the Landers and Big Bear earthquakes, compared to the previous earthquakes,  $f_1$  did not change much, with is consistent with the fact that these earthquakes (which occurred about 200 km away from the building) did not cause any damage. The study concluded that monitoring changes in  $f_{sys}$  can lead to false alarms about the occurrence of damage, and that  $f_1$ , as estimated from wave travel times by the proposed method, is a much more reliable estimator of damage.



Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 12  
 ICS building: Impulse response analysis and wave travel times for EW response, and for a virtual source at the ground floor (left) and at the roof (right)



Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 13  
 ICS building: reduction of floor stiffness versus time [54]

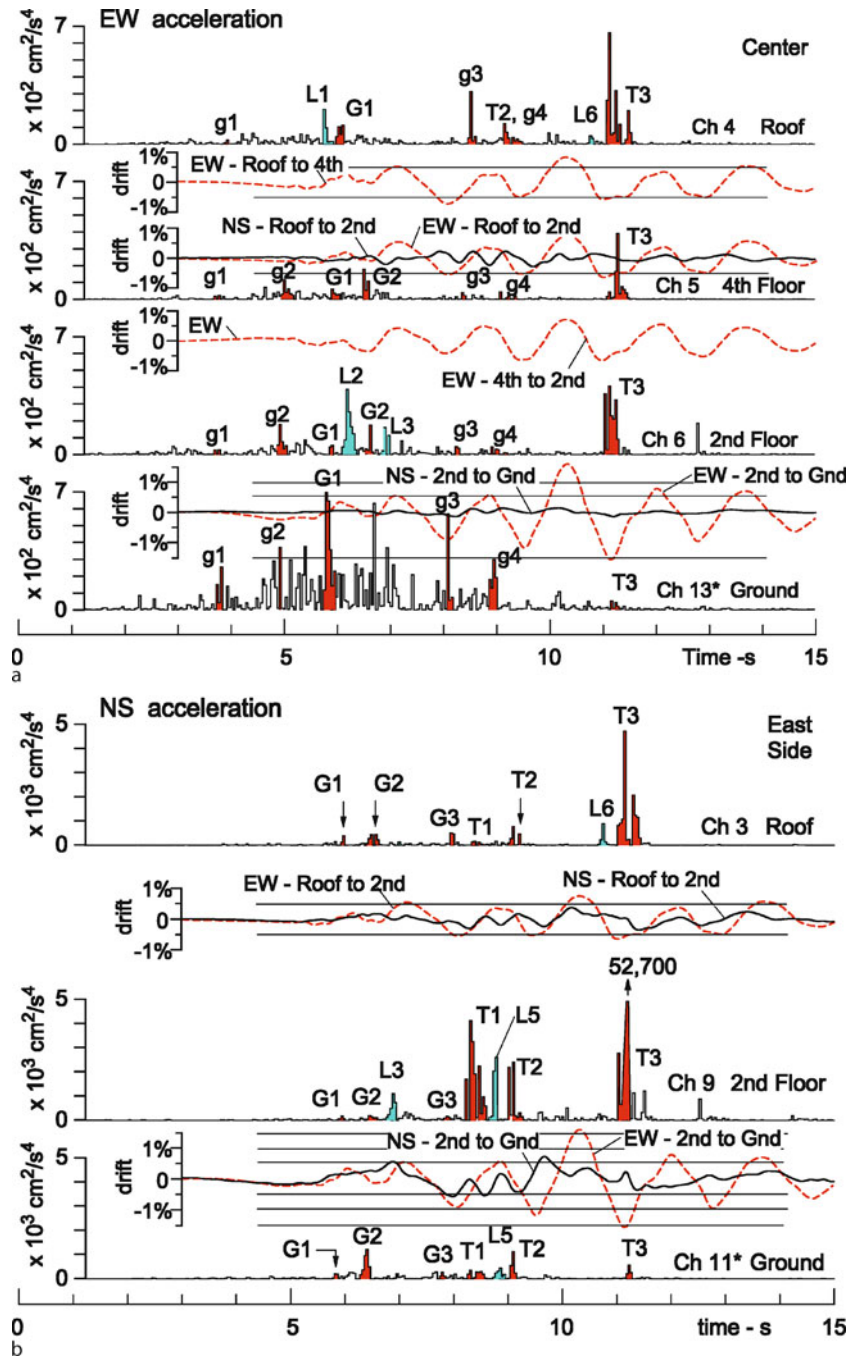
**Future Directions**

A successful system for earthquake damage detection and early warning would involve applications of technologies in fields other than structural mechanics and engineering, such as sensing, data communication, signal processing, artificial intelligence, and decision analysis. The end of the 20th and the beginning of the 21st centuries have been marked by a revolution in the development and af-

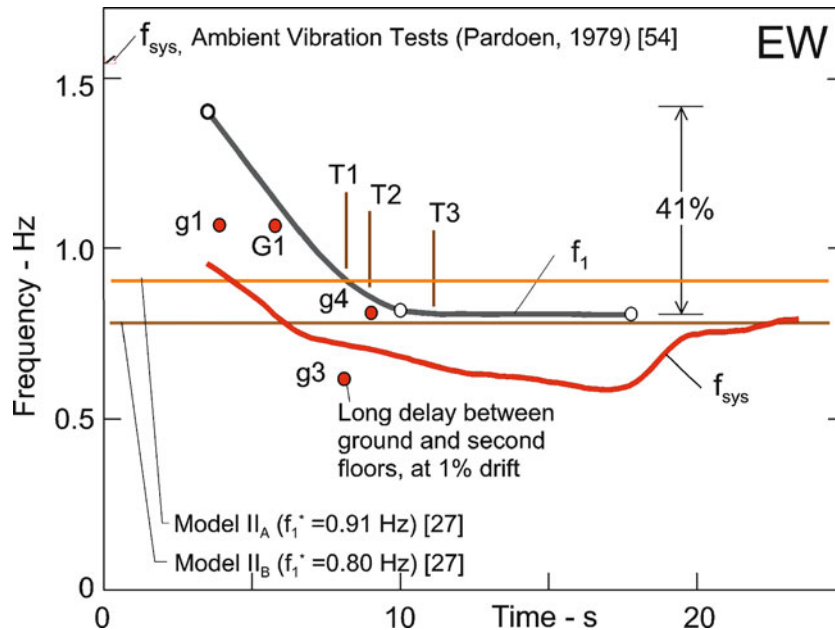
fordability of the technologies in these other fields. Much research in structural health monitoring for civil structures has been directed towards *adaptation* of these technologies to civil structures. The remaining challenge is to develop a system that is robust, redundant and well calibrated, which will neither miss significant damage nor produce many false alarms. Achieving this would require focusing the efforts and resources to further develop those methodologies that are robust when applied to real structures and data, and to calibrate them using documented full-scale data. Further enhancement of the spatial resolution of such methods would benefit from inexpensive and reliable new sensors.

All this will have to be accomplished by continuously expanding our experience in dealing with the complexities of metastable damage states of engineering structures, which will gradually become more feasible with the formulation of realistic physical models. Nevertheless, the practical outcome of most approaches in engineering will probably remain empirical. Also, the art of dynamical modeling will have to be further developed, especially for the assessment of the damaged states of engineering structures that are highly nonlinear and chaotic. In the end, in structural health monitoring, and in design of earthquake resistant structures, the fact that some modeling problems





Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 14  
**a** ICS building: novelties analysis of the EW accelerations at the center of the building. **b** Same as Fig. 13a but for the NS accelerations at the east end of the building



Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 15

ICS building: comparison of results for EW motions from different methods. System frequency  $f_{sys}$  from time frequency analysis, fixed base frequency  $f_1$  from wave travel time analysis, and times of occurrence of novelties, T1, T2, and T3

will remain will have to be accepted. However, considerable progress will be achieved if the success is gauged by the degree to which the predictions match observations in the full-scale structures, contributing towards safety and minimization of disruption and productivity of society in seismically active regions.

## Bibliography

- Applied Technology Council (1989) Procedures for post-earthquake safety evaluation of buildings. Report ATC-20. Redwood City
- Beltrami E (1987) Mathematics for Dynamic Modeling. Wiley, New York
- Carder DS (1936) Vibration observations. In: Earthquake Investigations in California 1934–1935. US Dept. of Commerce, Coast and Geological Survey, Special Publication No 201. Washington DC, pp 49–106
- Carden EP, Fanning P (2004) Vibration Based Condition Monitoring: a Review. Struct Health Monit 3(4):355–377. doi:10.1177/1475921704047500
- Celebi M, Sanli A (2002) GPS in pioneering dynamic monitoring of long-period structures. Earthq Spectr 18(1):47–61
- Celebi M, Sanli A, Sinclair M, Gallant S, Radulescu D (2004) Real-time seismic monitoring needs of a building owner—and the solution: a cooperative effort. Earthq Spectr 20(2):333–346
- Chang PC, Flatau A, Liu SC (2003) Review paper: health monitoring of civil infrastructure. Struct Health Monit 2(3):257–267
- Clinton JF, Bradford SK, Heaton TH, Favela J (2006) The observed wander of the natural frequencies in a structure. Bull Seism Soc Am 96(1):237–57
- Crawford R, Ward HS (1968) Determination of the natural periods of building. Bull Seism Soc Am 54(6A):1743–1756
- Doebling SW, Farrar CR, Prime MB (1998) A summary review of vibration-based damage identification methods. Shock Vib Dig 30(2):91–105. doi:10.1177/058310249803000201
- Doebling SW, Farrar CR, Prime MB, Shevitz DW (1996) Damage identification and health monitoring of structural and mechanical systems from changes in their vibration characteristics: a literature review. Report LA-13070-MS. Los Alamos National Laboratory, Los Alamos
- Farrar CR, Worden K (2007) An introduction to structural health monitoring. Phil Trans R Soc A 365:303–315. doi:10.1098/rsta.2006.1928
- Foutch DA, Luco JE, Trifunac MD, Udawadia FE (1975) Full-scale three-dimensional tests of structural deformations during forced excitation of a nine-story reinforced concrete building. Proc. of the US National Conference on Earthquake Engineering. Ann Arbor, pp 206–215
- Ghobarah A (2004) On drift limits associated with different damage levels. Proc. of the International Workshop on Performance-Based Design, 28 June–1 July 2004, Bled, Slovenia, pp 4321–332
- Graizer VM (1991) Inertial seismometry methods, Izvestiya. Earth Phys Akad Nauk SSSR 27(1):51–61
- Graizer VM (2005) Effect of tilt on strong motion data processing. Soil Dyn Earthq Eng 25:197–204
- Hera A, Hou Z (2004) Application of wavelet approach for ASCE structural health monitoring benchmark studies. J Eng Mech ASCE 130(1):96–104

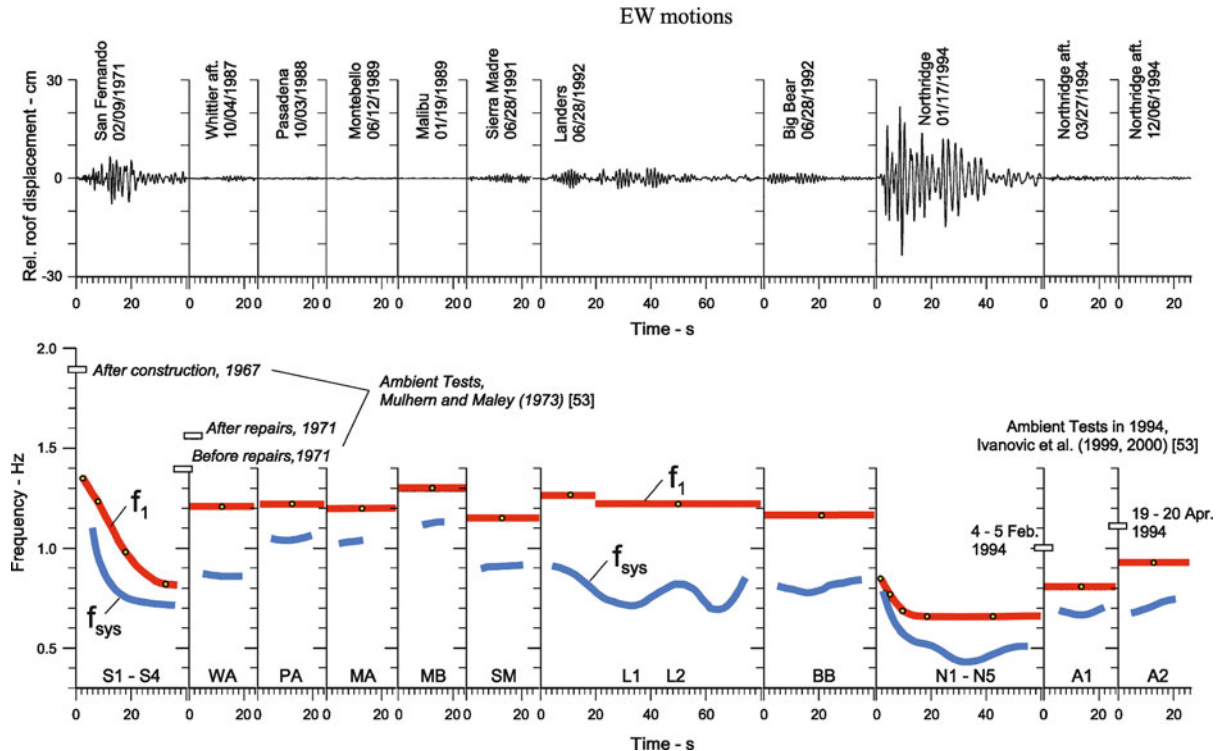
## Van Nuys Hotel

**Data: 11 earthquakes and 5 ambient vibration tests in 24 years**

**$f_{sys}$**  – form time-freq. energy distribution,  **$f_1$**  – from wave travel times.

**$f_1$  decrease:** 1971 San Fernando - by ~40%. 1994 Northridge - by ~22%.

Difference between  $f_1$  and  $f_{sys}$  is not constant (see Landers and Big Bear).



Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 16

Variations of  $f_1$  and  $f_{sys}$  in the Van Nuys building during the 11 earthquakes, between February of 1971 and December of 1994. Measured values of  $f_{sys}$  during five ambient vibration tests: (1) in 1967, following construction, (2) in 1971, after San Fernando earthquake and before repairs, (3) in 1971 after the repairs, (4) in January of 1994, 18 days after the Northridge earthquake, and (5) in April of 1994, after the building was restrained by wooden braces

18. Hou Z, Noori M, Amand R (2000) Wavelet-based approach for structural damage detection. *J Eng Mech ASCE* 126(7): 677–683
19. Hudson DE (1970) Dynamic tests of full scale structures. In: Wiegel RL (ed) *Earthquake Engineering*. Prentice Hall, pp 127–149
20. Ivanović SS, Trifunac MD, Todorovska MI (2001) On identification of damage in structures via wave travel times. In: Erdik M, Celebi M, Mihailov V, Apaydin N (eds) *Proc. of the NATO Advanced Research Workshop on Strong-Motion Instrumentation for Civil Engineering Structures*, 2–5 June, 1999. Kluwer, Istanbul, pp 447–468
21. Kalkan E, Graizer V (2007) Multi-component ground motion response spectra for coupled horizontal, vertical, angular accelerations and tilt. *Indian J Earthq Technol (special issue on Response Spectra)* 44(1):259–284
22. Kanai K (1965) Some new problems of seismic vibrations of a structure. *Proc. of the Third World Conf. Earthquake Eng*, 22 January, 1 February, 1965. Auckland and Wellington, New Zealand, pp II-260–II-275
23. Kapitaniak T (1991) *Chaotic Oscillations in Mechanical Systems*. Manchester Univ. Press, Manchester
24. Kawakami H, Oyunchimeg M (2003) Normalized input-output minimization analysis of wave propagation in buildings. *Eng Struct* 25(11):1429–1442
25. Kawakami H, Oyunchimeg M (2004) Wave propagation modeling analysis of earthquake records for buildings. *J Asian Archit Build Eng* 3(1):33–40
26. Kohler MD, Heaton T, Bradford SC (2007) Propagating waves in the steel, moment-frame Factor building recorded during earthquakes. *Bull Seism Soc Am* 97(4):1334–1345
27. Kojić S, Trifunac MD, Anderson JC (1984) A post earthquake

- response analysis of the Imperial County Services building in El Centro. Report CE 84-02. University of Southern California, Department of Civil Engineering, Los Angeles
28. Lee VW, Trifunac MD (1990) Automatic digitization and processing of accelerograms using PC. Dept. of Civil Eng. Report CE 90-03. Univ. Southern California, Los Angeles
  29. Lee WHK, Celebi M, Todorovska MI, Diggles MF (eds) (2007) Rotational Seismology and Engineering Applications. Online Proceedings for the First International Workshop, September 18–19 September, Menlo Park. US Geological Survey, Open-File Report 2007-1144 <http://pubs.usgs.gov/of/2007/1144>
  30. Lighthill J (1994) Chaos: A historical perspective. In: Newman WI, Gabrielov A, Turcotte D (eds) Nonlinear Dynamics and Predictability of Geophysical Phenomena. Geophysical Monograph 83, IUGG, vol 18 pp 1–5
  31. Liu SC, Tomizuka M, Ulsoy G (2006) Strategic issues in sensors and smart structures. Struct Control Health Monit 13:946–957
  32. Luco JE, Trifunac MD, Udwadia FE (1975) An experimental study of ground deformations caused by soil-structure interaction. Proc. US National Conf. on Earthq. Eng. Ann Arbor, MI, pp 136–145
  33. Luco JE, Trifunac MD, Wong HL (1987) On the apparent change in the dynamic behavior of a nine-story reinforced concrete building. Bull Seism Soc Am 77(6):1961–1983
  34. Luco JE, Trifunac MD, Wong HL (1988) Isolation of soil-structure interaction effects by full-scale forced vibration tests. Earthq Eng Struct Dyn 16:1–21
  35. Ma J, Pines DJ (2003) Damage detection in a building structure model under seismic excitation using dereverberated wave machines. Eng Struct 25:385–396
  36. Mallat SG (1989) Multiresolution approximations and wavelet orthonormal bases of  $L_2(\mathbb{R})$ . Trans Am Math Soc 315:69–87
  37. Oyunchimeg M, Kawakami H (2003) A new method for propagation analysis of earthquake waves in damaged buildings: Evolutionary Normalized Input-Output Minimization (NIOM). J Asian Archit Build Eng 2(1):9–16
  38. Rezai M, Rahmatian P, Ventura C (1998) Seismic data analysis of a seven-storey building using frequency response function and wavelet transform. Proc. of the NEHRP Conference and Workshop on Research on the Northridge, California Earthquake, 17 January, 1994. CUREe, Oakland, pp 421–428
  39. Snieder R, Şafak E (2006) Extracting the building response using interferometry: theory and applications to the Millikan Library in Pasadena, California. Bull Seism Soc Am 96(2):586–598
  40. Sohn H, Farrar CR, Hemez FM, Shunk DD, Stinemates DW, Nadler BR (2003) A Review of Structural Health Monitoring Literature: 1996–2001, Report LA-13976-MS. Los Alamos National Laboratory
  41. Şafak E (1998) Detection of seismic damage in multi-story buildings by using wave propagation analysis. Proc. of the Sixth US National Conf. on Earthquake Eng. EERI, Oakland, Paper No 171, pp 12
  42. Şafak E (1999) Wave propagation formulation of seismic response of multi-story buildings. J Struct Eng ASCE 125(4):426–437
  43. Todorovska MI (1998) Cross-axis sensitivity of accelerographs with pendulum like transducers: mathematical model and the inverse problem. Earthq Eng Struct Dyn 27:1031–1051
  44. Todorovska MI (2009) Seismic interferometry of a soil-structure interaction model with coupled horizontal and rocking response. Bull Seism Soc Am 99-2A (in press)
  45. Todorovska MI (2008) Soil-structure system indetification of Millikan Library north-south response during four earthquakes (1970–2002): what caused the observed wandering of the system frequencies? Bull Seism Soc Am 99-2A (in press)
  46. Todorovska MI, Al Rjoub Y (2006) Effects of rainfall on soil-structure system frequency: examples based on poroelasticity and a comparison with full-scale measurements. Soil Dyn Earthq Eng 26(6–7):708–717
  47. Todorovska MI, Al Rjoub Y (2008) Environmental effects on measured structural frequencies – model prediction of short term shift during heavy rainfall and comparison with full-scale observations. Struct Control Health Monit (in press)[doi:10.1002/stc.260](https://doi.org/10.1002/stc.260)
  48. Todorovska MI, Lee VW (1989) Seismic waves in buildings with shear walls or central core. J Eng Mech ASCE 115(12):2669–2686
  49. Todorovska MI, Trifunac MD (1989) Antiplane earthquake waves in long structures. J Eng Mech ASCE 115(12):2687–2708
  50. Todorovska MI, Trifunac MD (1990) A note on the propagation of earthquake waves in buildings with soft first floor. J Eng Mech ASCE 116(4):892–900
  51. Todorovska MI, Trifunac MD (2005) Structural Health Monitoring by Detection of Abrupt Changes in Response Using Wavelets: Application to a 6-story RC Building Damaged by an Earthquake. Proc. of the 37th Joint Panel Meeting on Wind and Seismic Effects, 16–21 May, 2005. Tsukuba, Japan. US Japan Natural Resources Program (UJNR), pp 20
  52. Todorovska MI, Trifunac MD (2007) Earthquake damage detection in the Imperial County Services Building I: the data and time-frequency analysis. Soil Dyn Earthq Eng 27(6):564–576
  53. Todorovska MI, Trifunac MD (2008) Impulse response analysis of the Van Nuys 7-storey hotel during 11 earthquakes and earthquake damage detection. Struct Control Health Monit 15(1):90–116. [doi:10.1002/stc.208](https://doi.org/10.1002/stc.208)
  54. Todorovska MI, Trifunac MD (2008) Earthquake damage detection in the Imperial County Services Building III: analysis of wave travel times via impulse response functions. Soil Dyn Earthq Eng 21(5):387–404. [doi:10.1016/j.soildyn.2007.07.001](https://doi.org/10.1016/j.soildyn.2007.07.001)
  55. Todorovska MI, Trifunac MD (2008) Earthquake damage detection in the Imperial County Services Building II: analysis of novelties via wavelets. Struct Control Health Monit (submitted for publication)
  56. Todorovska MI, Trifunac MD, Ivanović SS (2001) Wave propagation in a seven-story reinforced concrete building, Part I: theoretical models. Soil Dyn Earthq Eng 21(3):211–223
  57. Todorovska MI, Trifunac MD, Ivanović SS (2001) Wave propagation in a seven-story reinforced concrete building, Part II: observed wave numbers. Soil Dyn Earthq Eng 21(3):225–236
  58. Trifunac MD (2007) Early History of the Response Spectrum Method, Dept. of Civil Engineering, Report CE 07-01. Univ. Southern California, Los Angeles, California
  59. Trifunac MD, Todorovska MI (2001) A note on the useable dynamic range of accelerographs recording translation. Soil Dyn Earthq Eng 21(4):275–286
  60. Trifunac MD, Todorovska MI (2001) Evolution of accelerographs, data processing, strong motion arrays and amplitude

- and spatial resolution in recording strong earthquake motion. *Soil Dyn Earthq Eng* 21(6):537–555
61. Trifunac MD, Todorovska MI (2001) Recording and interpreting earthquake response of full-scale structures: In Erdik M, Celebi M, Mihailov V, Apaydin N (eds) *Proc. of the NATO Advanced Research Workshop on Strong-Motion Instrumentation for Civil Engineering Structures*, 2–5 June, 1999. Kluwer, Istanbul, p 24
  62. Trifunac MD, Ivanovic SS, Todorovska MI (1999) Experimental evidence for flexibility of a building foundation supported by concrete friction piles. *Soil Dyn Earthq Eng* 18(3):169–187
  63. Trifunac MD, Todorovska MI, Hao TY (2001) Full-scale experimental studies of soil-structure interaction – a review. *Proc. of the 2nd US Japan Workshop on Soil-Structure Interaction*, 6–8 March, 2001. Tsukuba City, Japan, pp 52
  64. Trifunac MD, Ivanović SS, Todorovska MI (2003). Wave propagation in a seven-story reinforced concrete building, Part III: damage detection via changes in wave numbers. *Soil Dyn Earthq Eng* 23(1):65–75
  65. Trifunac MD, Todorovska MI, Manić MI, Bulajić BĐ (2008) Variability of the fixed-base and soil-structure system frequencies of a building – the case of Borik-2 building. *Struct Control Health Monit* (in press). doi:10.1002/stc.277
  66. Udawadia FE, Jerath N (1980) Time variations of structural properties during strong ground shaking. *J Eng Mech Div ASCE* 106(EM1):111–121
  67. Udawadia FE, Marmarelis PZ (1976) The identification of building structural systems I. The linear case. *Bull Seism Soc Am* 66(1):125–151
  68. Udawadia FE, Marmarelis PZ (1976) The identification of building structural systems II. The nonlinear case. *Bull Seism Soc Am* 66(1):153–171
  69. Udawadia FE, Trifunac MD (1974) Time and amplitude dependent response of structures. *Earthq Eng Struct Dyn* 2:359–378
  70. Ward HS, Crawford R (1966) Wind induced vibrations and building modes. *Bull Seism Soc Am* 56(4):793–813
  71. Wong HL, Trifunac MD, Luco JE (1988) A comparison of soil-structure interaction calculations with results of full-scale forced vibration tests. *Soil Dyn Earthq Eng* 7(1):22–31

## Earthquake Early Warning System in Southern Italy

ALDO ZOLLO<sup>1</sup>, GIOVANNI IANNACONE<sup>2</sup>,  
VINCENZO CONVERTITO<sup>2</sup>, LUCA ELIA<sup>2</sup>,  
IUNIO IERVOLINO<sup>3</sup>, MARIA LANCIERI<sup>2</sup>,  
ANTHONY LOMAX<sup>4</sup>, CLAUDIO MARTINO<sup>1</sup>,  
CLAUDIO SATRIANO<sup>1</sup>, EMANUEL WEBER<sup>2</sup>,  
PAOLO GASPARINI<sup>1</sup>

<sup>1</sup> Dipartimento di Scienze Fisiche, Università di Napoli “Federico II” (RISSC-Lab), Napoli, Italy

<sup>2</sup> Osservatorio Vesuviano, Istituto Nazionale di Geofisica e Vulcanologia (RISSC-Lab), Napoli, Italy

<sup>3</sup> Dipartimento di Ingegneria Strutturale, Università di Napoli “Federico II”, Napoli, Italy

<sup>4</sup> Alomax Scientific, Mouans-Sartoux, France

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Earthquake Potential and Seismic Risk in the Campania Region](#)

[Seismic Network Architecture and Components](#)

[Real-Time Data Transmission System](#)

[Network Management and Data Archiving](#)

[Real-Time Earthquake Location and Magnitude Estimation](#)

[Real-Time Hazard Analysis for Earthquake Early Warning](#)

[Future Directions](#)

[Bibliography](#)

### Glossary

**Data transmission system** A multi-component device aimed at the transmission of seismic signals over a distance, also denoted as a telecommunication system. Each data transmission system consists of two basic elements: a transmitter that takes information and converts it to an electromagnetic signal and a receiver that receives the signal and converts it back into usable information.

Modern telecommunication systems are two-way and a single device, a transceiver, acts as both a transmitter and receiver. Transmitted signals can either be analogue or digital. In an analogue signal, the signal is varied continuously with respect to the information. In a digital signal, the information is encoded as a set of discrete, binary values. During transmission, the in-

formation contained in analogue signals will be degraded by noise, while, unless the noise exceeds a certain threshold, the information contained in digital signals will remain intact. This represents a key advantage of digital signals over analogue signals. A collection of transmitters, receivers or transceivers that communicate with each other is a telecommunication network. Digital networks may consist of one or more routers that route data to the correct user.

### Earthquake early warning system (EEWS)

A real-time, modern information system that is able to provide rapid notification of the potential damaging effects of an impending earthquake, through rapid telemetry and processing of data from dense instrument arrays deployed in the source region of the event of concern (regional EEWS) or surrounding the target infrastructure (site-specific EEWS). A “regional” EEWS is based on a dense sensor network covering a portion or the entirety of an area that is threatened by earthquakes. The relevant source parameters (event location and magnitude) are estimated from the early portion of recorded signals and are used to predict, with a quantified confidence, a ground motion intensity measure at a distant site where a target structure of interest is located. On the other hand, a “site-specific” EEWS consists of a single sensor or an array of sensors deployed in the proximity of the target structure that is to be alerted, and whose measurements of amplitude and predominant period on the initial *P*-wave motion are used to predict the ensuing peak ground motion (mainly related to the arrival of *S* and surface waves) at the same site.

**Earthquake location** An earthquake location specifies the spatial position and time of occurrence for an earthquake. The location may refer to the earthquake hypocenter and corresponding origin time, a mean or centroid of some spatial or temporal characteristic of the earthquake, or another property of the earthquake that can be spatially and temporally localized.

**Earthquake magnitude** The magnitude is a parameter used by seismologists to quantify the earthquake size. The Richter magnitude scale, or more correctly, local magnitude *ML* scale, assigns a single number to quantify the amount of seismic energy released by an earthquake. It is a base-10 logarithmic scale obtained by calculating the logarithm of the combined horizontal amplitude of the largest displacement from zero on a seismometer output. Measurements have no limits and can be either positive or negative.

Introduced by the Japanese seismologist Aki in 1962, the seismic moment is the present-day physical pa-

parameter used to characterize the earthquake strength. It represents the scalar moment of one of the couples of forces producing the dislocation at an earthquake fault and it is measured from the asymptotic DC level on displacement Fourier spectra of recorded seismic signals.

**Probability density function – PDF** A function in one or more dimensional space  $\mathbf{X}$  that (i) when integrated over some interval  $\Delta x$  in  $\mathbf{X}$  gives a probability of occurrence of any event within  $\Delta x$ , and (ii) has unit integral over space  $\mathbf{X}$ , where  $\mathbf{X}$  represents a space of possible events.

**Seismic data-logger** A core element of a digital seismic station, whose aim is to record the analogue signals from seismic sensors and convert them in digital form with an assigned sampling frequency. Ground motion signals acquired by seismic sensors are pre-amplified and anti-aliasing filtered in a data-logger before they are digitalized through an AD (analog-to-digital) converter. The main technical features of a modern data-logger are the number of available channels, the allowed sampling frequencies, the dynamic range, the digitizer clock type, the storage capacity (PCMCIA, internal flash and/or hard disk, USB, ...), network interfaces (ethernet, wireless lan, or ppp) and power consumption.

**Seismic hazard** The probability that at a given site, a strong motion parameter (generally the peak ground acceleration) exceeds an assigned value in a fixed time period. When the seismic hazard is computed for an extended region it is generally represented as a map. The hazard map is commonly computed for a constant probability level (10%, 5% or 2%) and a given time window (50 years). It represents the spatial variation of the peak ground acceleration (expressed in percentage of gravity  $g$ ) to be exceeded in the given period with the chosen probability level.

Earthquake early warning systems can provide a mean for the evaluation of real-time hazard maps which evolve with time, as new information about source location, magnitude and predicted peak ground motion parameters are available soon after the earthquake occurrence.

**Seismic sensors** Instruments used to record the ground vibration produced by natural and artificial sources, generally denoted as seismometers. A seismometer measures the relative motion between its frame and a suspended mass. Early seismometers used optics, or motion-amplifying mechanical linkages. The motion was recorded as scratches on smoked glass, or exposures of light beams on photographic paper. In modern

instruments the proof mass is held motionless by an electronic negative feedback loop that drives a coil. The distance moved, speed and acceleration of the mass are directly measured. Most modern seismometers are broadband, working on a wide range of frequencies (0.01–100 Hz). Another type of seismometer is a digital strong-motion seismometer, or accelerometer, which measures soil acceleration. Due to its relatively high dynamic range, the accelerometer can record unsaturated strong amplitude signals at close distances from a large earthquake. This data is essential to understand how an earthquake affects human structures.

### Definition of the Subject

The origin of the term “early warning” probably goes back to the first decades of the last century. However, the first practical use of an “early warning” strategy was military and it was developed during the “cold war” years as a countermeasure to the potential threat from inter-continental ballistic missiles. The objective of these systems was to give an alert to target areas as soon as a missile was detected by a radar system or a launch was detected by a satellite system. In this context the term “lead time” was defined as the time elapsing between the detection of the missile and the estimated impact on the target.

In the last decades the use of the term “early warning” greatly expanded. It is used with small, but significant, variations in various types of risks, from epidemiological, to economic, social, and of course all the types of natural and environmental risks.

In fact, in these contexts, including some natural risks such as hydro-geological and volcanic, the warning is not given at the onset of the catastrophic phenomenon, but after the occurrence of some precursory phenomena which can trigger a catastrophic event (for instance intensive rainfall for hydrological risk, earthquakes and/or ground deformation for volcanic risk). The main consequence of this difference is an increase in the probability of issuing false alarms.

The case of earthquake early warning is similar to missile early warning. The alert is given after an earthquake is detected by a network of seismometers. An earthquake early warning is based on the fact that most of the radiated energy is contained in the slower traveling phases (S- and surface waves traveling at about 3.5 km/s or less) which arrive at any location with a delay with respect to small amplitude higher velocity phases (P-waves, travelling at about 6–7 km/s) or to an electromagnetically transmitted (EM) signal giving the warning.

## Introduction

Many regions in the world are affected by natural hazards such as earthquakes, tsunamis, volcanoes, floods, storms, landslides, etc., each of which can have devastating socio-economic impacts. Among these natural events, earthquakes, have been among the most recurrent and damaging hazards during last few decades, resulting in large numbers of casualties, and massive economic losses [30].

The problem of earthquake risk mitigation is faced using different approaches, depending upon the time scale being considered. Whilst over time scales of decades it is of utmost importance that land use regulations and building/infrastructure codes are continuously updated and improved, for time scales of a few years, the main risk mitigation actions are at the level of information and education in order to increase individual and social community awareness about potentially damaging hazards. Over shorter time scales (months to hours), it would naturally be of great benefit to society as a whole if the capability to accurately predict the time, location and size of a potentially catastrophic natural event were available. However, due to the great complexity of the natural processes of concern, such predictions are currently not possible.

On the other hand, on very short time scales (seconds to minutes), new strategies for earthquake risk mitigation are being conceived and are under development worldwide, based on real-time information about natural events that is provided by advanced monitoring infrastructures, denoted as “early warning systems”.

## Regional and On-site Early Warning Systems

Earthquake Early Warning Systems (EWS) are modern, real-time information systems that are able to provide rapid notification of the potential damaging effects of an impending earthquake through the rapid telemetry and processing of data from dense instrument arrays deployed in the source region of the event of concern. Such systems allow mitigating actions to be taken before strong shaking and can significantly shorten the time necessary for emergency response and the recovery of critical facilities such as roads and communication lines.

Advances have been made towards the implementation of operational systems in Japan, Taiwan, and Mexico using two different approaches, i. e., “regional warning” and “onsite warning” [25]. A regional warning system is based on a dense sensor network covering a portion or the entire area that is threatened by earthquakes. The relevant source parameters (earthquake location and magnitude) are estimated from the early portion of recorded signals and are used to predict, with a quantified confidence,

a ground motion intensity measure at a distant site where a target structure of interest is located. Alternatively, “onsite warning” systems consist of a single sensor or an array of sensors deployed in the proximity of the target structure that is to be alerted, and whose measurements on the initial *P*-wave motion are used to predict the ensuing peak ground motion (mainly related to the arrival of *S* and surface waves) at the same site.

## Implementation of Early Warning Systems Worldwide

In Japan, since the 1965, the JNR (Japanese National Railway) has developed and operated the Urgent Earthquake Detection and Alarm System (UrEDAS), which is an onsite warning system along the Shinkansen (bullet train) railway. UrEDAS is based on seismic stations deployed along the Japanese Railway with an average distance of 20 km. An alert is issued if the horizontal ground acceleration exceeds 40 cm/s<sup>2</sup>. In the 1996, the UrEDAS was combined with a new seismometer called “compact UrEDAS” [31,32,33].

On the other hand, for about one decade the Japanese Meteorological Agency (JMA) has been developing and experimenting with a mixed single station and network based early warning system to generate immediate alerts after earthquakes with JMA Intensity greater than “lower 5” (approximately  $M > 6$ ) [24]. During a testing period from February 2004 to July 2006, the JMA sent out 855 earthquake early warnings, only 26 of which were recognized as false alarms [40]. On October 1, 2007 the broadcast early warning system developed by the Japanese Meteorological Agency (JMA) became operative. In this system, the first warning is issued 2 s after the first *P* phase detection, if the maximum acceleration amplitude exceeds the threshold of 100 cm/s<sup>2</sup>.

In the United States the first prototype of an early warning system was proposed by Bakun et al. [4] and developed for mitigating earthquake effects in California. It was designed to rapidly detect the Loma Prieta aftershocks and send an alert when the estimated magnitude was greater than 3.7, in order to reduce the risk of the crews working in the damaged area. The system is composed of four components: ground motion sensors deployed in the epicentral area, a central receiver, radio repeaters and radio receivers. The prototypical system worked for 6 months, during which time 19 events with  $M > 3.5$  occurred, 12 alerts were issued with only 2 missed triggers and 1 false alarm.

Based on pioneering work by Allen and Kanamori [2] seismologists across California are currently planning real-time testing of earthquake early warning across the



state using the ElarmS (Earthquake Alarms Systems) methodology [1]. The approach uses a network of seismic instruments to detect the first-arriving energy at the surface, the *P*-waves, and translate the information contained in these low amplitude waves into a prediction of the peak ground shaking that follows. Wurman et al. [47] illustrated the first implementation of ElarmS in an automated, non-interactive setting, and the results of 8 months of non-interactive operation in northern California.

Since 1989, in Mexico, the civil association CIRES (Centro de Instrumentacion y REgistro Sismico) with the support of Mexico City Government Authorities, developed and implemented the Mexican Seismic Alert System (SAS) [15]. The SAS is composed of (a) a seismic detection network, 12 digital strong motion stations deployed along 300 km of the Guerrero coast, (b) a dual communication system: a VHF central radio relay station and three UHF radio relay stations, (c) a central control system which continuously controls the operational status of the seismic detection and communication system and, when an event is detected, automatically determines the magnitude and issues the alarm, and (d) a radio warning system for broadcast dissemination of the alarm to end users. After 11 years, the SAS system recorded 1373 events in the Guerrero coast, it issued 12 alerts in Mexico city, with only one false alarm.

In Taiwan, the Taiwan Central Weather Bureau (CWB) developed an early warning system based on a seismic network consisting of 79 strong motion stations installed across Taiwan and covering an area of  $100 \times 300 \text{ km}^2$  [44]. Since 1995 the network has been able to report event information (location, size, strong motion map) within 1 min after an earthquake occurrence [39]. To reduce the report time, Wu and Teng [44] introduced the concept of a virtual sub-network: as soon as an event is triggered by at least seven stations, the signals coming from the stations less distant than 60 km from the estimated epicenter are used to characterize the event. This system successfully characterized all the 54 events occurred during a test period of 7 months (December 2000 – June 2001), with an average reporting time of 22 s.

In Europe, the development and testing of EEWS is being carried out in several active seismic regions. Europe is covered by numerous high-quality seismic networks, managed by national and European agencies, including some local networks specifically designed for seismic early warning around, for example, Bucharest, Cairo, Istanbul and Naples.

In Turkey, an EEWS is operative, called PreSEIS (pre-seismic shaking), to provide rapid alert for Istanbul and surrounding areas. It consists of 10 strong motion sta-

tions located along the border of the Marmara sea along an arc of about 100 km, close to the seismogenetic zone of the Great Marmara Fault Zone with real time data transmission to Kandilli-Observatory [7,14]. An alarm is issued when a threshold amplitude level is exceeded.

In Romania, the EEWS is based on three tri-axial strong motion sensors deployed in the Vrancea area with a satellite communication link to the Romanian Data Center at NIEP in Bucharest [7,42]. The system is based on first *P* wave detection and prediction of the peak horizontal acceleration recorded in Bucharest, allowing for a warning time of about 25 s.

On 2006 the European Union launched the 3-year project SAFER (Seismic Early Warning for Europe), which is a cooperative scientific program aimed at developing technological and methodological tools that exploit the possibilities offered by real-time analysis of signals coming from these networks for a wide range of actions, performed over time intervals of a few seconds to some tens of minutes. The project includes the participation of 23 research groups from several countries of Europe. The primary aim of SAFER is to develop tools that can be used by disaster management authorities for effective earthquake early warning in Europe and, in particular, its densely populated cities.

### **The Development of an Early Warning System in Campania Region, Southern Italy**

The present article is focused on the description of technologies and methodologies developed for the EEWS under construction in southern Italy.

With about 6 million inhabitants, and a large number of industrial plants, the Campania region (southern Italy), is a zone of high seismic risk, due to a moderate to large magnitude earthquake on active fault systems in the Apenninic belt. The 1980,  $M = 6.9$  Irpinia earthquake, the most recent destructive earthquake to occur in the region, caused more than 3000 casualties and major, widespread damage to buildings and infrastructure throughout the region.

In the framework of an ongoing project financed by the Regional Department of Civil Protection, a prototype system for seismic early and post-event warning is being developed and tested, based on a dense, wide dynamic seismic network under installation in the Apenninic belt region (ISNet, Irpinia Seismic Network).

Considering an earthquake warning window ranging from tens of seconds before to hundred of seconds after an earthquake, many public infrastructures and buildings of strategic relevance (hospitals, gas pipelines, railways,

railroads, ...) in the Campania region can be considered as potential EEWS target-sites for experimenting with innovative technologies for data acquisition, processing and transmission based on ISNet. The expected time delay to these targets for the first energetic *S* wave train is around 30 s at about 100 km from a crustal earthquake occurring in the source region. The latter is the typical time window available for mitigating earthquake effects through early warning in the city of Naples (about 2 million inhabitants, including suburbs).

This article illustrates the system architecture and operating principles of the EEWS in the Campania region, focusing on its innovative technological and methodological aspects. These are relevant for a reliable real-time estimation of earthquake location and magnitude which are used to predict, with quantified confidence, ground motion intensity at a distant target site.

The system that we describe in this article uses an integrated approach from real time determination of source parameters to estimation of expected losses.

This problem must be dealt in an evolutionary (i. e., time-dependent) and probabilistic framework where probability density functions (PDFs) for earthquake location, magnitude and attenuation parameters are combined to perform a real-time probabilistic seismic hazard analysis.

### Earthquake Potential and Seismic Risk in the Campania Region

The southern Apennines are an active tectonic region of Italy that accommodates the differential motions between the Adria and Tyrrhenian microplates [23]. The majority of the seismicity in this region can be ascribed to this motion. These earthquakes mainly occur in a narrow belt along the Apennine chain and are associated with young faults, with lengths ranging from 30 to 50 km, and mainly confined to the upper 20 km of the crust [28,41].

Recent stress and seismic data analyzed by [29] using earthquake locations and fault mechanisms show that the southern Apennines are characterized by an extensional stress regime and normal-fault earthquakes. However, the occurrence of recent (e. g., 5 May, 1990, Potenza,  $M$  5.4; 31 October – 1 November, 2002, Molise,  $M$  5.4) and historic (e. g., 5 December, 1456,  $M$  6.5) earthquakes do not exclude other mechanisms such as strike-slip faulting.

There have been numerous large and disastrous events in the southern Apennines, including those which occurred in 1694, 1851, 1857 and 1930. The location of historical earthquakes retrieved from the CFTI (Catalogo dei Forti Terremoti in Italia, Catalogue of Strong Earthquakes

in Italy) database [6] is shown in Fig. 1. The most recent and well documented event is the complex normal-faulting  $M$  6.9 Irpinia earthquake of 23 November, 1980 [5,43].

As recently indicated in the study by Cinti et al. [9], the southern Apennines has a high earthquake potential with an increasing probability of occurrence for  $M \geq 5.5$  earthquakes in the next decade. The new national hazard map (Gruppo di lavoro MPS, 2004), indicates that the main towns of the region fall in a high seismic hazard area, where it is expected that a peak ground acceleration value ranging between 0.15 and 0.25 g will be exceeded in 475 years.

These aspects make the Campania region a suitable experimental site for the implementation and testing of an early warning system. A potential application of an early warning system in the Campania region should consider an expected time delay to the first energetic *S* wave train varying between 14–20 s at 40–60 km distance to 26–30 s at about 80–100 km, from a crustal earthquake occurring along the Apenninic fault system. Based on those delay times, a large number of civil and strategic infrastructures located in the Campania region are eligible for early warning applications, as shown in Fig. 2.

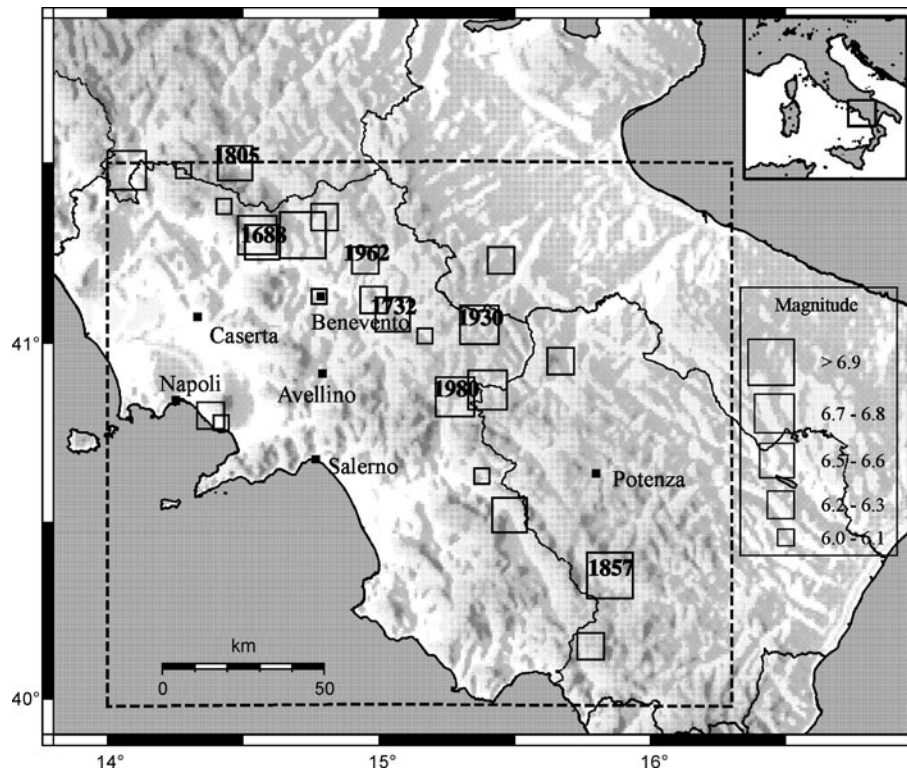
### Seismic Network Architecture and Components

The Irpinia Seismic Network (ISNet) is a local network of strong motion, short period and broadband seismic stations deployed along the southern Apenninic chain covering the seismogenic areas of the main earthquakes that occurred in the region in the last centuries, including the  $M_s = 6.9$ , 23 November 1980 event.

The seismic network is composed of 29 stations organized in six sub-nets, each of them composed of a maximum of 6–7 stations (Fig. 3). The stations of a given sub-net are connected with real-time communications to a central data-collector site called the Local Control Center (LCC).

The different LCCs are linked to each other and to a Network Control Center (NCC) with different types of transmission systems. The whole data transmission system is fully digital over TCP/IP, from the data-loggers, through the LCC, to the NCC, located in the city of Naples, 100 km away from the network center.

To ensure a high dynamic recording range, each seismic station is equipped with a strong-motion accelerometer and a three-component velocity meter (natural period = 1 s). In five station locations the seismometers are replaced by broadband (0.025–50 Hz) sensors to guarantee good-quality recording of teleseismic events. Data acquisition at the seismic stations is performed by the inno-



Earthquake Early Warning System in Southern Italy, Figure 1

Location of the main historic earthquakes retrieved from the CFTI database using as region of interest that defined by the external rectangle. The box dimensions are proportional to magnitude. The best constrained historic earthquakes are reported along with their date of occurrence

vative data-logger Osiris-6, produced by Agecodagis sarl. The hardware/software characteristics of the system allow it to install self-developed routines to perform real time specific analysis.

The data-loggers are remotely controlled through a configuration tool accessible via TCP/IP, managing sampling rate, gain, application of calibration signal to the resets of disks, GPS, etc. Furthermore, a complete station health status is available, which helps in the diagnosis of component failure or data-logger malfunction. The data-loggers store the data locally or send it to each LCC where the real-time data management system Earthworm (developed at USGS-United State Geological Survey) is operating.

A calibration unit is installed at each seismic station to automatically provide a periodic calibration signal to seismic sensors in order to verify the correct response curve of the overall acquisition chain.

The power supply of the seismic station is provided by two solar panels (120 W peak, with 480 Wh/day), two 130 Ah gel cell batteries, and a custom switching circuit board between the batteries. With this configura-

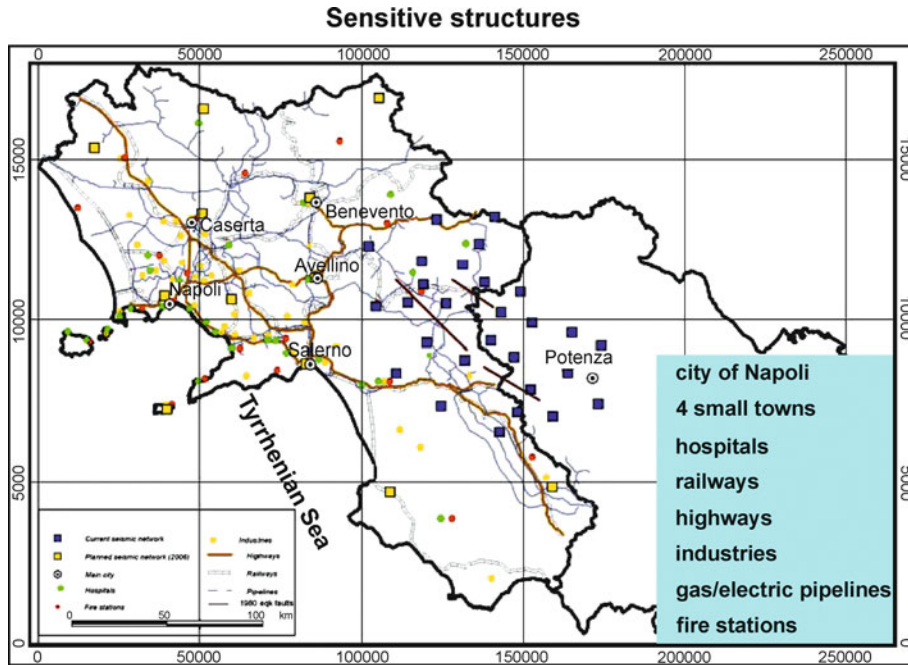
tion, 72-h autonomy is ensured for the seismic and radio communication equipment. Each site is also equipped with a GSM/GPRS programmable control/alarm system connected to several environmental sensors and through which the site status is known in real time. With SMS (Short Message Service) and through the programmable GSM controller, the seismic equipment can be completely reset remotely with a power shutdown/restart. The GSM also controls the device start/stop release procedure when the battery goes over/under a predefined voltage level.

Unlike the seismic stations, LCCs, which host the data server and transmission system instruments, are AC power supplied with back-up gel batteries guaranteeing 72-h stand-by power.

### Real-Time Data Transmission System

ISNet has a distributed star topology that uses different types of data transmission systems.

The seismic stations are connected via spread-spectrum radio bridges to the LCCs. Data transmission between LCCs from the local control center to the network



Earthquake Early Warning System in Southern Italy, Figure 2  
 Distribution of the sensitive structures, potential candidates for an early warning system in the Campania-Lucania region



Earthquake Early Warning System in Southern Italy, Figure 3  
 Topology of the communication system of ISNet showing the extended-star configuration of the seismic network. Symbols explanation: green squares – seismic stations; blue squares – Local Control Centres (LCC); yellow lines – WLAN radio linkconnecting seismic stations and LCC; white segments – SDH carrier-class radio; red triangles – radio link repeaters; red circle – Network Control Centre RISS in Naples; yellow squares – main cities

control center in Naples is performed through different technologies and media types as shown in Table 1.

To transmit waveforms in real time from the seismic stations to the LCCs, a pair of outdoor Wireless LAN

bridges operating in the 2.4 GHz ISM band are used. Our tests have shown that these instruments operate continuously without any radio link failure due to adverse weather conditions (snow, heavy rain).

Earthquake Early Warning System in Southern Italy, Table 1  
Specification of the ISNet data communication links

Type	Frequency (GHz)	Bandwidth (Mbps)	# Number of		Comments
			Stations	LCCs	
Spread spectrum Radio	2.45	54	27 <sup>3</sup>	–	Throughput around 20–24 Mbps for links between 10–15 km (based on ethernet packets with an average size of 512 bytes).
Ethernet	–	100	2 <sup>3</sup>		Stations connected with ethernet cable to LCC infrastructure.
Wireline SHDSL over Frame Relay	–	2.048	–	2	At the central site (RISSC) the CIR <sup>1</sup> is maximum 1.6 Mbps depending upon number of PVCs <sup>2</sup> . At the remote (LCC) site the bandwidth is 640/256 kbps with CIR of 64 kbps in up and download, over ADSL with ATM ABR service class.
Microwave Radio SDH	7	155	–	6	Carrier-class microwave link. Connect six LCC with 155 Mbps (STM-1) truly full bandwidth available. First link constructed for early warning applications.
Microwave Radio HyperLAN/2	5.7	54	–	2	The true usable maximum throughput of HyperLAN/2 is 42 Mbps.

<sup>1</sup> CIR Committed Information Rate.

<sup>2</sup> PVC permanent virtual circuit.

<sup>3</sup> Not included stations hosted by LCCs.

The two primary backbone data communication systems of the central site use Symmetrical High-speed Digital Subscriber Line (SHDSL) technology over a frame-relay protocol. Frame relay offers a number of significant benefits over analogue and digital point-to-point leased lines. With the latter, each LCC requires a dedicated circuit between the LCCs and NCC. Instead, the SHDSL frame relay is a packet-switched network, which allows a site to use a single frame-relay phone circuit to communicate with multiple remote sites through the use of permanent virtual circuits. With virtual circuits, each remote site is seen as part of a single private LAN, simplifying IP address scheme maintenance and station monitoring.

Each seismic site has a real-time data flow of 18.0 kbps (at 125 Hz sampling rate for each physical channel), and the overall data communication bandwidth that is needed is around 540 kbps for 30 stations. ISNet supports this throughput under the worst conditions seen and it has been designed to guarantee further developments, such as the addition of further seismic or environmental sensors, without the need for larger economic and technological investment.

## Network Management and Data Archiving

### The Network Manager Application and Implementation Overview

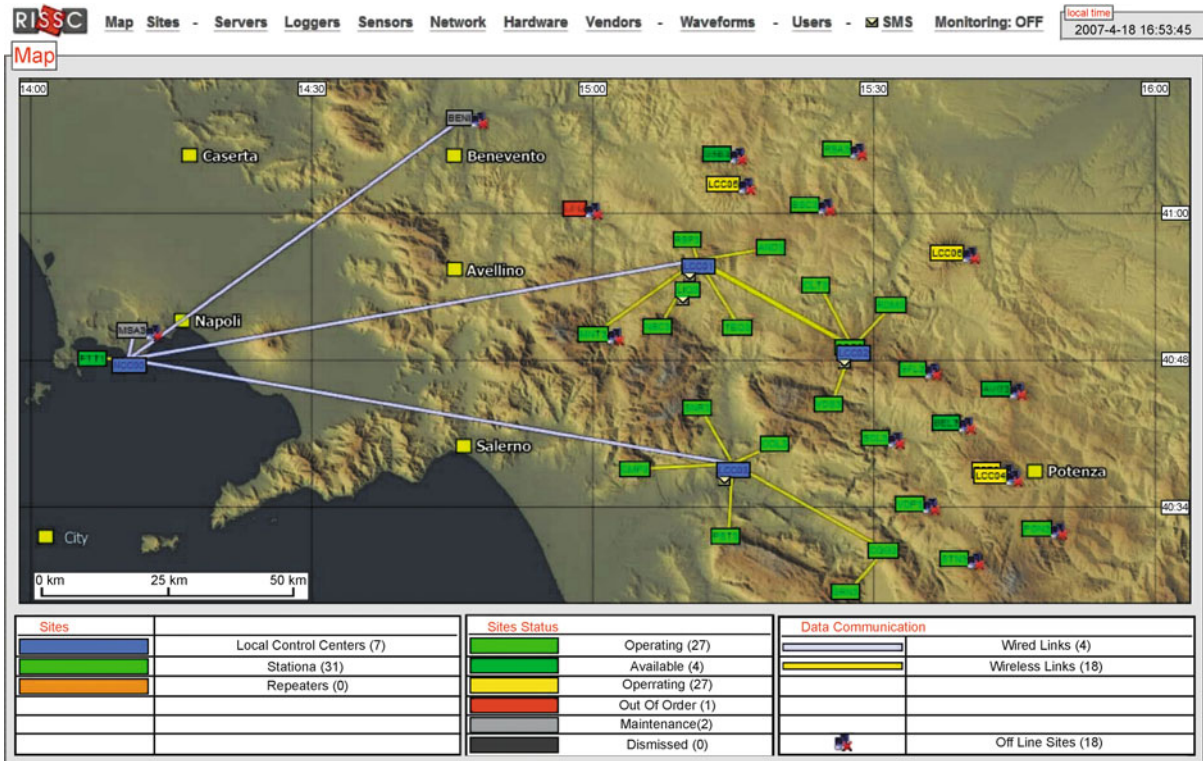
As seen in the previous paragraphs, ISNet is a complex infrastructure, and thus needs a suitable software application

in order to be effectively managed: a front-end to users and administrators with an interface that is simple to use. To this aim we developed a server-client database-driven application, dubbed *SeismNet Manager*, to keep track of the several components that comprise or are produced by the network, such as stations, devices and recorded data. This application, whose front page is shown in Fig. 4, lets the administrators manage (insert, edit, view and search) the details of (a) seismic stations and Local Control Centers (sites), (b) data communication links between sites (wired or wireless), instruments and devices (sensors, loggers, network hardware), and (c) recorded and computed data (waveforms, events). *SeismNet Manager* also keeps an historical record of the installations and configurations of the above elements.

All of the mentioned components are handled by leveraging an instrumental database, a flexible repository of information that was implemented by using PostgreSQL, a robust and feature-rich Database Management System available as open source.

### The Instrumental Database

The instrumental database is a web-oriented application tool where, at the top level, the network is modeled as a set of sites, with installed loggers, sensors, data acquisition servers, network hardware and generic hardware, in a given configuration. Each of the mentioned entities is mirrored by a different class of objects in the database,



Earthquake Early Warning System in Southern Italy, Figure 4

The front page of SeismNet Devices Manager. This page is meant to convey the state of the whole network at a glance. Each node (station or LCCs) is shown along with its operating state, data links of different types to nearby nodes, whether it's currently on-line or not, along with eventual alarms still pending

where the relevant details are stored and then presented to the users as interactive web pages. As an example, see the page for a typical seismic station in Fig. 5.

The instrumental database was implemented with a layer of abstraction that lets one easily perform complex queries and hides the actual implementation details of the underlying structure to a possible client. There are both *stored procedures*, i. e., functions that perform complex tasks given simple inputs, and *views*, i. e., virtual database tables that collect the most important pieces of information about an object, physically scattered in many tables, in a single place and make it possible to easily query, for example, for all the details of the correctly operating sensors installed one year ago at stations with a working wireless link to a given LCC server.

This abstract interface makes the devices database a central repository for effectively cross-correlating the seismic data recorded at any given place and time with the details of the instrument(s) that recorded them, and the configuration details of the systems that ultimately made them available. The interface approach also makes

it easier to change the implementation details without the need to update the web application, or any external client procedures that need to interact with the instrumental database.

### Automatic Monitoring of the Devices and Automatic Data Retrieval

All of the details about the network described so far are provided by the administrators of the system and are manually updated every time the configuration of something in the network changes, e. g., after installing a new sensor or replacing some faulty hardware at a station. This manual input is needed for “dumb” devices, such as sensors. “Smart” devices, i. e., computers with an IP address (loggers, bridges and Earthworm servers), on the other hand, can be queried about their actual configuration from time to time. The web application can plot the temporal evolution of some selected parameters as graphs, spanning a period ranging from hours to years (Fig. 6). This is useful to correlate issues spotted on the recorded seis-

Status	
State	Operating
Visibility	Private
Online	<input checked="" type="checkbox"/>

Description	
Code (4+ char.)	SNR3
Network	ISNET
Type	Station
Comment	

Location	
Extended Location Name	Senerchia
Longitude E (deg.)	15.1925
Latitude N (deg.)	40.7361
Elevation (m)	998

GSM terminal	
<input checked="" type="checkbox"/> SIM telephone number	+39 3358028209 <a href="#">View SMS</a>

History	
Begin Date (yyyy-mm-dd)	2005-11-11
End Date (yyyy-mm-dd)	

Data Links				
Destination	Location Name	Distance (km)	Technology	Down/Up (Mb/s)
LCC03	Contursi Terme	9.9	Wi-Fi 802.11g	54

Loggers				
Model	IP address	Serial number	Storage medium	Recording type
OSIRIS6	10.37.37.20	370020	CompactFlash	Continuous

Sensors				
Model	Type	Serial number	Connected to	Components
CMG-5T	Accelerometer	T5744	OSIRIS6 370020 channels 0.1.2	TripleComponent
S13J	Velocimeter	V397	OSIRIS6 370020 channel 3	Vertical
S13J	Velocimeter	H490	OSIRIS6 370020 channel 4	NorthSouth
S13J	Velocimeter	H505	OSIRIS6 370020 channel 5	EastWest


[Components History](#)

Network Hardware				
Model	IP address	Type	Serial number	
AIR-BR1310G-A-K9-R	192.168.3.37	BRIDGE WIRELESS	FTX0905U0DE	

**Photo**



**SAR**

upload/site39filename\_sar.doc

[Satellite Map](#)   [Roads Map](#)

Waveforms		
Date	Mag	Files
2007-03-15	1.1 MI	6
2007-03-14	2.5 MI	12
2007-03-11	1.8 MI	6
2007-03-08	0.8 MI	6
2007-03-07	3.1 MI	18
2007-03-05	2.1 MI	12
2007-03-03	0.8 MI	6
2007-03-02	2 MI	6
2007-02-28	2.6 MI	6
2007-02-25	3.1 MI	12

740 more file(s)...

**Notes & Files**

2007-01-12

In data 11-01-06 e' stato sostituito il cavo S13J

---

2006-11-10

Sostituito cavo S-13J quello attuale permette la calibrazione da remoto e ha resistenza di smo ...

---

2006-10-09

Installato S13J proviene da USA. Bloccati i cavi di S13J e GURALP. La stazione osiris è st ...

---

2006-09-26

Logger OSIRIS Accelerometer: Guralp  
CMG-5T Velocimeter: Geotech S13J

Earthquake Early Warning System in Southern Italy, Figure 5

The page relative to a seismic station. This page is a collection of all the pieces of information linked to a particular site: location details and map; some pictures and notes; recently received warning messages; currently installed devices and their configurations and mutual connections; data links to other stations; most recent waveforms recorded. Every device at a site also has an associated installation object, that records the configuration parameters and the physical connections to other nearby devices, valid over a period of time. Some elements, such as the data storage servers and the loggers, also need some further configuration parameters, that are independent of their actual physical installation, for things like firmware release and versions of the software packages run

mic data (typically, "holes" in the stream of data) to hardware problems. It is possible to inspect the whole chain of data transfers to pinpoint the source of the problem (e. g., low batteries on a logger due to a faulty inverter, low signal of a wireless connection due to harsh weather conditions).

There are both automatic and manual procedures to insert new events and data files in the system. The automatic procedures make use of several sources of events to process, such as: INGV (Istituto Nazionale di Geofisica e Vulcanologia, Italy) bi-weekly bulletins; INGV real time alerts; our early warning system. Likewise, they exploit several sources of recorded seismic data in order to provide a SAC file (SAC – seismic analysis code from

Lawrence Livermore National Laboratory) spanning the period from just before the arrival time, up to the end of the event.

Sensor data are retrieved from: (1) a repository of files from the internal mass storage of the loggers; (2) a local Earthworm Wave Server that caches older data collected from all the LCCs; (3) the most recent real time recording from the remote Wave Servers. The instrumental database is used to determine which sites/sensors/configurations recorded each event and to fill the headers of the files using the standard SAC format. The waveforms and events database, on the other end, is used by the automatic procedures to know which pieces of data are still missing for already recorded events (due to e. g., the temporary un-



Earthquake Early Warning System in Southern Italy, Figure 6

Health graphs of the devices. A device can be marked for monitoring and its internal state, or “health”, gets polled at regular intervals. Several of its internal variables are then retrieved and stored into the database, and their temporal evolution can be plotted as a graph. In this case both the internal temperature and CPU load of an OSISRIS data logger are shown, over periods ranging from one hour to one year

availability of one or more seismic data sources) and need to be collected.

**The Waveforms and Events Database: Searching and Visualizing the Seismic Data**

We also built a waveform and event database, the natural complement to the instrumental database. It keeps track of the events detected by the network and the relative waveforms recorded by the sensors. This database stores objects for events, origin estimations (time and location), magnitude estimations and waveforms. Several origins can be attached to a single event, as different algorithms and different institutions provide different estimations. Likewise, several magnitude types and estimations are attached to each origin. A waveform object for each sensor that recorded the earthquake is also linked to the event object, and stores a pointer to a SAC file, and its source (site and channel). The latter records are then used to gather, from the instrumental database, the actual details of the instruments that recorded the data.

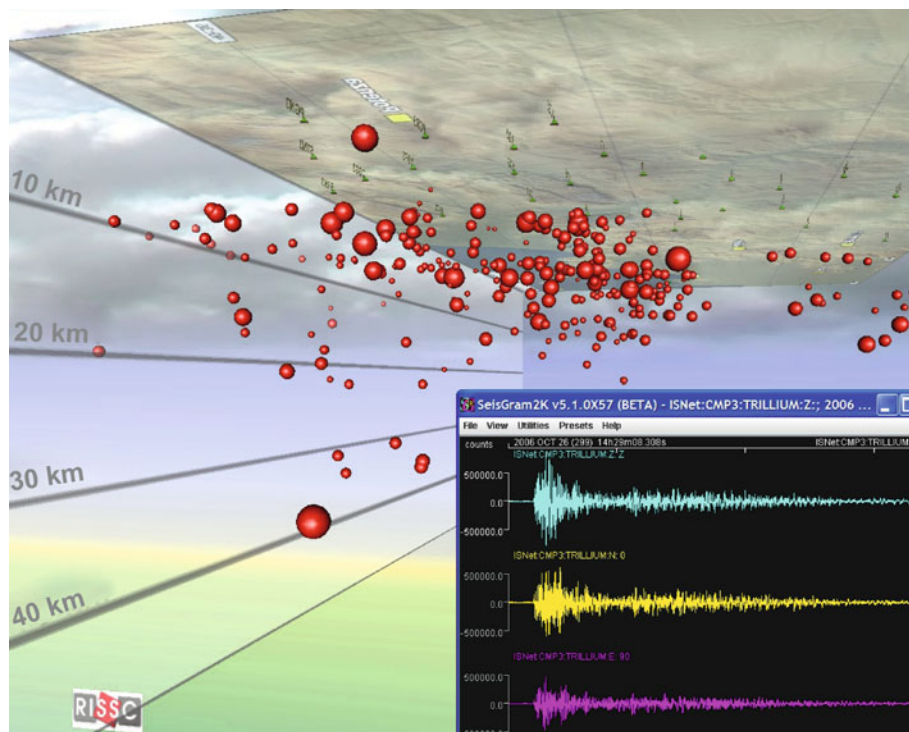
An interface for searching both events and waveforms is provided, as pictured in Fig. 7. Events can be filtered on origin time and location, magnitude, and distance to the stations. Waveforms can be filtered on station, component, instrument and quality.

**Real-Time Earthquake Location and Magnitude Estimation**

**Real-Time Earthquake Location**

**Previous Related Studies** There are many methodologies for standard earthquake location, performed when most or all the phase arrival times for an event are available. Standard analysis techniques are generally not suited for early warning applications, since they typically need the seismic event to be fully recorded at several stations, leaving little or no lead time for the warning [25]. For this reason, a different strategy is required, where the computation starts when a few seconds of data and a small number of recording stations are available, and the results are updated with time.





Earthquake Early Warning System in Southern Italy, Figure 7

Visualization of the seismic data. This is the graphical presentation of data recorded by ISNet. The waveforms matching the user's search criteria can be viewed on-line via Seisgram2K (where they can also be processed), while the events are rendered via VRML as a fully interactive 3D scene in the browser itself

Previous work on earthquake location for early warning includes several approaches to gain constraints on the location at an earlier time and with fewer observations than for standard earthquake location.

In the ElarmS methodology [47], when the first station triggers, the event is temporarily located beneath that station; after a second station trigger the location moves to a point between the two stations, based on the timing of the arrivals; with three or more triggered arrivals, the event location and origin time is estimated using trilateration and a grid search algorithm.

Horiuchi et al. [18] combine standard L2-norm event location, equal differential-time (EDT) location on quasi-hyperbolic surfaces, and the information from not-yet arrived data to constrain the event location beginning when there are triggered arrivals from two stations. The two arrival times define a hyperbolic surface, which contains the event location. This solution is further constrained by EDT surfaces constructed using the current time ( $t_{\text{now}}$ ) as a substitute for future, unknown arrival times at the stations, which have not yet recorded arrivals. The constraint increases as  $t_{\text{now}}$  progresses, even if no further stations record an arrival.

Rydelek and Pujol [36], applying the approach of Horiuchi et al. [18], show that useful constraints on an event location can be obtained with only two triggered stations. Cua and Heaton [12], generalized the approach by Rydelek and Pujol in order to start the location with one single triggering station.

The real-time location technique described in this paper is based on the equal differential-time (EDT) formulation [16,27] for standard earthquake location. The EDT location is given by the point traversed by the maximum number of quasi-hyperbolic surfaces, on each of which the difference in calculated travel-time to a pair of stations is equal to the difference in observed arrival times for the two stations. The EDT location determination is independent of origin time and reduces to a 3D search over latitude, longitude and depth. Furthermore, EDT is highly robust in the presence of outliers in the data [27]. This robustness is critical for the problem of earthquake location for seismic early warning, since we will often work with small numbers of data and may have outlier data such as false triggers, picks from other events, and misidentified picks from energetic, secondary phases.

Assuming that a dense seismic network is deployed around the fault zone, we define as the “*evolutionary approach*” a type of analysis where the estimates of earthquake location and size, and their associated uncertainty, evolve with time as a function of the number of recording stations and of the length of the portion of signal recorded at each station.

A direct implication of the evolutionary strategy is that each algorithm must be capable of *real-time* operation, i. e., its computational time must be smaller than the rate at which data enters the system.

Furthermore, since each algorithm starts processing a limited amount of information, the estimated earthquake parameter must be provided, at each time step, as a *probability density function* (PDF) which incorporates in its definition the uncertainties related both to the model employed and to the available data.

**The Real-Time Earthquake Location Method** The methodology is related to that of Horiuchi et al. [18], which has been extended and generalized by (a) starting the location procedure after only one station has triggered, (b) using the equal differential-time approach proposed by Font [16] to incorporate the triggered arrivals and the not-yet-triggered stations, (c) estimating the hypocenter probabilistically as a PDF instead of as a point, and (d) applying a full, non-linearized, global-search for each update of the location estimate.

We assume that a seismic network has known sets of operational and non-operational stations (Fig. 8a), that when an earthquake occurs, triggers (first *P*-wave arrival picks) will become available from some of the operational stations, and that there may be outlier triggers which are not due to *P* arrivals from the earthquake of interest.

Let's denote the operational stations as  $(S_0, \dots, S_N)$ , and consider a gridded search volume  $V$  containing the network and target earthquake source regions, and the travel times from each station to each grid point  $(i, j, k)$  in  $V$  computed for a given velocity model.

The standard EDT approach states that, if the hypocenter  $(i, j, k)$  is exactly determined, then the difference between the observed arrival times  $t_n$  and  $t_m$  at two stations  $S_n$  and  $S_m$  is equal to the difference between calculated travel times  $tt_n$  and  $tt_m$  at the hypocentral position, since the observed arrival times share the common earthquake origin time. In other words, the hypocenter must satisfy the equality:

$$(tt_m - tt_n)_{i,j,k} = t_m - t_n; \quad m \neq n \quad (1)$$

for each pair of triggering stations  $S_n$  and  $S_m$ . For a constant velocity model, this equation defines a 3D hyperbolic

surface whose symmetry axis goes through the two stations. Given  $N$  triggering stations,  $N(N-1)/2$  surfaces can be drawn; the hypocenter is defined as the point crossed by the maximum number of EDT surfaces.

Following an evolutionary approach, the method evaluates, at each time step, the EDT equations considering not only each pair of triggered stations, but also those pairs where only one station has triggered.

Therefore, when the first station,  $S_n$ , triggers with an arrival at  $t_n = t_{\text{now}}$  ( $t_{\text{now}}$  is the current clock time), we can already place some limit on the hypocenter position (Fig. 8b). These limits are given by EDT surfaces defined by the condition that each operational but not-yet-triggered station  $S_l$  will trigger in the next time instant,  $t_l \geq t_n$ . That is:

$$(tt_l - tt_n)_{i,j,k} = t_l - t_n \geq 0; \quad l \neq n. \quad (2)$$

On these conditional EDT surfaces, the *P* travel time to the first triggering station  $tt_n$  is equal to the travel-time to each of the not-yet-triggered stations,  $tt_l$ ,  $l \neq n$ . These surfaces bound a volume (defined by the system of inequalities) which must contain the hypocenter. In the case of a homogeneous medium with constant *P*-wave speed, this hypocentral volume is the Voronoi cell around the first recording station, defined by the perpendicular bisector surfaces with each of the immediate neighboring stations.

As the current time  $t_{\text{now}}$  progresses, we gain the additional information that the not-yet-triggered stations can only trigger with  $t_l > t_{\text{now}}$ . Thus the hypocentral volume is bounded by conditional EDT surfaces that satisfy the inequality:

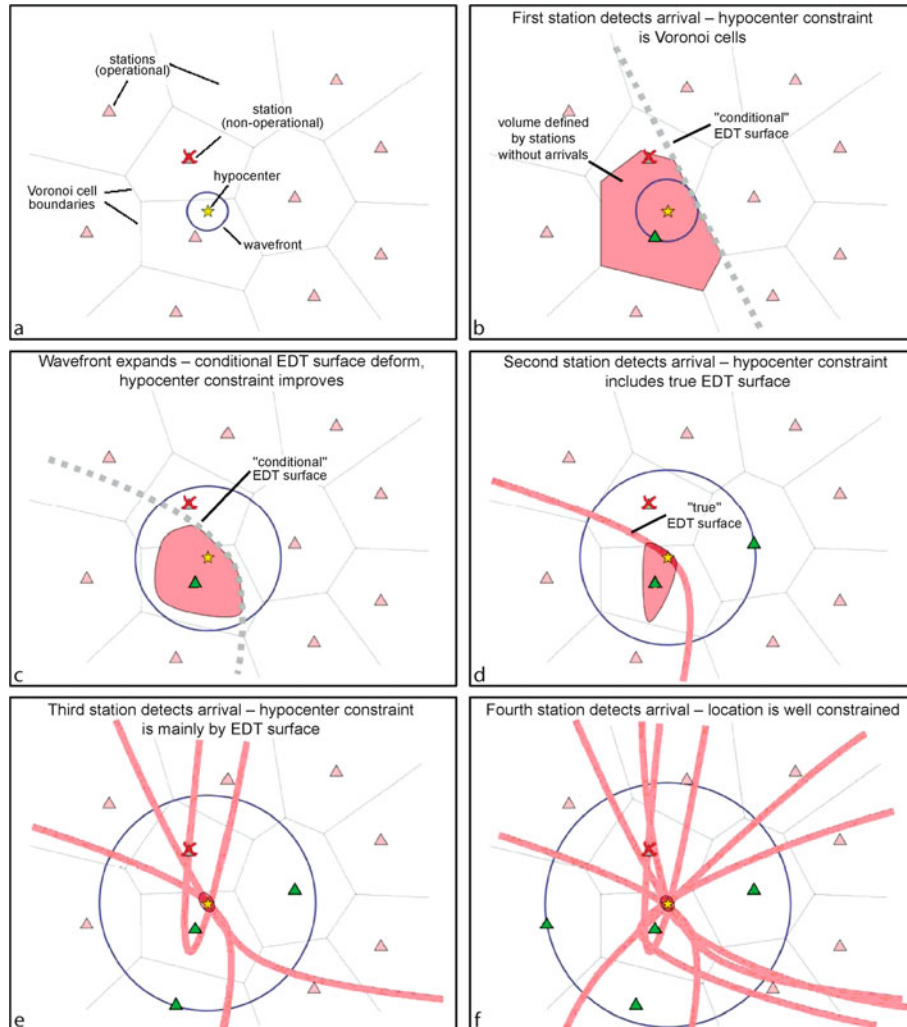
$$(tt_l - tt_n)_{i,j,k} \geq \delta t_{n,l}; \quad l \neq n. \quad (3)$$

$\delta t$  is the time interval between the arrival time at station  $S_n$  and the latest time for which we have information from station  $S_l$ ,

$$\delta t_{n,l} = t_{\text{now}} - t_n, \quad (4)$$

where  $t_n$  is the observed arrival time at station  $S_n$ .

The system (3) defines the volume, bounded by the conditional EDT surfaces, in which the hypocenter may be located given that, at current time  $t_{\text{now}}$ , only the station  $S_n$  has triggered. When  $\delta t = 0$  the system (3) reduces to the system (2); for  $\delta t > 0$ , the hypocentral volume will be smaller than the previous one, since the updated, conditional EDT surfaces tend to fold towards and around the first triggered station (Fig. 8c).



Earthquake Early Warning System in Southern Italy, Figure 8

Schematic illustration of the evolutionary earthquake location algorithm. For clarity, only a map view with the epicentral location is represented. **a** Given a seismic network with known sets of operational and non-operational stations, we can define a priori the Voronoi cell associated to each station. **b** When the first station triggers, we can define a volume that is likely to contain the location, this volume is limited by conditional EDT surfaces on which the  $P$  travel time to the first triggering station is equal to the travel-time to each of the operational but not-yet-triggered stations. **c** As time progresses, we gain additional information from the stations that have not yet triggered, the EDT surfaces move towards and bend around the first triggering station, and the likely-location volume decreases in size. **d** When the second station triggers, we can define a true EDT surface; the hypocenter is on the intersection between this surface and the volume defined by the conditional EDT surfaces, which continues decreasing in size. **e** When a third station triggers, we can define two more true EDT surfaces, further increasing the constraint on hypocenter position. **f** As more stations trigger, the location converges to the standard EDT location composed entirely of true EDT surfaces

We interpret the hypocentral volume in a probabilistic way by defining, for each inequality in (3), a value  $p_{n,l}(i, j, k)$  which is 1 if the inequality is satisfied and 0 if not. Then we sum the  $p_{n,l}(i, j, k)$  over stations  $l$  at each grid point, obtaining a non-normalized probability density  $P(i, j, k)$ , where  $P(i, j, k) = N - 1$  for grid points where

all the inequalities are satisfied and a value less than  $N - 1$  elsewhere.

When the second and later stations trigger, we first re-evaluate the system (3) for all pairs of triggered stations  $S_n$  and all not-yet-triggered stations  $S_l$ . Secondly, we construct standard, true EDT surfaces (see Eq. 2) between

each pair  $S_n, S_m$  of the triggered stations, by evaluating for each grid point the quantity:

$$q_{n,m}(i, j, k) = \exp \left\{ -\frac{[(tt_n - tt_m)_{i,j,k} - (t_n - t_m)]^2}{2\sigma^2} \right\};$$

$$n \neq m. \quad (5)$$

The expression between square brackets at the exponent is the standard EDT Eq. 2 whose solutions are quasi-hyperbolic surfaces; in practice all true EDT surfaces are given a finite width by including the uncertainty  $\sigma$  in the arrival time picking and the travel-time calculation.

The quantity  $q_{n,m}(i, j, k)$  has values between 0 and 1. We sum the  $q_{n,m}(i, j, k)$  with the  $p_{n,l}(i, j, k)$  obtained from the re-evaluation of (4) to obtain a new  $P(i, j, k)$ .

Starting from  $P$ , we define a value:

$$Q(i, j, k) = \left( \frac{P(i, j, k)}{P_{\max}} \right)^N, \quad (6)$$

which forms a relative probability density function (PDF, with values between 0 and 1) for the hypocenter location within the grid cell  $(i, j, k)$ . The function  $Q(i, j, k)$  may be arbitrarily irregular and may have multiple maxima.

At predetermined time intervals, we evaluate (3) and (5) to obtain  $Q(i, j, k)$  in the search volume, using the Oct-tree importance sampling algorithm ([13,27], <http://www.alomax.net/nlloc/octtree>). This algorithm uses recursive subdivision and sampling of rectangular cells in 3D space to generate a cascade structure of sampled cells, such that the spatial density of sampled cells follows the target function values. The Oct-tree search is much faster than a simple or nested grid search (factor 10–100 faster) and more global and complete than stochastic search methods algorithms such as simulated annealing and genetic algorithms [13]. For each grid point, an origin time estimate can be obtained from the observed arrival times and the calculated travel times.

As more stations trigger, the number of not-yet-triggered stations becomes small, and the location converges towards the hypocentral volume that is obtained with standard EDT location using the full set of data from all operational stations (Fig. 8d–f).

If there are uncorrelated outlier data (i.e., triggers that are not compatible with  $P$  arrivals from a hypocenter within or near the network), then the final hypocentral volume will usually give an unbiased estimate of the hypocentral location, as with standard EDT location. However, if one or more of the first arrival times is an outlier, then the earliest estimates of the hypocentral volume

may be biased. Synthetic tests have shown that, if  $N_{\text{out}}$  is the number of outlier data, the bias reduces significantly after about  $4 + N_{\text{out}}$  arrivals have been obtained, and then decreases further with further arrivals, as the solution converges towards a standard EDT location [37].

We performed several synthetic tests using the geometry of the ISNet network. For each simulated event, we computed theoretical arrival picks using travel times obtained by the finite difference solution of the eikonal equation [35] for a 1D,  $P$ -wave velocity model. To reproduce uncertainties introduced by the picking algorithm, we add to each arrival time a random error following a Gaussian distribution with a variance of 0.02 s.

Here we use only  $P$  picks since currently most networks have poor capability to perform real-time  $S$  picking. Our tests consider an earthquake occurring at the center of the network at a depth of 10 km (Fig. 9) and an earthquake occurring outside the network at a depth of 10 km (Fig. 10). Each panel in Figs. 9 and 10 is a snapshot at a given time showing the marginal map (i.e., summed over  $i, j$  or  $k$ ) for  $Q(i, j, k)$  along the horizontal ( $x, y$ ) and the two vertical ( $x, z$  and  $y, z$ ) planes. The star shows the known, synthetic hypocentral location. In the first case, two seconds after the first trigger (5.03 s from the event origin), 9 stations have triggered and the location is already well constrained for early warning purposes.

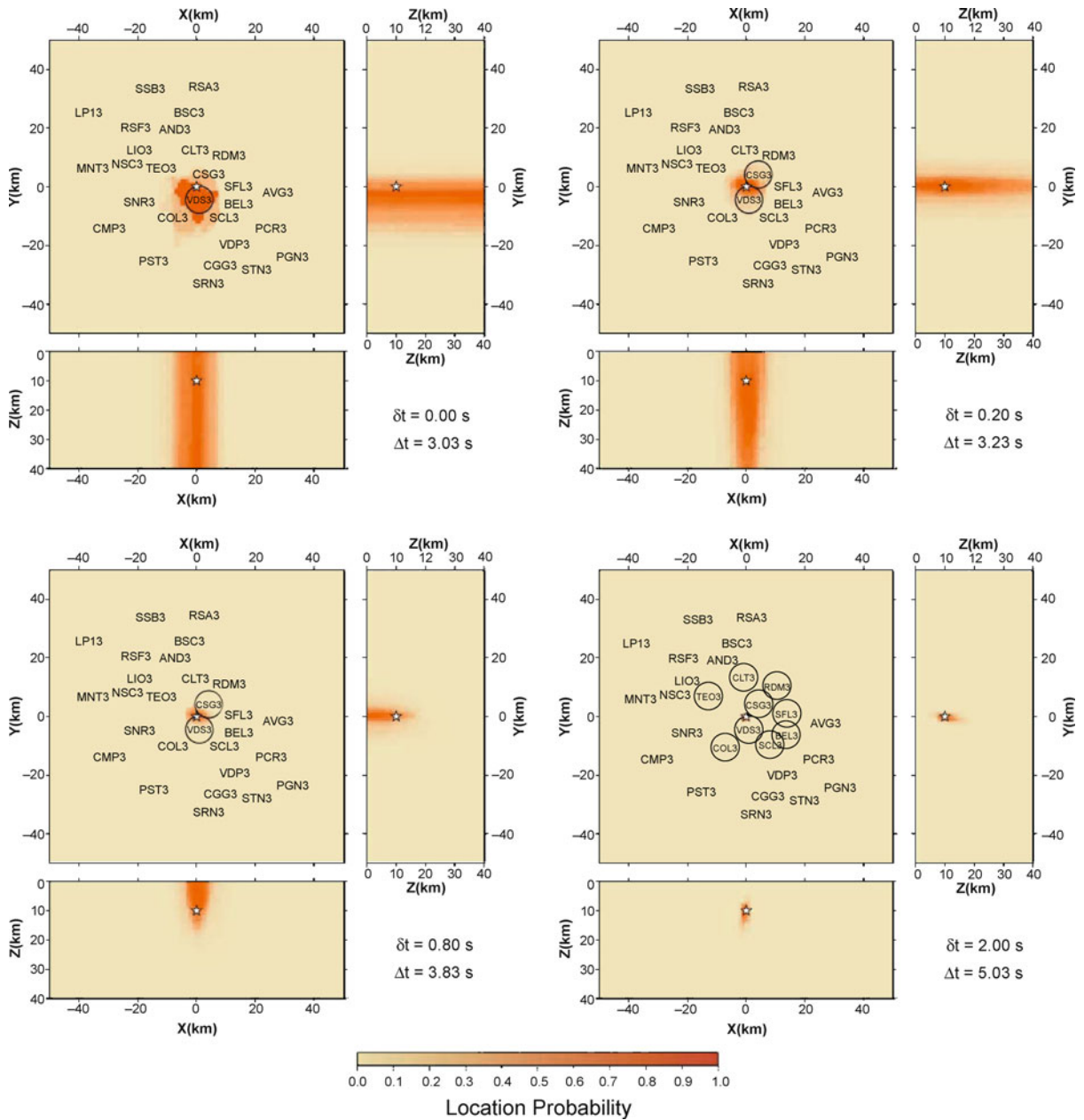
In the second case, at  $\Delta t = 11.76$  s, 2 s after the first event detection, the constraint on the location PDF improves further, but the PDF retains an elongated shape because of the poor azimuthal coverage of the network for this event. The event depth is only constrained by an upper bound, but the depth range includes the true value.

### Real-Time Magnitude Estimation Using a Bayesian, Evolutionary Approach

**Previous Related Studies** The problem of magnitude estimation from early seismic signal has been previously approached and analyzed by different authors.

Nakamura [31] first proposed the correlation between the event magnitude and the characteristic period of  $P$ -phase defined as the ratio between the energy of the signal and its first derivative.

Allen and Kanamori [2] modified the original Nakamura method and described the correlation between the predominant period and the event magnitude for Southern California events. Lockman and Allen (2007) studied the predominant period – magnitude relations for the Pacific Northwest and Japan. They also investigated the sensitivity of such relations using different frequency bands.

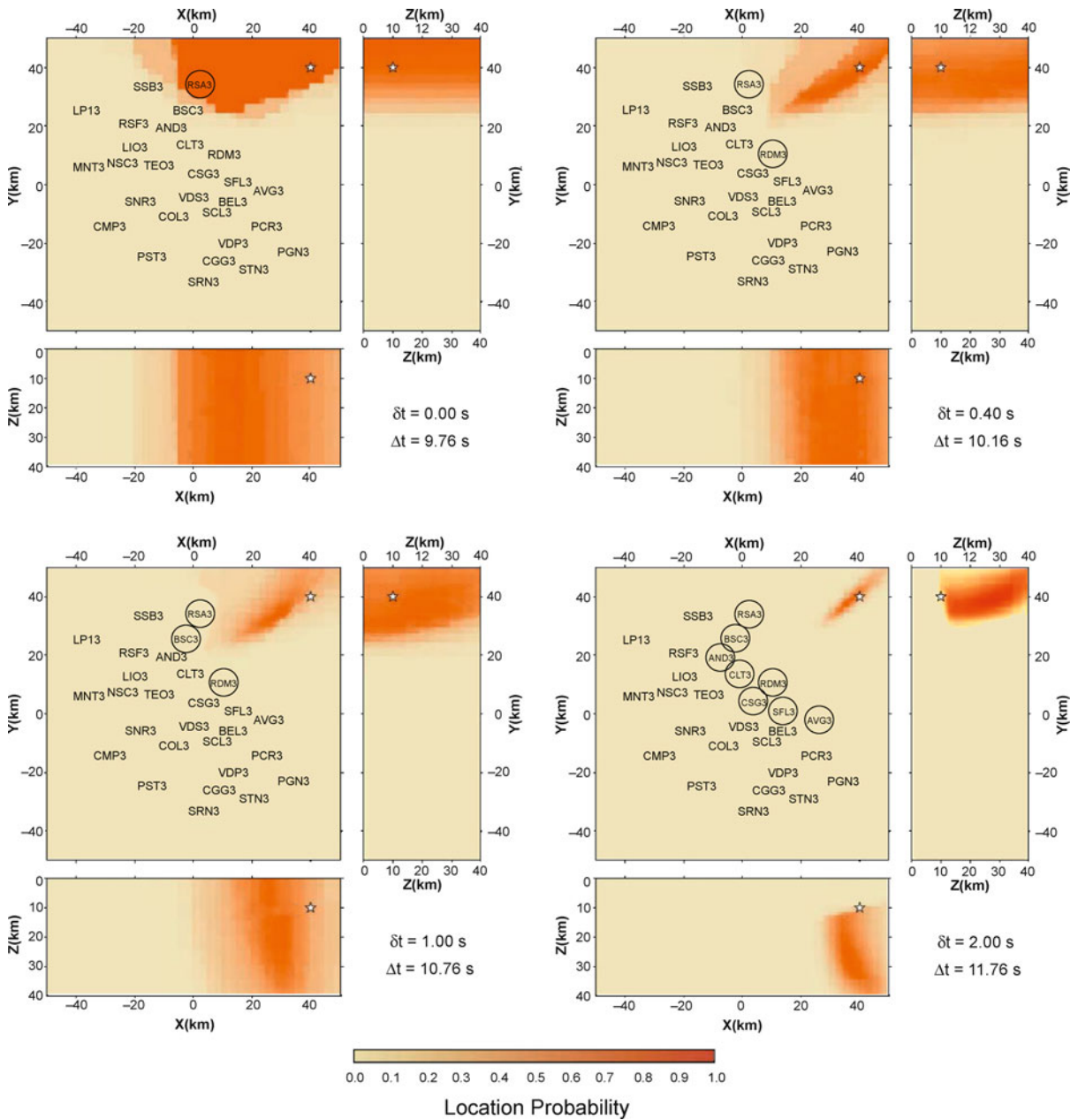


Earthquake Early Warning System in Southern Italy, Figure 9

Location test for a synthetic event occurring at the center of the Irpinia Seismic Network (ISNet). The three orthogonal views show marginal values of the probability function  $Q(i, j, k)$ . The true hypocenter is identified by a star.  $\delta t$  is the time from the first trigger,  $\Delta t$  is the time from event origin. For each snapshot, stations that have triggered are marked with a circle

Using a complementary approach, Wu and Kanamori [45] investigated the feasibility of an on-site EEWs for Taiwan region based on prediction of earthquake damage, based on measurements of the predominant period and peak displacement on early  $P$ -wave signals detected at the network.

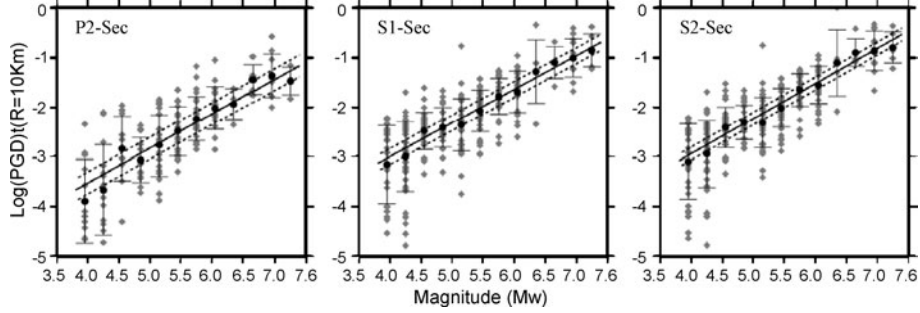
Odaka et al. (2003) proposed a single station approach for the real-time magnitude estimation. The authors fit the initial part of waveform envelope and showed a relation between the final event magnitude, the envelope shape coefficient and the maximum  $P$  amplitude measured in a 3 s time window.



Earthquake Early Warning System in Southern Italy, Figure 10  
 Location test for a synthetic event occurring outside the ISNet network (see Fig. 9 for explanation)

Wu and Zhao [46] and Zollo et al. [49] (Fig. 11) demonstrated the existence of a correlation between the event magnitude and the peak displacement measured a few seconds after the *P* arrival based on massive analysis of Southern Californian and Euro-Mediterranean earthquake records. In particular, Zollo et al. showed that

both *P* and *S* wave early phases have the potential for real time estimation of magnitude up to about *M* 7. Zollo et al. [50] and Lancieri and Zollo [26] extended this observation to Japanese earthquake records, showing that a possible saturation effect may exist at about *M* 6.5 for *P* measurements in 2 s windows while it vanishes when a larger,



Earthquake Early Warning System in Southern Italy, Figure 11

Correlation between low-pass filtered peak ground motion value and moment-magnitude for earthquakes occurred in the Euro-Mediterranean region (after [49]). The panels show the logarithm of peak ground displacement normalized at a reference distance of 10 km as a function of Mw in time windows of (left) 2 s length from the first *P*-arrival and (middle) 1- and (right) 2-s from the first *S*-arrivals. *P*- and *S*-data are measured on vertical and root-squared sum of horizontal components, respectively. Each panel shows the best fit regression line (solid line) along with 1-WSE limits (dashed lines)

4 s window is considered. The scaling of displacement peak with magnitude, instead, appears at even shorter (1 s) time lapses after the first *S*-arrival.

Using an alternative method, Simmons [38] proposed a new algorithm based on discrete wavelet transforms able to detect first *P* arrival and to estimate final magnitude analyzing first seconds of *P*-wave.

**The Real-Time Magnitude Estimation Method** The real time and evolutionary algorithm for magnitude estimation presented in this paper is based on a magnitude predictive model and a Bayesian formulation. It is aimed at evaluating the conditional probability density function of magnitude as a function of ground motion quantities measured on the early part of the acquired signals [19].

The predictive models are empirical relationships which correlate the final event magnitude with the logarithm of quantities measured on first 2–4 s of record.

The first prediction model, based on the predominant period of *P*-phase ( $\tau_P$ ), has been introduced by Allen and Kanamori [2]. Recently, Wu and Zhao [46] showed the existence of a correlation between magnitude, distance and peak displacement measured in a 2–4 s window after *P*-phase.

Zollo et al. [49,50] refined this correlation and extended the observation on the peaks measured in 2 s after the *S*-phase arrival through the analysis of the European and Japanese strong motion data-bases (Ambraseys et al. [3], K-NET www service of NIED – National Research Institute for Earth Science and Disaster Prevention, Japan).

The method therefore assumes that the linear relationship between the logarithm of the observed quantity and

magnitude is known, along with standard errors of the predictive models.

At each time step  $t$  from the first station trigger, the conditional PDF of magnitude  $M$  given the observed data vector  $\underline{d} = \{d_1, d_2, \dots, d_n\}$  is expressed via the Bayes theorem as:

$$f(m|\underline{d}) = \frac{f(\underline{d}|m)f(m)}{\int_{M_{\text{MIN}}}^{M_{\text{MAX}}} f(\underline{d}|m)f(m)dM}, \quad (7)$$

where  $f(m)$  is the a priori distribution which incorporates the information available before the experimental data are collected through a truncated exponential functional form, derived by the Gutenberg–Richter recurrence relationship,

$$f(m) : \begin{cases} \frac{\beta e^{-\beta m}}{e^{-\beta M_{\text{min}}} - e^{-\beta M_{\text{max}}}} & M_{\text{min}} \leq m \leq M_{\text{max}} \\ 0 & m \notin [M_{\text{min}}, M_{\text{max}}] \end{cases}, \quad (8)$$

where  $\{\beta, M_{\text{min}}, M_{\text{max}}\}$  depend on the seismic features and on the detection threshold of the seismic network of the considered region.

The conditional probability  $f(\underline{d}|m)$  contains all the information concerning the magnitude as retrievable from the data acquired at time  $t$ .

Assuming that components of the observed data vector  $\underline{d}$  have a lognormal distribution, and that they are stochastically independent and identically distributed random variables of parameters  $\mu_{\log(d)}$  and  $\sigma_{\log(d)}$ , then the likelihood is written as:

$$f(\underline{d}|m) = \prod_{i=1}^v \frac{1}{\sqrt{2\pi}\sigma_{\log(d)}d_i} e^{-\frac{1}{2}\left(\frac{\log(d_i) - \mu_{\log(d)}}{\sigma_{\log(d)}}\right)^2}, \quad (9)$$

where  $\nu$  is the number of stations acquiring at the instant  $t$ ;  $\mu_{\log(d)}$  and  $\sigma_{\log(d)}$  are the mean and the standard deviation of the logs of  $d_i$ , respectively.

Substituting Eq. 8 and Eq. 9 into Eq. 7,  $f(m|d)$  results as in Eq. 10 where it depends on data only through  $\sum_{i=1}^{\nu} \log(d_i)$  and  $\nu$ , which therefore are jointly sufficient statistics for the estimation of magnitude [21]:

$$f(m|d) = f\left(m \mid \sum_{i=1}^{\nu} \log(d_i)\right) = \frac{e^{\left(2\mu_{\log(d)}\left(\sum_{i=1}^{\nu} \log(d_i)\right) - \nu\mu_{\log(d)}^2\right) / 2\sigma_{\log(d)}^2} e^{-\beta m}}{\int_{M_{\text{MIN}}}^{M_{\text{MAX}}} e^{\left(2\mu_{\log(d)}\left(\sum_{i=1}^{\nu} \log(d_i)\right) - \nu\mu_{\log(d)}^2\right) / 2\sigma_{\log(d)}^2} e^{-\beta m} dm} \quad (10)$$

As just outlined,  $f(m|d)$  depends on  $\sum_{i=1}^{\nu} \log(d_i)$  and on the number of stations triggered,  $\nu$ , at the time of the estimation and, consequently, on the amount of information available. As more stations are triggered, and provide more measures of  $d$ , the estimation improves.

The described technique is evolutionary in the sense that  $f(m|d)$  depends on time, i. e., as time passes, additional stations provide new observations (predominant period and/or  $P$ -,  $S$ -peaks), which are used to refine the probabilistic estimation of magnitude.

### Magnitude Estimation from Peak Displacement Measurements

The empirical relationships between low-pass filtered, initial  $P$ - and  $S$ -peak displacement amplitudes and moment magnitude (e. g. [49]) can be used as predictive models for the real-time estimation of magnitude using the Bayesian approach described above.

While the  $P$ -wave onset is identified by an automatic picking procedure, the  $S$ -onset can be estimated from an automatic  $S$ -picking or from a theoretical prediction based on the hypocentral distance given by the actual earthquake location. At a given time step after the first  $P$ -wave detection at the network, progressively refined estimates of magnitude are obtained from  $P$ - and  $S$ -peak displacement data. These are preliminarily corrected for distance amplitude effects through an empirical attenuation relationship obtained from available strong motion records [46,49]:

$$f(M, R) = A_{\text{phase}} + B_{\text{phase}}M + C_{\text{phase}} \log(R), \quad (11)$$

where the constants  $A_{\text{phase}}$ ,  $B_{\text{phase}}$  and  $C_{\text{phase}}$  are determined through a best-fit regression with a retrieved standard error of  $SE_{\text{phase}}^{\text{PMR}}$  and  $R$  is the hypocentral distance.

Following the procedure described in [49], the relationship (11) is used to correct observed peaks for the distance effect, by normalizing them to a reference distance (e. g.,  $R = 10$  km) and to determine a new best fit regression between the distance corrected peak value  $(PD_{\text{phase}})^{10 \text{ km}}$  and the final magnitude:

$$\log\left(PD_{\text{phase}}^{10 \text{ km}}\right) = \log\left(PD_{\text{phase}}^R\right) - C_{\text{phase}} \log\left(\frac{R}{10}\right) \quad (12)$$

$$\log\left(PD_{\text{phase}}^{10 \text{ km}}\right) = A'_{\text{phase}} + B'_{\text{phase}}M. \quad (13)$$

Assuming a standard error of  $SE_{\text{phase}}^{\text{PM}}$  on peak displacements retrieved from (13) and combining the Eqs. (11) and (13), the mean values and standard deviation of quantity  $\log(PD_{\text{phase}})$ , can be written as:

$$\begin{aligned} \mu_{\log(PD_{\text{phase}})} &= B'_{\text{phase}}M + A'_{\text{phase}} + C_{\text{phase}} \log\left(\frac{R}{10}\right) \\ \sigma_{\log(PD_{\text{phase}})} &= SE_{\text{phase}}^{\text{PM}} + \log\left(\frac{R}{10}\right) \Delta C_{\text{phase}} \\ &\quad + C_{\text{phase}} \frac{1}{R} \Delta R, \end{aligned} \quad (14)$$

where  $R$  is estimated with an error of  $\Delta R$  and  $\Delta C_{\text{phase}}$  is the error on the  $C_{\text{phase}}$  coefficient in Eq. (12).

The values of coefficients in (14) used for real time magnitude estimates at ISNet are obtained from the regression analysis based on records from the European Strong Motion Database [49] and given in Table 2.

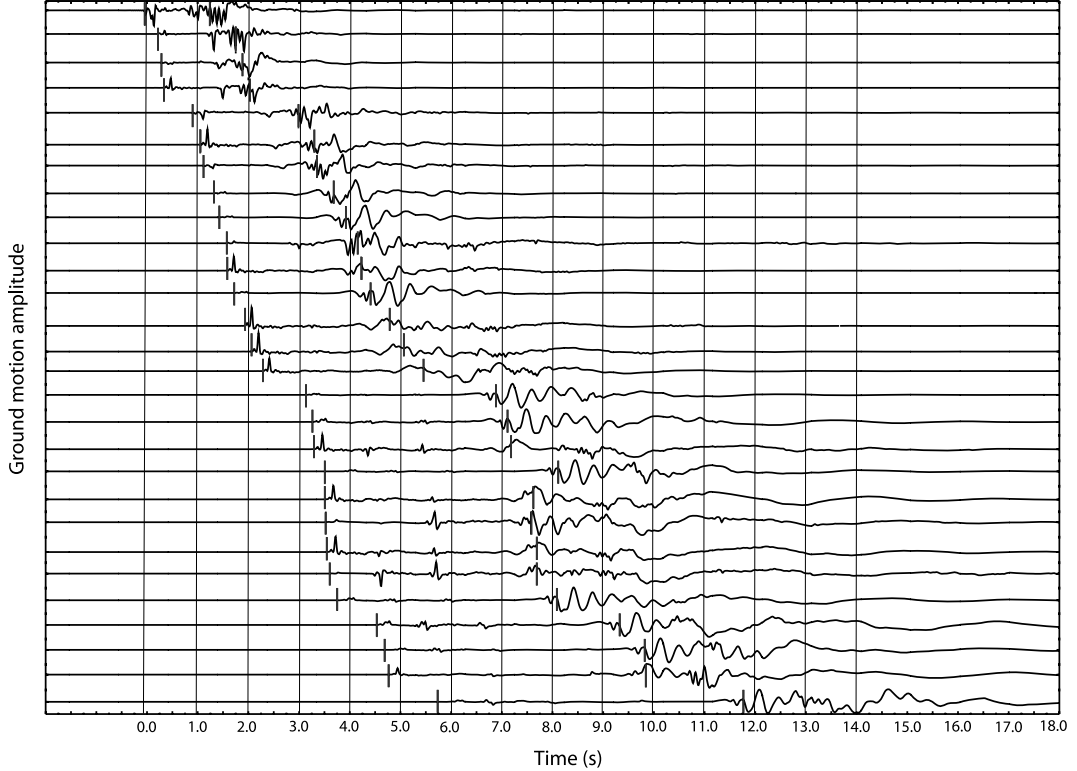
Figure 12 illustrates an example of real time magnitude estimation on a simulated event with  $M = 7.0$ , whose epicenter is located along the 1980 Irpinia earthquake faulting system. Synthetic seismograms have been computed by using the discrete wave-number method of Bouchon [8] and Coutant (1989) for a extended source model in a flat-layered velocity model.

Figure 13a shows the probability density function defined in Eq. (8) evaluated at each time step. Time zero is assigned to the first  $P$  detection at the network. As time evolves the PDF tightens around the predicted magnitude value, indicating a more refined, probabilistic estimate of magnitude.

Earthquake Early Warning System in Southern Italy, Table 2  
Coefficients of the empirical regression relationships between low-pass filtered  $P$  and  $S$  displacement peaks and magnitude

Phase	$A'_{\text{phase}}$	$B'_{\text{phase}}$	$C_{\text{phase}}$	$SE_{\text{phase}}^{\text{PM}}$	$\Delta C_{\text{phase}}$
2P	-6.31	0.70	-1.05	0.22	0.30
2S	-5.77	0.71	-0.71	0.13	0.16





Earthquake Early Warning System in Southern Italy, Figure 12

Synthetic seismograms for a  $M 7.0$  earthquake at the center of the network (see Fig. 9). The seismograms are computed using a line source, rupture model (constant rupture velocity) while complete wavefield green's functions in a flat-layered model are computed by using the discrete wavenumber summation method of Bouchon [8]. Each vertical line indicates the 1 s signal packets examined at each time step. This plot allows us to understand seconds after seconds which stations are acquiring and what sort of input ( $P$  or  $S$  peak) they are giving to the real time system. For example after three seconds to the first  $P$  phase picking thirteen stations are acquiring, the 2 s  $S$ -phase peak is available at the nearest stations. This observation motivates the use of the  $S$  phase information in a real time information. If a dense network is deployed in the epicentral area the nearest station will record the  $S$ -phase before the  $P$  phase arrives to the far ones, as seen in previous example, and this is perfectly compatible with the real time analysis

By defining  $F_t(m)$  as the cumulative PDF at time  $t$ , it is possible to estimate a magnitude range of variation  $[M_{\min}; M_{\max}]$  whose limits are defined based on the shape of the  $F_t(m)$  function:

$$\begin{aligned} M_{\min} &: \int_{-\infty}^{M_{\min}} f_t(m|\underline{d}) dm = \alpha, \\ M_{\max} &: \int_{-\infty}^{M_{\max}} f_t(m|\underline{d}) dm = 1 - \alpha. \end{aligned} \quad (15)$$

For example, if we assume  $\alpha = 1\%$ , then  $M_{\min}$  and  $M_{\max}$  will be, respectively, the  $F_t(m)$  evaluated at 0.01 and 0.99.

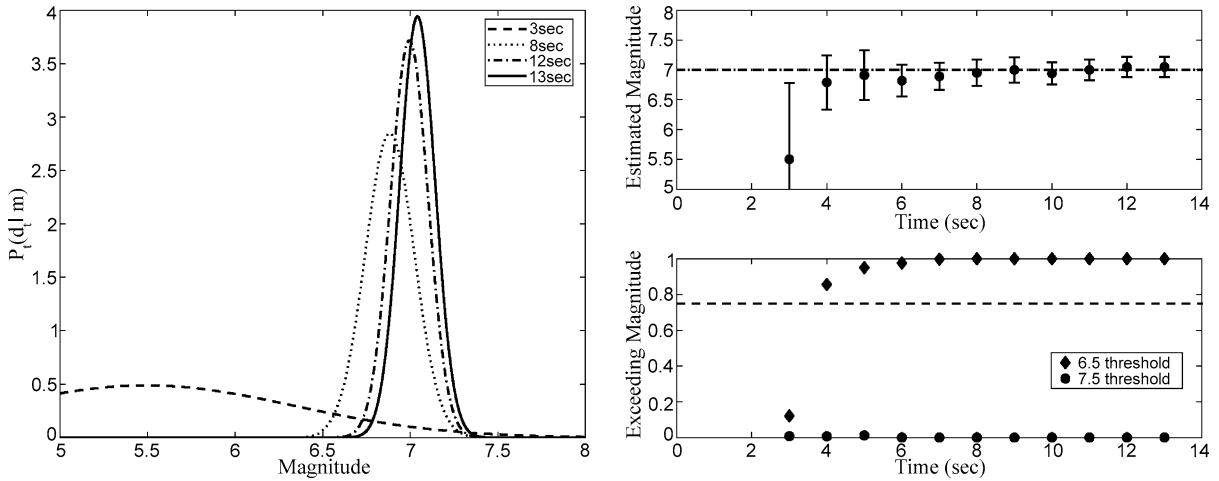
In Fig. 13b the estimates of magnitude uncertainty bounds are reported as a function of time. After three iterations (corresponding to a time of 9 s from the event origin time and 4 s after the first  $P$ -phase arrival at the network) the magnitude estimation converges to the true magnitude

value. In fact, due to the high density of seismic station in the epicentral area, at that time most of seismic station contributes to the magnitude estimation with peaks read on  $P$ -phase windows (Fig. 14), while a further refinement of magnitude estimate is due to the near source  $S$ -wave arrivals.

### Real-Time Hazard Analysis for Earthquake Early Warning

#### The Real-Time Hazard Determination

Using the methods previously described for estimating in real-time the event magnitude and location, it is possible to perform a real-time hazard analysis [19]. This analysis is based on the extension of classical Probabilistic Seismic Hazard Analysis (PSHA) proposed by Cornell [11] that is generally used for long-term probabilistic hazard



Earthquake Early Warning System in Southern Italy, Figure 13  
 Application of the method for real time magnitude estimate to a  $M 7$  simulated event occurring within the area covered by the ISNet network. *Left panel.* PDF distribution at several time steps measured from the first  $P$ -phase picking. *Right top,* magnitude estimation with uncertainties as a function of time. The *dashed line* refers to the actual magnitude value, the errors represent the 95% of confidence bound evaluated as cumulative PDF integral in the 5–95% range. *Right bottom,* probability to exceed magnitude 6.5 and magnitude 7.5 thresholds in function of time. The *dashed line* is the 75% probability level

assessment. Classical PSHA integrates data from existing seismic catalogs both in terms of magnitude, location and recorded strong ground motion values in addition to the information concerning seismogenic areas of interest (expected maximum magnitude,  $b$ -value of the Gutenberg Richter relationship, etc.) to provide the hazard curve as the final outcome. Each point on that curve corresponds to the value of a ground motion intensity measure (IM) (e. g., peak ground acceleration, PGA, peak ground velocity, PGV or the spectral acceleration,  $S_a$ ), having a given probability or frequency of exceedance in a fixed period of time for a site of interest.

The probabilistic framework of the PSHA, specifically the hazard integral, can be used for real-time hazard if the PDFs of magnitude and source-to-site distance are replaced with those depending on the data gathered by the EEWs during the occurrence of a specific earthquake.

This is the case, for example, of the PDF on the source-to-site distance whose statistical moments evolve with real-time earthquake location. As a consequence, this PDF does not depend on the seismic potential of the area of interest (as in the case of the classical PSHA, which accounts for the occurrence of all the earthquake in a fixed range of magnitude), but rather depends on the time evolving event location provided by the EEWs. The same considerations apply to the PDF on the magnitude as described in the following sections whose statistical moment, at a given

time, depends on the number of triggered stations at that time.

In this theoretical framework the real-time hazard integral can be written as:

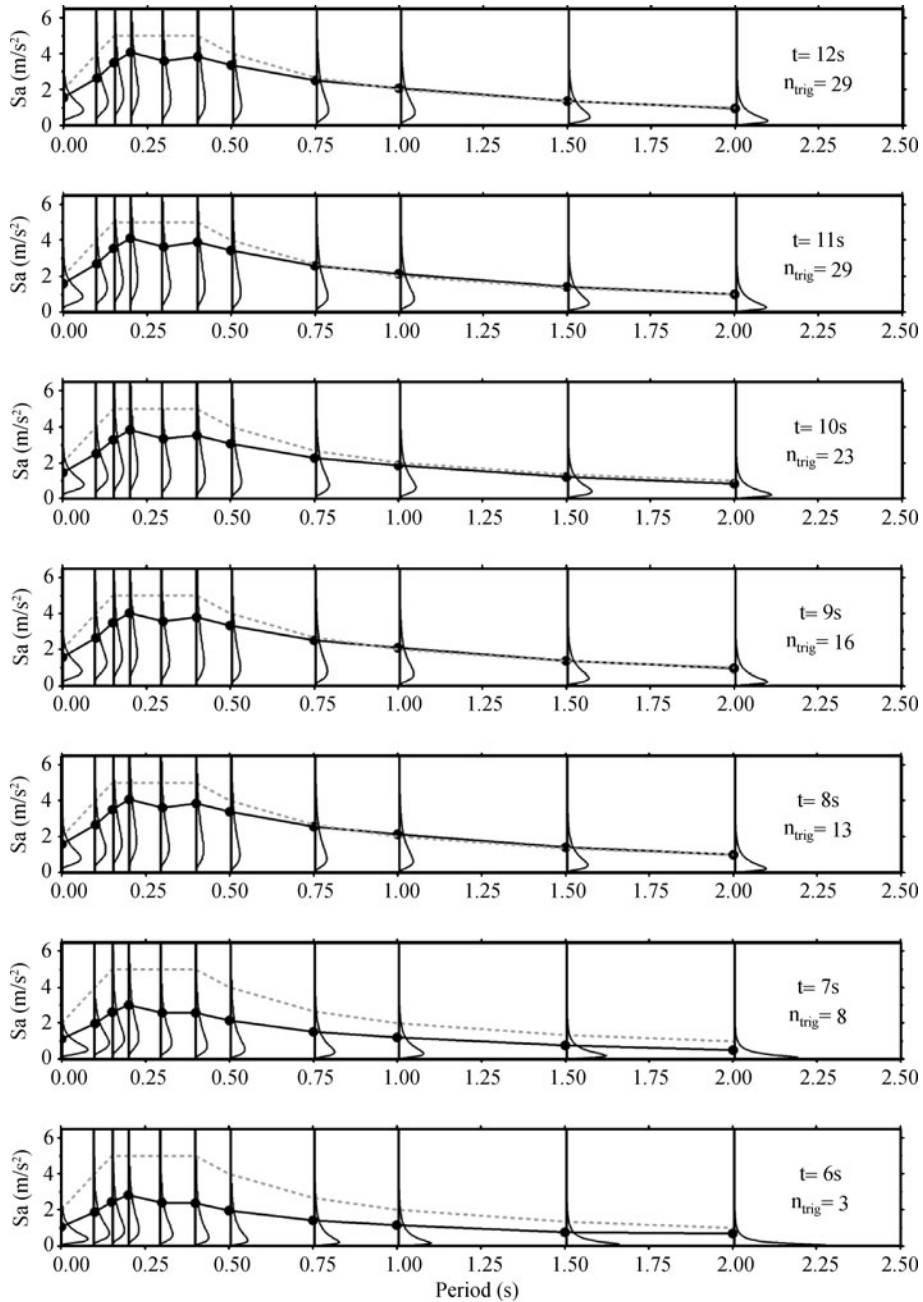
$$f(\text{IM}|\underline{d}, \underline{s}) = \int_M \int_R f(\text{IM}|m, r) f(m|\underline{d}) f(r|\underline{s}) dM dR, \quad (16)$$

where  $f(r|\underline{s})$  is the PDF of distance  $r$ , which eventually depends only on the triggering sequence of the stations in the network, where  $\underline{s} = \{s_1, \dots, s_N\}$  is such a sequence. This renders also the PDF of  $r$  time dependent.

Given that for each point in a volume containing the earthquake hypocenter, the probability of that point being coincident with the true hypocenter is calculated via a rapid location technique, a simple geometrical transformation allows one to obtain the probabilistic distribution of the source-to-site distance.

The PDF  $f(\text{IM}|m, r)$ , is given, for example, by an ordinary attenuation relationship. It is worth to recall that the computed hazard refers to a particular set of triggered stations and, consequently, it depends on the information available at time  $t$  from the first detection of the event.

Figure 14 illustrates, as an example, the estimation of spectral acceleration ordinates for different periods, for a  $M 7.0$  event located at an epicentral distance of 50 km from the early warning target site [10].



Earthquake Early Warning System in Southern Italy, Figure 14

Real-time estimation of spectral ordinates' distributions as function of the number of stations triggered for a  $M$  7.0 event with an epicentral distance of 50 km from the early warning target site. The parameter  $n_{trig}$  in the figure is equivalent to the number of stations  $\nu$  in the text. The acceleration spectrum (*black curve*) was obtained by choosing at each period the spectral value with 20% exceedance probability according to the corresponding distribution, so it is analogous to a uniform hazard spectrum with the exception that it is computed in real-time. The *grey dashed line* is the Italian code spectrum assigned for building design in the target location at the town of Avellino, 40 km distant from the earthquake epicenter, and is reported for comparison purposes (after [10])

We note the evolution of Sa predictions via the corresponding PDFs. The different panels correspond to increasing times from the earthquake origin and, therefore, to different numbers of stations triggered.

### The False Alarm Issue

Once the EEWS provides a probability distribution of the ground motion intensity measure (IM) at the target site (e. g., peak ground acceleration or velocity), a decisional condition has to be checked in order to decide whether to alert or not.

Several options are available to formulate a decisional rule, for example the alarm may be issued if the probability of the predicted IM exceeding a critical threshold ( $IM_C$ ) is greater than a reference value ( $P_c$ ):

$$\text{Alarm if: } \int_0^{IM_C} f(IM|\underline{d}, \underline{s})d(IM) = P[IM > IM_C] > P_c. \quad (17)$$

The efficiency of the decisional rule may be evaluated in terms of false and missed alarms probabilities (known as the “cry wolf” issue, e. g., [20]). The false alarm occurs when, on the basis of the information processed by the EEWS, the alarm is issued while the intensity measure at the site  $IM_T$  (T subscript means “true”, indicating the realization of the IM to be distinguished from the prediction  $IM_C$ ) is smaller than the threshold  $IM_C$ . A missed alarm corresponds to not launching the alarm if needed,

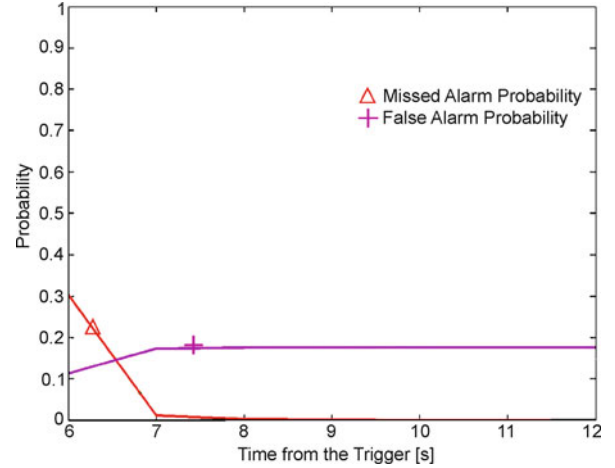
$$\begin{aligned} \text{Missed Alarm: } & \{\text{No Alarm} \cap IM_T > IM_C\} \\ \text{False Alarm: } & \{\text{Alarm} \cap IM_T \leq IM_C\}. \end{aligned} \quad (18)$$

It has been discussed above how the information and the uncertainties on earthquake location and magnitude are dependent on the number of stations triggered at a certain time.

Therefore, in principle, the decisional rule may be checked at any time after the first station has triggered and, consequently, the false and missed alarm probabilities are also time dependent.

Using the decisional rule of (18) and considering PGA as IM, the time evolution of false/missed alarm probabilities has been simulated for the Campania EEWS, given the occurrence of a  $M 7$  earthquake, and a target site at an epicentral distance of 110 km.

Figure 15 reports the missed and false alarm probabilities as a function of time from the first trigger at the ISNet network.



Earthquake Early Warning System in Southern Italy, Figure 15 Example of estimation of false and missed alarm probabilities as function of the time from the first trigger for a  $M 7.0$  event with an epicentral distance of 110 km from the early warning target site. For the decisional rule adopted in this case the threshold is  $PGAc = 0.3 \text{ m/s}^2$  and the limit probability is  $P_c = 0.2$  (after [19])

### A Loss Estimation Approach to Early Warning

Magnitude and distance distributions conditioned to the measurements of the seismic network can also be used for a real-time estimation of risk, which includes losses produced by the earthquake [21]. Based on the real-time risk assessment, a security action aimed at risk mitigation is undertaken if the alarm is issued.

For example, some critical system could shut down or people in buildings may shelter themselves if the warning time is not sufficient to evacuate the dangerous buildings. More complex security measures may be related to the semi-active control of buildings [22].

Therefore, if an EEWS exists, it may trigger a security procedure in case of warning. The estimation of the expected losses for a specific building may be computed, for the case of warning issued and not issued respectively:

$$\begin{aligned} E^W[L|\underline{d}, \underline{s}] = & \int_L \int_{DM} \int_{EDP} \int_{IM} l f^W(l|dm) f(dm|edp) \\ & \times f(edp|im) f(im|\underline{d}, \underline{s}) dL dDM dEDP dIM, \end{aligned} \quad (19)$$

where  $f^W(l|dm)$  is the PDF of the loss ( $L$ ) given the structural and non-structural damage vector ( $DM$ ) reflecting the risk reduction in the case of warning; and  $f^W(l|dm)$  is the loss function if no alarm is issued (no security action is undertaken);  $f(dm|edp)$  is the joint PDF of damages given the Engineering Demand Parameters (EDP), proxy for the structural response;  $f(edp|im)$  is the joint PDF of

the EDPs conditioned to a vector of ground motion intensity measures (IM);  $f(im|\underline{d}, \underline{s})$  is the real-time hazard expressed by (16) in the case of a scalar IM.

Being able to compute, before the ground motion hits the site, the expected losses in case of warning ( $W$ ) or not ( $\bar{W}$ ), is relevant for taking the optimal decision, i. e., to alarm if this reduces the expected losses and to not issue any warning otherwise:

$$\text{to alarm if } E^W[L|\underline{d}, \underline{s}] \leq E^{\bar{W}}[L|\underline{d}, \underline{s}]$$

Optimal decision: (20)

$$\text{to not alarm if } E^W[L|\underline{d}, \underline{s}] > E^{\bar{W}}[L|\underline{d}, \underline{s}]$$

which is a better decisional rule in respect to that of (18).

Computing and comparing expected losses, conditioned to the real-time information coming from the EEWS, in the case of alarming or not, allows the determination of the alarm threshold above which it is convenient to issue the warning according to the optimally maximum criterion.

Assessment of average loss reduction determined by issuing an Early Warning provides a quantitative tool to evaluate the efficiency and feasibility of an EEWS.

Other potential advantages given by this approach are that: (a) the threshold may be set on a statistic (i. e., the summation of the logs) inferred from seismic network measurements, dramatically reducing the required computational effort for real-time decision making; (b) it minimizes the cry wolf issue reducing the probability of false and missed alarms thanks to threshold optimization. In fact, although the number of MA and FA depend on the decisional rule adopted to issued the alarm, the approach developed in Iervolino et al. [20,21,22] avoids explicitly considering the missed and false alarm rates associated with the decision, as the choice to alarm or not is taken based on the expected economic loss (not on the estimation of peak ground motion). In other words, if in computing the expected loss one accounts for the costs of false and missed alarms, there is no need to optimize the *cry wolf* issue, and MA and FA rates are at their values determined by the respective costs, and in this sense are optimal.

### Future Directions

We have analyzed and illustrated the main scientific and technological issues related to the implementation and management of an earthquake early warning system under development in the Campania region of southern Italy.

The system is designed for early warning alert notification at distant coastal targets based on a dense, wide-dynamic seismic network (accelerometers, seismometers

and broadband sensors) deployed in the Apenninic belt region (ISNet – Irpinia Seismic Network). It can therefore be classified as a regional Early Warning System consisting of a wide seismic sensor network covering a portion or the entire area which is threatened by a quake's strike.

According to [25], real-time estimates of earthquake location and magnitude are needed for regional warning systems (EEWS), i. e., dense seismic networks covering all or a portion of an area of interest. However the alarm decision in an early warning system is based, rather, on the prediction, with quantified confidence, of a ground motion intensity at a distant target site (where a sensitive structure is located). This problem needs an evolutionary (i. e., time-dependent) and probabilistic frame where pdfs for earthquake location, magnitude and attenuation parameters are combined to perform a real-time probabilistic seismic hazard analysis (e. g., [19]).

Considering the peak displacement amplitude and/or predominant frequency measured in the early portion of  $P$ -waves, we have shown that suitable probability density functions for the earthquake location and magnitude parameters can be constructed and used for real-time probabilistic assessment of false alarms and loss estimation, which are the key elements based on which automatic actions can be undertaken to mitigate earthquake effects.

Based on the analysis of acceleration records of Euro-Mediterranean and Japanese earthquakes, Zollo et al. [49,50] have shown the advantages of using near source strong motion records for real time estimation of earthquake magnitude. In fact they provide unsaturated recordings of moderate to large earthquakes and, in case of dense station coverage of the source area, the combination of both  $P$ - and  $S$ -wave amplitude information can be used to get fast and robust earthquake location and magnitude estimates.

We support the use of  $S$ -waves recorded in the near-source of an impending earthquake for earthquake early warning, especially in view of the excellent correlation that  $S$ -peaks show with magnitude up to about  $M = 7$  for Euro-Mediterranean and Japanese earthquakes [49,50]. Dense accelerometric networks now operating in Europe, USA, Taiwan, Japan and other seismic regions in the world can provide a sufficient number of records at distances smaller than 20–30 km from potentially damaging crustal earthquakes so that  $S$ - $P$  times are expected to be smaller than 2–3 s. A magnitude estimation using  $S$ -waves could be therefore available 4–5 s after the first  $P$ -wave is recorded, which is still useful for sending an alert to distant target sites.

Although relatively few magnitude 7 and larger earthquakes have hit the Apenninic belt, and generally the

Mediterranean region, during the last century, there have been many instances of damaging quakes in the magnitude 6 range.

Earthquake early warning systems have the potential to mitigate the effects of moderate size earthquakes ( $M = 6-7$ ), which can produce severe damage in densely urbanized areas and places where old structures were not built to current standards. This has been the case for a significant number of earthquakes occurred in the Mediterranean basin during last decades: the 1976 Friuli ( $M = 6-6.5$ ) and 1997 Colfiorito ( $M = 6$ ) in Italy, 1999 Athens ( $M = 5.9$ ) in Greece, 2002 Nahrin, in Afghanistan ( $M = 6.1$ ), 2003 in Algeria ( $M = 6.7$ ), 2003 Bam ( $M = 6.3$ ) in Iran, 2004 in Morocco ( $M = 6.4$ ).

An earthquake early warning system can be effective for mitigating the effects of moderate earthquakes. For moderate size events, early warning systems could also mitigate earthquake effects in terms of infrastructure operability (e.g., hospitals, firehouses, telecommunication hubs, ...) during the post-event emergency phase and rescue operations. For instance, in tall buildings, the higher floors generally sway much more than those near ground level, so that even a moderate earthquake could cause severe damage to a high rise. Therefore, even at 70–80 km distance from its epicenter, a magnitude 6 quake could affect hospital operating rooms and other critical installations.

Installations as close as 50 km from the epicenter could receive an earthquake warning 10 s prior to the arrival of the more energetic waves ( $S$  and surface waves) of an earthquake. To take advantage of this brief warning period, automated systems would have to be created that respond instantly to notification alert signals, and they would have to be carefully calibrated to avoid false or missed alarms. Closer to the epicenter, a magnitude 6 or higher earthquake can damage critical infrastructures, such as telephone lines, gas pipelines, highways, and railroads, as well as airport runways and navigation systems. These disruptions would have a domino effect in more distant areas, which could be mitigated by an early warning alert system, based on the earliest primary wave data to arrive at recording stations close to the epicenter.

Finally, we note that earthquake early warning systems can also help mitigate the effects of such earthquake-induced disasters as fires, explosions, landslides, and tsunamis, which can in many cases be more devastating than the earthquake itself. Systems could be installed at relatively low cost in developing countries, where moderate sized earthquakes can cause damage comparable to that caused by much larger earthquakes in developed countries.

## Bibliography

### Primary Literature

- Allen RM (2007) The ElarmS earthquake early warning methodology and its application across California. In: Gasparini P, Manfredi G, Zschau J (eds) Earthquake early warning systems. Springer, Berlin, pp 21–44. ISBN-13 978-3-540-72240-3
- Allen RM, Kanamori H (2003) The potential for earthquake early warning in Southern California. *Science* 300:786–789. doi:10.1126/science.1080912
- Ambraseys N, Smit P, Douglas J, Margaris B, Sigbjornsson R, Olafsson S, Suhadolc P, Costa G (2004) Internet site for European strong-motion data. *Boll Geofis Teor Appl* 45(3):113–129
- Bakun W, Fischer HF, Jensen E, VanSchaack J (1994) Early warning system for aftershocks. *Bull Seismol Soc Am* 84(2):359–365
- Bernard P, Zollo A (1989) The Irpinia (Italy) 1980 earthquake: detailed analysis of a complex normal fault. *J Geophys Res* 94:1631–1648
- Guidoboni E, Ferrari G, Mariotti D, Comastri A, Tarabusi G, Valensise G (2007) CFT4Med, Catalogue of Strong Earthquakes in Italy (461 B.C.-1997) and Mediterranean Area (760 B.C.-1500). INGV-SGA. Available from <http://storing.ingv.it/cft4med/>
- Bose M, Ionescu C, Wenzel F (2007) Earthquake early warning for Bucharest, Romania: Novel and revised scaling relations. *Geophys Res Lett* 34:L07302. doi:10.1029/2007GL029396
- Bouchon M (1979) Discrete wave number representation of elastic wave fields in three-space dimensions, *J Geophys Res* 84:3609–3614
- Cinti FR, Faenza L, Marzocchi W, Montone P (2004) Probability map of the next  $M \geq 5.5$  earthquakes in Italy. *Geochem Geophys Geosyst* 5:Q1103. doi:10.1029/2004GC000724.
- Convertito V, Iervolino I, Giorgio M, Manfredi G, Zollo A (2008) Prediction of response spectra via real-time earthquake measurements. *Soil Dyn Earthq Eng* 28(6):492–505. doi:10.1016/j.soildyn.2007.07.006
- Cornell CA (1968) Engineering seismic hazard analysis. *Bull Seismol Soc Am* 59(5):1583–1606
- Cua G, Heaton T (2007) The virtual seismologist (VS) method: A Bayesian approach to earthquake early warning. In: Gasparini P, Manfredi G, Zschau J (eds) Earthquake early warning systems. Springer, Berlin. doi:10.1007/978-3-540-72241-0\_7
- Curtis A, Lomax A (2001) Prior information, sampling distributions and the curse of dimensionality. *Geophysics* 66:372–378. doi:10.1190/1.1444928
- Erdik M, Fahjan Y, Ozel O, Alcik H, Mert A, Gul M (2003) Istanbul earthquake rapid response and the early warning system. *Bull Earthquake Eng* 1(1):157–163. doi:10.1023/A:1024813612271
- Espinosa-Aranda JM, Jimenez A, Ibarrola G, Alcantar F, Aguilar A, Inostroza M, Maldonado S (1995) Mexico City seismic alert system. *Seismol Res Lett* 66:42–53
- Font Y, Kao H, Lallemand S, Liu C-S, Chiao L-Y (2004) Hypocentral determination offshore Eastern Taiwan using the maximum intersection method. *Geophys J Int* 158(2):655–675. doi:10.1111/j.1365-246X.2004.02317.x
- Grasso V, Allen RM (2005) Earthquake warning systems: Characterizing prediction uncertainty. *Eos Trans AGU* 86(52), Fall Meet. Suppl., Abstract S44B-03

18. Horiuchi S, Negishi H, Abe K, Kamimura A, Fujinawa Y (2005) An automatic processing system for broadcasting earthquake alarms. *Bull Seism Soc Am* 95(2):708–718. doi:10.1785/0120030133
19. Iervolino I, Convertito V, Giorgio M, Manfredi G, Zollo A (2006) Real time risk analysis for hybrid earthquake early warning systems. *J Earthq Eng* 10(6):867–885
20. Iervolino I, Convertito V, Giorgio M, Manfredi G, Zollo A (2007a) The cry wolf issue in seismic early warning applications for the campania region. In: Gasparini P et al (eds) *Earthquake early warning systems*. Springer, Berlin. doi:10.1007/978-3-540-72241-0\_11
21. Iervolino I, Giorgio M, Manfredi G (2007b) Expected loss-based alarm threshold set for earthquake early warning systems. *Earthq Eng Struc Dyn* 36(9):1151–1168. doi:10.1002/eqe.675
22. Iervolino I, Manfredi G, Cosenza E (2007c) Earthquake early warning and engineering applications prospects. In: Gasparini P, Manfredi G, Szchau J (eds) *Earthquake early warning systems*. Springer, Berlin, doi:10.1007/978-3-540-72241-0\_12
23. Jenny S, Goes S, Giardini D, Kahle H-G (2006) Seismic potential of Southern Italy. *Tectonophysics* 415:81–101. doi:10.1016/j.tecto.2005.12.003
24. Kamigaichi O (2004) JMA earthquake early warning. *J Japan Assoc Earthq Eng* 3:134–137
25. Kanamori H (2005) Real-time seismology and earthquake damage mitigation. *Ann Rev Earth Planet Sci* 33:195–214. doi:10.1146/annurev.earth.33.092203.122626
26. Lancieri M, Zollo A (2008) A bayesian approach to the real-time estimation of magnitude from the early *P*- and *S*-wave displacement peaks. *J Geophys Res*. doi:10.1029/2007JB005386, in press
27. Lomax A (2005) A Reanalysis of the hypocentral location and related observations for the great 1906 California earthquake. *Bull Seism Soc Am* 95(3):861–877. doi:10.1785/0120040141
28. Meletti C, Patacca E, Scandone P (2000) Construction of a seismotectonic model: The case of Italy. *Pure Appl Geophys* 157:11–35
29. Montone P, Mariucci MT, Pondrelli S, Amato A (2004) An improved stress map for Italy and surrounding regions (central Mediterranean). *J Geophys Res* 109:B10410. doi:10.1029/2003JB002703
30. Munich Re (eds) (2005) *Environmental report – perspectives – Today's ideas for tomorrow's world*, WKD-Offsetdruck GmbH, München
31. Nakamura Y (1988) On the urgent earthquake detection and alarm system (UrEDAS). *Proc 9th World Conf Earthquake Eng VII*, Toyko, 673–678
32. Nakamura Y (1989) Earthquake alarm system for Japan railways. *Japan Railway Eng* 109:1–7
33. Nakamura Y (2004) Uredas, urgent earthquake detection and alarm system, now and future. *13th World Conference on Earthquake Engineering* 908
34. Okada T et al (2003) A new method of quickly estimating epicentral distance and magnitude from a single seismic record. *Bull Seismol Soc Am* 93(1):526–532. doi:10.1785/0120020008
35. Podvin P, Lecomte I (1991) Finite difference computations of traveltimes in very contrasted velocity models: a massively parallel approach and its associated tools. *Geophys. J Int.* 105:271–284
36. Rydelek P, Pujol J (2004) Real-time seismic warning with a 2-station subarray. *Bull Seism Soc Am* 94(4):1546–1550. doi:10.1785/012003197
37. Satriano C, Lomax A, Zollo A (2008) Real-time evolutionary earthquake location for seismic early warning. *Bull Seism Soc Am* 98(3):1482–1494. doi:10.1785/0120060159
38. Simons F, Dando JB, Allen R (2006) Automatic detection and rapid determination of earthquake magnitude by wavelet multiscale analysis of the primary arrival. *Earth Planet Sci Lett* 250:214–223. doi:10.1016/j.epsl.2006.07.039
39. Teng TL, Wu Y-M, Shin TC, Tsai YB, Lee WHK (1997) One minute after: strong-motion map, effective epicenter, and effective magnitude. *Bull Seism Soc Am* 87(5):1209–1219
40. Tsukada S (2006) Earthquake early warning system in Japan. *Proc 6th Joint Meeting UJNR Panel on Earthquake Research*, Tokushima, Japan
41. Valensise G, Amato A, Montone P, Pantosti D (2003) Earthquakes in Italy: Past, present and future. *Episodes* 26(3):245–249
42. Wenzel FM et al (1999) An early warning system for Bucharest. *Seismol Res Lett* 70(2):161–169
43. Westaway R, Jackson J (1987) The earthquake of 1980 November 23 in Campania-Basilicata (southern Italy). *Geophys J R Astron Soc* 90:375–443. doi:10.1111/j.1467-6435.1999.tb00581.x
44. Wu Y-M, Teng T (2002) A virtual subnetwork approach to earthquake early warning. *Bull Seismol Soc Am* 92(5):2008–2018. doi:10.1785/0120040097
45. Wu Y-M, Kanamori H (2005) Experiment on onsite early warning method for the Taiwan early warning system. *Bull Seismol Soc Am* 95(1):347–353. doi:10.1785/0120040097
46. Wu YM, Zhao L (2006) Magnitude estimation using the first three seconds of *p*-wave amplitude in earthquake early warning. *Geophys Res Lett* 33:L16312. doi:10.1029/2006GL026871
47. Wurman G, Allen RM, Lombard P (2007) Toward earthquake early warning in Northern California. *J Geophys Res* 112:B08311. doi:10.1029/2006JB004830
48. Zhou H (1994) Rapid 3-D hypocentral determination using a master station method. *J Geophys Res* 99(B8):15439–15455
49. Zollo A, Lancieri M, Nielsen S (2006) Earthquake magnitude estimation from peak amplitudes of very early seismic signals on strong motion records. *Geophys Res Lett* 33:L23312. doi:10.1029/2006GL027795
50. Zollo A, Lancieri M, Nielsen S (2007) Reply to comment by P. Rydelek et al on "Earthquake magnitude estimation from peak amplitudes of very early seismic signals on strong motion records". *Geophys Res Lett* 34:L20303. doi:10.1029/2007GL030560.

## Books and Reviews

- Berger JO (1985) *Statistical decision theory and Bayesian analysis*. Springer, New York
- Coutant O (1989) *Program de simulation numerique AXITRA*. Rapport LGIT, Grenoble, France
- Gruppo di Lavoro MPS (2004) *Redazione della mappa di pericolosità sismica prevista dall'Ordinanza PCM 3274 del 20 marzo (2003) Rapporto Conclusivo per il Dipartimento della Protezione Civile, INGV, Milano-Roma, aprile (2004) 65 pp + 5 appendici*
- Milne J (1886) *Earthquakes and other earth movements*. Appelton, New York, p 361

## Earthquake Engineering, Non-linear Problems in

MIHAILO D. TRIFUNAC  
Department of Civil Engineering,  
University of Southern California, Los Angeles, USA

### Article Outline

Glossary  
Definition of the Subject  
Introduction  
Vibrational Representation of Response  
Response in Terms of Wave Propagation –  
An Example  
Observations of Nonlinear Response  
Future Directions  
Bibliography

### Glossary

**Meta-stability of man-made structures** is the consequence of their upright construction above ground. For excessive dynamic (earthquake) loads, when the lateral deflection exceeds some critical value (this is normally accompanied by softening nonlinear behavior of the structural members), the overturning moment of the gravity forces becomes larger than the restoring moment, and the structure becomes unstable and moves exponentially toward collapse.

**Complex and evolving structural systems** are structures with a large number of degrees of freedom and many structural members, which for given loads experience softening nonlinear deformations. During strong excitation, continuous changes (typically decreases) in effective stiffness and time-dependent changes in boundary conditions result in a system whose properties are changing with time.

**Soil–structure interaction** is a process in which the soil and the structure contribute to mutual deformations while undergoing dynamic response. In time, with continuously changing contact area between the foundation and the soil (opening and closing of gaps), when the deformations are large, soil–structure interaction is characterized by nonlinear geometry and nonlinear material properties in both the soil and in the structure.

### Definition of the Subject

Nonlinear problems in structural earthquake engineering deal with the dynamic response of meta-stable, man-

made buildings subjected to strong earthquake shaking. During earthquakes, structures constructed on soft sediments and soils deform together with the underlying soil in the dynamic process called soil–structure interaction. Strong shaking forces the soil–structure systems to evolve through different levels of nonlinear response, with continuously changing properties that depend upon the time history of excitation and on the progression and degree of damage. Thus far, the analyses of this response have used the vibrational approach and lumped mass discrete models to represent real structures. Loss of life and property, however, continue to be high during strong shaking in the vicinity of the faults responsible for earthquakes. This calls for new, more physically refined methods of analysis, which can be based on nonlinear wave propagation, and for balancing of the structural capacities with the power carried by the earthquake waves.

After a brief discussion of the literature on the complex and chaotic dynamics of simple mechanical oscillators, the dynamic characteristics and governing equations in the meta-stable structural dynamics of earthquake engineering are introduced. The nature of the solutions of the governing equations in terms of both the vibrational and the wave representations is discussed, and the dynamic instability, material and geometric nonlinearities, and complexities of the governing equations associated with nonlinear soil–structure interaction are described. Collectively, the examples presented reflect the complex physical nature of meta-stable structural systems that experience nonlinear dynamic response, the characteristics of which change and evolve during earthquake excitation.

### Introduction

Earthquake engineering, through a cooperation of structural and geotechnical engineers with seismologists and geologists, aims to develop methods for safer design of man-made structures to withstand shaking near intermediate and large earthquakes. This requires addressing the problems of predictability of the response of complicated nonlinear systems, which is one of the important subjects of modern nonlinear science. Through the studies of the dynamic response, earthquake engineers address complex physical problems and issues with important social implications.

The completeness and beauty of the linear differential equations appear to have led to their dominance in the mathematical training of engineers and scientists during most of the 20th century. The recognition that chaotic dynamics is inherent in all nonlinear physical phenomena, which has created a sense of revolution in applied me-



chanics and physics today, so far has had little if any effect on the research and design of earthquake-resistant structures. In the past, the designs in structural engineering and control systems were kept within the realm of linear system dynamics. However, the needs of modern technology have pushed the design into the nonlinear regimes of large deformations, which has increased the possibility of encountering chaotic dynamic phenomena in structural response. Even a cursory review of papers on chaotic vibrations in mechanical systems leads to the conclusion that chaotic dynamics is not a small, insignificant class of motions and that chaotic oscillations occur in many nonlinear systems and for a wide range of values of the parameters.

If an engineer chooses parameters that produce chaotic output, then he or she loses predictability. However, the chaotic behavior of nonlinear systems does not exclude predictability of the response but rather introduces upper bounds (prediction horizons) [31] and renders the predictions probabilistic. The important question is then over what time-scale are the forecasts reliable, given the current state and knowledge of the system. Another key ingredient for prediction is an adequate physical model. At present, because of the multitude of interacting phenomena and the absence of physically complete equations of motion, there exists no adequate general model of the complete earthquake response process. While the practical outcome of most work in earthquake engineering remains empirically based, the nonlinear methods are gaining popularity, aiming to decipher the governing phenomena and to assess the reliability of the models. It appears now that the broad-based revolution in the worldview of science that begun in the twentieth century will be associated with chaotic dynamics [43]. This revolution should eventually also contribute to better understanding and more complete representation of the response analyses in earthquake engineering.

It has been argued that major changes in science occur not so much when new theories are advanced but when the simple models with which scientists conceptualize a theory are changed [24]. In vibrations, such a conceptual model that embodies the major features of a whole class of problems is the spring-mass system. Lessons emerging from studies of the spring-mass model and several other relevant models can serve as conceptual starting points for generalizations and also as a guide to further studies of more complex models in earthquake engineering and structural dynamics.

Studies of forced vibrations of a pendulum have revealed complex dynamics and chaotic vibrations [14,15]. A simply supported beam with sub-buckling axial com-

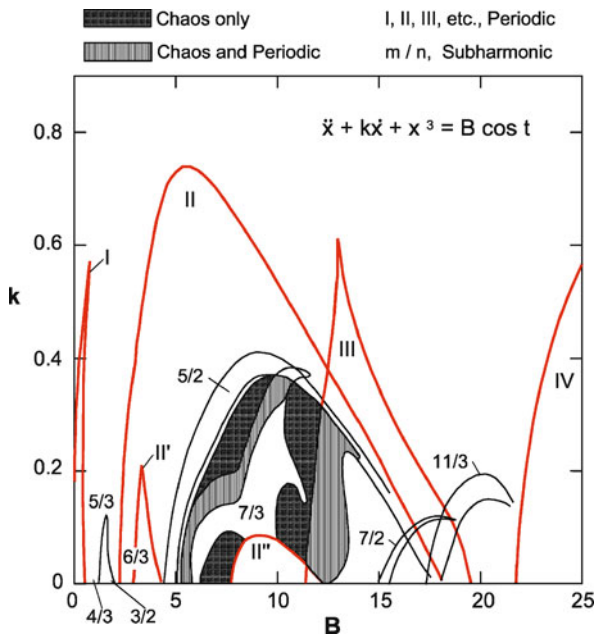
pression modeled by a single mode approximation yields a Mathieu type equation and for certain values of the parameters leads to unstable solutions. When nonlinearities are added, these vibrations result in a limit cycle. A related problem is a classical pendulum with a vibrating pivot support, which also leads to chaotic vibrations [29,34]. Chaotic motions in a double pendulum have been studied by Richter and Scholz [45], and the complex dynamics and chaotic solutions for a spherical pendulum with two degrees of freedom have been described by Miles [35].

Impact-type problems result in explicit difference equations or maps, which can yield chaotic vibrations for certain values of the governing parameters [30]. A mass vibrating in a gap between two stiff springs on either side [17,46,47] is a simple related model, which suggests a starting point for research in nonlinear vibration of piles, and for impact-type interaction of adjacent buildings, excited by strong earthquake ground motion. The reader can find examples of such problems in the description of damage in Mexico City, for example, during several earthquakes [32].

Chaotic motions of an elasto-plastic arch have been studied by Poddar et al. [42]. Forced vibrations of a buckled beam, modeled by the Duffing equation, showed that chaotic vibrations are possible [16]. Forced vibrations described by a Duffing equation with viscous damping and nonlinear (cubic) elastic (stiffening) spring were studied by Ueda [67]. Figure 1 summarizes his results and describes the regions of chaotic, periodic (I, II, etc.), and subharmonic (m/n) motions as functions of the damping and forcing amplitudes. This simple equation, representing a hardening spring system, has direct analogues in the dynamics of piles and in the rocking of buildings, both following the strong-motion phase of earthquake shaking after horizontal gaps have been created between the pile (foundation walls) and the soil [63].

A mechanical system with a nonlinear restoring force and with a control force added to move the system according to some prescribed signal has been studied by Holmes and Moon [19] and Holmes [18]. It was shown that such a system exhibits both periodic limit-cycle oscillation and chaotic motions. Chaotic vibrations in continuous beams have been studied for nonlinear body forces and nonlinear boundary conditions (that depend on the motion), and for motions large enough for the nonlinear terms in the equations of motion to be significant [37,38,39,40,41]. Forced planar vibrations of nonlinear elastica [35,36], were shown to become unstable and exhibit chaotic motions under certain conditions.

The above-mentioned studies imply that there is a conflict in the classical engineering description of the world.



Earthquake Engineering, Non-linear Problems in, Figure 1  
Chaos diagram showing regions of chaotic, chaotic and periodic, periodic (I, II, III, etc.), and sub-harmonic ( $4/3$ ,  $3/2$ ,  $5/3$ , etc.) motions for a nonlinear equation as functions of non-dimensionalized damping and forcing amplitude (from [67])

One aspect of this conflict is the assumption that nature is a deductive system, moving forward in time according to deterministic laws. Another aspect is that a scientist attempting to model portions of the world from finite data projects unverifiable structure onto the local environment. The conflict is that these two views do not match, leaving us with a question: what are models good for? There are many systems in nature that are observed to be chaotic, and for which no adequate physical model exists. Whether a model is adequate or not depends, of course, on the questions asked [7]. Unfortunately, the art of dynamical modeling is often neglected in discussions of nonlinear and chaotic systems, in spite of its crucial importance [1]. In the following, the modeling problem in earthquake engineering will be illustrated using two common approaches to the solution, one based on an equivalent oscillator and the other one using wave representation.

Stochastic processes have been developed to describe irregular phenomena in deterministic systems that are too complicated or have too many variables to be fully described in detail. For example, stochastic processes have been used to model the response of structures to earthquake and wind forces, which are deterministic, and in principle could be completely described. In practice, the

stochastic modeling has been used also as an approximate description of a deterministic system that has unknown initial conditions and may be highly sensitive to the initial conditions. In trying to model real systems, as a result of the modeling process, we sometimes obtain a model that shows very regular behavior, while the real system has very irregular behavior. In that case, random noise is added to the model, but this represents no more than our lack of knowledge of the system structure or the inadequacy of the identification procedure [22].

In earthquake engineering, the complexity of the multi-dimensional real world is reduced to a sub-space, which is defined by (1) the dimensions and properties of the adopted mathematical models, (2) the nature of the adopted boundary conditions, and (3) the method of solution. A linear mechanical system cannot exhibit chaotic vibrations, and for periodic inputs it produces periodic outputs. The chaotic system must have nonlinear elements or properties, which can include, for example, (1) nonlinear elastic or spring elements; (2) nonlinear damping (such as stick-slip friction); (3) backlash, play, or bilinear springs; and (4) nonlinear boundary conditions. The nonlinear effects can be associated with the material properties, with the geometric effects, or both. In the following, the consequences of unorthodox boundary conditions and nonlinear waves in a building will be used to illustrate the extensions and complexities associated with evolving systems. The utility of this complexity can be viewed as the arbiter of the order and randomness.

### Vibrational Representation of Response

The first modern uses of mechanics in problems of earthquake engineering appeared during the early 1900s, following the earthquake disasters in San Francisco (1906), Messina-Reggio (1908), and Tokyo (1923) and the realization that something needed to be done to prevent such losses of life and property during future events. The first practical steps consisted of introducing the *seismic coefficient* (*shindo* in Japan, and *rapporto sismico* in Italy). This was followed by earthquake-resistant design codes, first adopted in Japan in 1923, and then in California in 1934 [44]. During the same period, there also appeared the first studies of the effects of earthquake shaking on structures in terms of simple mechanical oscillators [48], and in the early 1930s the modern theory based on the response spectrum method was introduced [2,3,4]. These early developments follow the deterministic formulations of Newtonian mechanics and employ linear models and equations of motion.

### Elementary Vibrational Representation of Response

The basic model employed to describe the response of a simple structure to only horizontal earthquake ground acceleration,  $\ddot{\Delta}_x$ , is a single-degree-of-freedom system (SDOF) that experiences rocking  $\psi_r$  relative to the normal to the ground surface. The model also assumes that the ground does not deform in the vicinity of the foundation—that is, it neglects the soil–structure interaction (Fig. 2). The rotation  $\psi_r$  is restrained by a spring with stiffness  $K_r$  and by a dashpot with rocking damping constant  $C_r$ , providing the fraction of critical damping  $\zeta_r$ . The natural frequency of this system is  $\omega_r = (K_r/h^2 m_b)^{1/2}$ , and for small rocking angles it is governed by the linear ordinary differential equation

$$\ddot{\psi}_r + 2\omega_r \zeta_r \dot{\psi}_r + \omega_r^2 \psi_r = -\ddot{\Delta}_x/h. \quad (1)$$

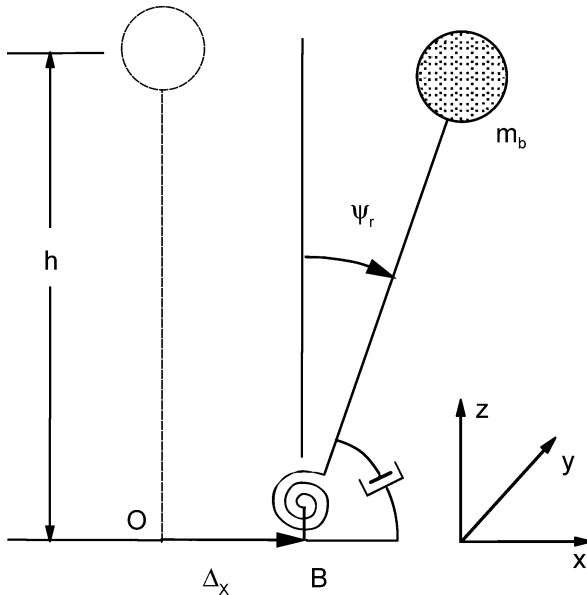
For any initial conditions, and for arbitrary excitation, this system always leads to a deterministic and predictable response. Equation (1) was used originally to develop the concept of relative response spectrum and continues to this day as the main vehicle in formulation of most earthquake engineering analyses of response [56]. If the gravity force is considered,  $\omega_r$  in Eq. (1) has to be reduced [5]. The system described by Eq. (1) is meta-stable for  $\psi_r$  smaller than its critical value. At the critical value of  $\psi_r$ , the over-

turning moment of the gravity force is just balanced by the elastic moment in the restraining spring, and for values greater than the critical value the system becomes unstable.

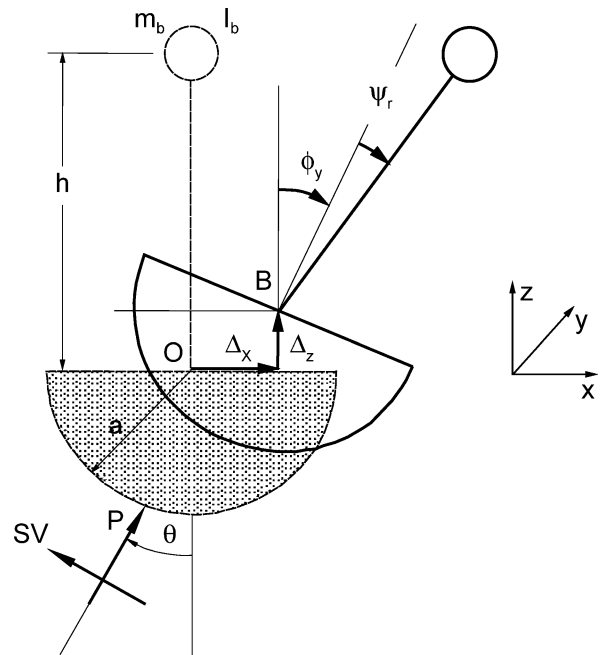
### Advanced Vibrational Representation of Response

In more advanced vibrational representations of the response, additional components of the earthquake excitation, structural dynamic instability, soil–structure interaction, spatial and temporal variations of the excitation, differential motions at different support points, and non-linear behavior of the stiffness  $K_r$  can be considered, but the structure usually continues to be modeled by mass-less columns, springs, and dashpots, and with a rigid mass  $m_b$ . In the following, we illustrate some of the above-mentioned cases.

**Dynamic Instability** An example of a simple model that includes instability is shown in Fig. 3. It experiences horizontal, vertical, and rocking excitations, which can result, for example, from incident P and SV waves. The structure



Earthquake Engineering, Non-linear Problems in, Figure 2  
Single-degree-of-freedom system (SDOF) representation of a building (inverted pendulum) with equivalent mass  $m_b$  and mass-less column of height  $h$ , experiencing rocking  $\psi_r$  due to horizontal motion of its base  $\Delta_x$



Earthquake Engineering, Non-linear Problems in, Figure 3  
Single-degree-of-freedom system (SDOF) representation of a building (inverted pendulum), with equivalent mass  $m_b$ , moment of inertia (about  $O$ )  $I_b$ , and a mass-less column of height  $h$ , experiencing relative rocking  $\psi_r$  due to horizontal, vertical, and rocking motions of its foundation ( $\Delta_x$ ,  $\Delta_z$ , and  $\phi_y$ ), which result from soil–structure interaction when excited by incident wave motion

is represented by an equivalent single-degree-of-freedom system, with a concentrated mass  $m_b$  at height  $h$  above the foundation. It has a radius of gyration  $r_b$  and a moment of inertia  $I_b = m_b r_b^2$  about point O. The degree-of-freedom in the model is chosen to correspond to the relative rocking angle  $\psi_r$ . This rotation is restrained by a spring with rocking stiffness  $K_r$  and by a dashpot with rocking damping  $C_r$  (both not shown in Fig. 3), and the gravitational force  $m_b g$  is considered. Taking moments about B results in the equation of motion

$$\ddot{\phi}_y + \ddot{\psi}_r + 2\omega_r \zeta_r \dot{\psi}_r + \omega_r^2 \psi_r = \{- (\ddot{\Delta}_x/a) \cos(\phi_y + \psi_r) + (\omega_r^2 \varepsilon_g + \ddot{\Delta}_z/a) \sin(\phi_y + \psi_r)\} / \varepsilon, \quad (2)$$

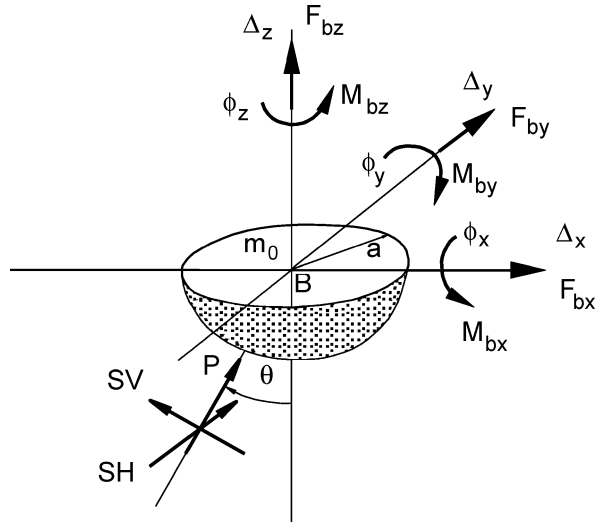
where  $\varepsilon = h(1 + (r_b/h)^2)/a$ ,  $\omega_r^2 = K_r/[m(h^2 + r_b^2)]$ ,  $\omega_r$  is the natural frequency of rocking,  $\zeta_r$  is a fraction of critical damping in  $2\omega_r \zeta_r = C_r/[m(h^2 + r_b^2)]$ , and  $\varepsilon_g = 2/\omega_r^2 a$ . Equation (2) is a differential equation coupling the rocking of the foundation,  $\phi_y$ , and of the structure,  $\psi_r$ , with the horizontal and vertical motions of the foundation. It is a nonlinear equation the solution to which requires numerical analysis. In this example, we will discuss only the case in which  $\phi_y + \psi_r$  is small. Then,

$$\ddot{\psi}_r + 2\omega_r \zeta_r \dot{\psi}_r + \{\omega_r^2(1 - \varepsilon_g/\varepsilon) - \ddot{\Delta}_z/\varepsilon a\} \psi_r = -\ddot{\phi}_y + \{-\ddot{\Delta}_x/a + (\omega_r^2 \varepsilon_g + \ddot{\Delta}_z/a) \phi_y\} / \varepsilon. \quad (3)$$

For steady-state excitation by incident P and SV waves with frequency  $\omega$ ,  $\Delta_x$ ,  $\phi_y$ , and  $\Delta_z$ , and therefore the forcing function of Eq. (3), will be periodic. Equation (3) is then a special form of the Hill's equation. Analysis of the stability of this equation can be found in the work of Lee [25]. For general earthquake excitation,  $\Delta_x$ ,  $\phi_y$ , and  $\Delta_z$  will be determined by the recorded components of motion, and in predictive analyses by simulated ground motions [27,28,70].

In Eq. (3),  $\phi_y$  describes rocking of the foundation to which the structure is attached. In analyses that do not consider soil-structure interaction,  $\phi_y$  will be determined directly by the rocking component of strong ground motion [21,28], and in studies that consider soil-structure interaction  $\phi_y$  will be one of the variables to be determined by the analysis [25].

**Soil-Structure Interaction** The problem of linear soil-structure interaction embodies the phenomena that result from (1) the presence of an inclusion (foundation, Fig. 4) in the soil [26], and (2) the vibration of the structure supported by the foundation, which exerts dynamic forces on the foundation [25]. Examples and a discussion of the non-



Earthquake Engineering, Non-linear Problems in, Figure 4 Six components of motion (three translations and three rotations)  $\{\Delta_x, \Delta_y, \Delta_z, \phi_x, \phi_y, \phi_z\}$  of point B, and six components of force (three forces and three moments)  $\{F_{ext}\} = \{F_{bx}, F_{by}, F_{bz}, M_{bx}, M_{by}, M_{bz}\}$ , that the structure exerts on the foundation at B

linear aspects of soil-structure interaction can be found in Gicev [9] and in a review of observations of response to earthquake shaking in full-scale structures in Trifunac et al. [63,64,65].

The dynamic response of a rigid, embedded foundation to seismic waves can be separated into two parts. The first part corresponds to the determination of the restraining forces due to the motion of the inclusion, usually assumed to be a rigid body. The second part deals with the evaluation of the driving forces due to scattering of the incident waves by the inclusion, which is presumed to be immobile. This can be illustrated by considering a foundation embedded in an elastic medium and supporting an elastic superstructure. The steady-state harmonic motion of the foundation having frequency  $\omega$  can be described by a vector  $\{\Delta_x, \Delta_y, \Delta_z, \phi_x, \phi_y, \phi_z\}^T$  (Fig. 4), where  $\Delta_x$  and  $\Delta_y$  are horizontal translations,  $\Delta_z$  is vertical translation,  $\phi_x$  and  $\phi_y$  are rotations about horizontal axes, and  $\phi_z$  is torsion about the vertical axis. Using superposition, displacement of the foundation is the sum of two displacements:

$$\{U\} = \{U^*\} + \{U_0\}, \quad (4)$$

where  $\{U^*\}$  is the foundation input motion corresponding to the displacement of the foundation under the action of the incident waves in the absence of external forces, and  $\{U_0\}$  is the relative displacement corresponding to the dis-

placement of the foundation under the action of the external forces in the absence of incident wave excitation.

The interaction force  $\{F_s\}$  generates the relative displacement  $\{U_0\}$ , which corresponds to the force that the foundation exerts on the soil and that is related to  $\{U_0\}$  by  $\{F_s\} = [K_s(\omega)]\{U_0\}$ , where  $[K_s(\omega)]$  is the  $6 \times 6$  complex stiffness matrix of the embedded foundation. It depends upon the material properties of the soil medium, the characteristics and shape of the foundation, and the frequency of the harmonic motion, and it describes the force-displacement relationship between the rigid foundation and the soil medium.

The driving force of the incident waves is equal to  $\{F_s^*\} = [K_s]\{U^*\}$ , where the input motion  $\{U^*\}$  is measured relative to an inertial frame. The “driving force” is the force that the ground exerts on the foundation when the rigid foundation is kept fixed under the action of the incident waves. It depends upon the properties of the foundation and the soil and on the nature of excitation.

The displacement  $\{U\}$  is related to the interaction and driving forces via  $[K_s]\{U\} = \{F_s\} + \{F_s^*\}$ . For a rigid foundation having a mass matrix  $[M_0]$  and subjected to a periodic external force,  $\{F_{ext}\}$ , the dynamic equilibrium equation is

$$[M_0]\{\ddot{U}\} = -\{F_s\} + \{F_{ext}\}, \tag{5}$$

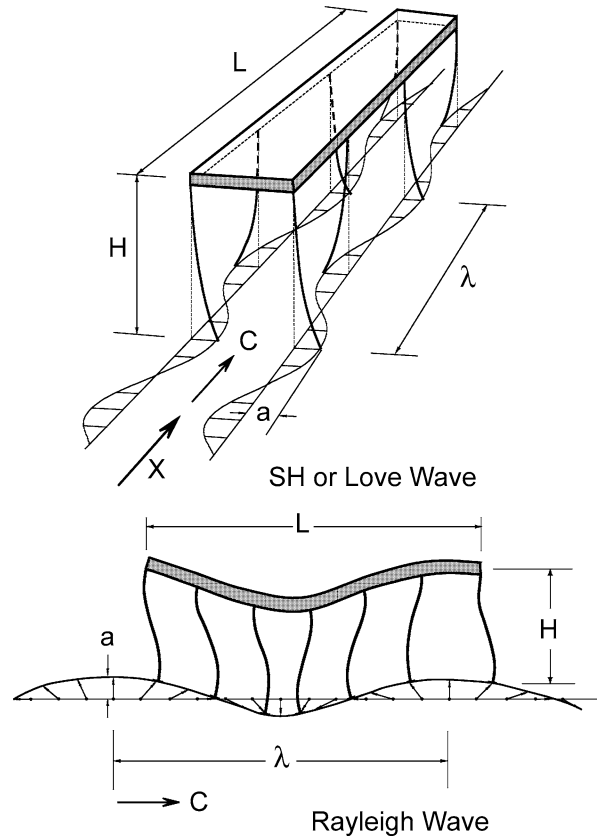
where  $\{F_{ext}\} = \{F_{bx}, F_{by}, F_{bz}, M_{bx}, M_{by}, M_{bz}\}$  is the force the structure exerts on the foundation (Fig. 4). Then, Eq. (5) becomes

$$[M_0]\{\ddot{U}\} + [K_s]\{U\} = \{F_s^*\} + \{F_{ext}\}. \tag{6}$$

The solution of  $\{U\}$  requires the determination of the mass matrix, the impedance matrix, the driving forces, and the external forces [25].

After the mass matrix  $[M_0]$ , the stiffness matrix  $[K_s]$ , and the force  $\{F_s^*\}$  have all been evaluated, they can be used to determine the foundation displacement  $\{U\}$ . For in-plane response excited by P and SV waves, for example, the relative response  $\psi_r$  is then given by Eq. (3).

**Differential Motions** Common use of the response spectrum method [56] and many dynamic analyses in earthquake engineering implicitly assume that all points of building foundations move synchronously and with the same amplitudes. This, in effect, implies that the wave propagation in the soil is neglected. Unless the structure is long (e. g., a bridge with long spans, a dam, a tunnel) or “stiff” relative to the underlying soil, these simplifications are justified and can lead to a selection of approximate design forces if the effects of soil-foundation interaction in the presence of differential ground motions can be



Earthquake Engineering, Non-linear Problems in, Figure 5 Schematic representation of the deformation of columns accompanying differential wave excitation of long structures for out-of-plane response (top) and in-plane response (bottom) when SH or Love waves (top) or Rayleigh waves (bottom) propagate along the longitudinal axis of a structure

neglected [6]. Simple analyses of two-dimensional models of long buildings suggest that when  $a/\lambda < 10^{-4}$ , where  $a$  is wave amplitude and  $\lambda$  is the corresponding wavelength, the wave propagation effects on the response of simple structures can be neglected [50].

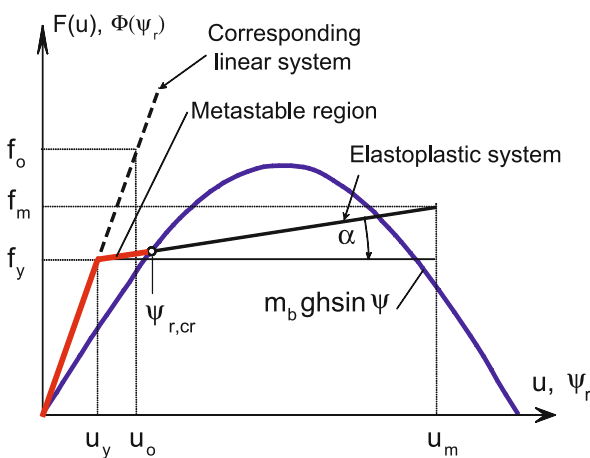
Figure 5 illustrates the “short” waves propagating along the longitudinal axis of a long building or a multiple-span bridge. For simplicity, the incident wave motion has been separated into out-of-plane motion (Fig. 5, top), consisting of SH and Love waves, and in-plane motion (Fig. 5, bottom) consisting of P, SV, and Rayleigh waves. The in-plane motion can further be separated into horizontal (longitudinal), vertical, and rocking components, while out-of-plane motion consists of horizontal motion in the transverse direction and torsion along the vertical axis. Trifunac and Todorovska [61] analyzed the effects of the horizontal in-plane components of differential motion

for buildings with models that are analogous to the sketch in Fig. 5 (bottom), and they showed how the response spectrum method can be modified to include the first-order effects of differential motions. Trifunac and Gicev [59] showed how to modify the spectra of translational motions, into a spectrum that approximates the total (translational and torsional) responses, and how this approximation is valid for strong motion waves an order of magnitude longer than the structure ( $\lambda \gg L$ ).

As can be seen from the above examples the differential motions lead to complex excitation and deformation of the structural members (columns, shear walls, beams, braces), increase the dimensions of the governing differential equations, lead to three-dimensional dynamic instability problems, and can lead to nonlinear boundary conditions. These are all conditions that create an environment in which, even with the most detailed numerical simulations, it is difficult to predict all of the complexities of the possible responses.

### Nonlinear Vibrational Analyses of Response

For engineering estimation of the maximum nonlinear response of a SDOF system,  $u_m$ , in terms of the maximum linear response,  $u_0$ , it is customary to specify a relation between  $u_m$  and  $u_0$  (Fig. 6). By defining the yield-strength reduction factor as  $R_y = u_0/u_y$ , where  $u_y$  is the yielding displacement of the SDOF system equivalent spring, and ductility as  $\mu = u_m/u_y$ , for the same ground motion the ratio  $u_m/u_0$  is then equal to  $\mu/R_y$ . Veletsos and



Earthquake Engineering, Non-linear Problems in, Figure 6 Bi-linear representation of stiffness (yielding at  $(u_y, f_y)$ ), overturning moment of gravity force ( $m_b g h \sin \psi$ ), critical rocking angle  $\psi_{r,cr}$ , and meta-stable region ( $0 < \psi_r < \psi_{r,cr}$ ) for an SDOF system

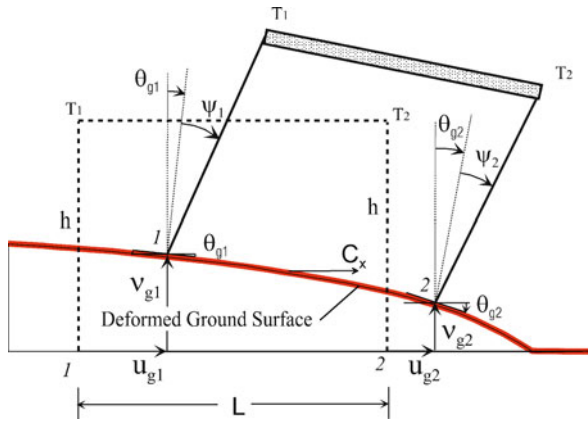
Newmark [68,69] showed that (1) for a long-period SDOF system when its natural period  $T_n = 2\pi/\omega_n$  becomes very long,  $u_m / u_0$  tends toward 1 and  $R_y$  approaches  $\mu$  (equal deformation rule); (2) for the response amplitudes governed mainly by the peak excitation velocities,  $u_m / u_0$  can be approximated by  $\mu/\sqrt{2\mu - 1}$  and  $R_y$  by  $\sqrt{2\mu - 1}$  (equal strain energy rule); and (3) for a high-frequency (stiff) system when  $T_n \sim 0$ ,  $R_y \sim 1$ .

**Complexities of Simultaneous Action of Dynamic Instability, Nonlinearity, and Kinematic Boundary Conditions – Example** The model we illustrate next is an SDOF when it is excited by synchronous horizontal ground motion at its two supports (1 and 2 in Fig. 7), but it behaves like a three-degree-of-freedom (3DOF) system when excited by propagating horizontal, vertical, and rocking ground motions. For such a system, the above classical equal energy and equal displacement rules for SDOF system will not apply.

The goals here are to describe the effects of differential motion on strength-reduction factors  $R_y$  of the simple structure shown in Fig. 7 when it is subjected to all of the components of near-source ground motions, and to illustrate the resulting complexities of nonlinear response. Analyses of the consequences of the differences in ground motion at structural supports, caused by non-uniform soil properties, soil-structure interaction, and lateral spreading, for example, will further contribute to the complexities of the response, but these factors will not be discussed here.

The original response spectrum method was formulated using a vibrational solution of the differential equation of an SDOF system excited by synchronous, and only horizontal (one component), ground motion. The consequences of simultaneous action of all six components of ground motion (three translations and three rotations) on the relative response of an SDOF system are still rarely considered in modern engineering design [58], even though it has been 75 years since the original response spectrum method was formulated and about 40 years since it became the principal tool in engineering design [56]. Because the response spectrum method has become an essential part of the design process and of the description of how strong motion should be specified for a broad range of design applications [52], we hope that the present examples will help to further understanding of the complexities of response in more realistic models of structures.

The nature of the relative motion of individual column foundations or of the entire foundation system will depend upon the type of foundation, the characteristics of the soil surrounding the foundation, the type of incident



Earthquake Engineering, Non-linear Problems in, Figure 7  
**The structure deformed by the wave, propagating from left to right, with phase velocity  $C_x$ , for the case of  $+v_{g_i}$  ("up" motion). Different column rotations  $\psi_1$  and  $\psi_2$  result from different translations and rotations at supports 1 and 2 (from [21])**

waves, and the direction of wave arrival, with the motion at the base of each column having six degrees of freedom. In the following example, we assume that the effects of soil–structure interaction are negligible; consider only the in-plane horizontal, vertical, and rocking components of the motion of column foundations; and show selected results of the analysis for a structure on only two separate foundations. We assume that the structure is near the fault and that the longitudinal axis of the structure ( $X$  axis) coincides with the radial direction ( $r$  axis) of the propagation of waves from the earthquake source, so that the displacements at the base of columns are different as a result of the wave passage alone. We suppose that the excitations at the piers have the same amplitude but different phases and that the phase difference (or time delay) will depend upon the distance between the piers and the horizontal phase velocity of the incident waves.

The simple model we consider, which is described in Fig. 7, represents a one-story structure consisting of a rigid mass,  $m$ , with length  $L$ , supported by two rigid, mass-less columns with height  $h$ , which are connected at the top to the mass and at the bottom to the ground by rotational springs (not shown in Fig. 7). The stiffness of the springs,  $k_\phi$ , is assumed to be elastic-plastic, as in Fig. 6, without hardening ( $\alpha = 0$ ). The mass-less columns are connected to the ground and to the rigid mass by rotational dashpots,  $c_\phi$ , providing a fraction of critical damping equal to 5 percent. Rotation of the columns,  $\phi_i = \theta_{g_i} + \psi_i$  for  $i = 1, 2$ , which is assumed to be not small, leads us to consider the geometric nonlinearity. The mass is acted upon by the acceleration of gravity,  $g$ , and is excited by differential hori-

zontal, vertical, and rocking ground motions,  $u_{g_i}$ ,  $v_{g_i}$ , and  $\theta_{g_i}$ ,  $i = 1, 2$  (Fig. 7) at the two bases, so that

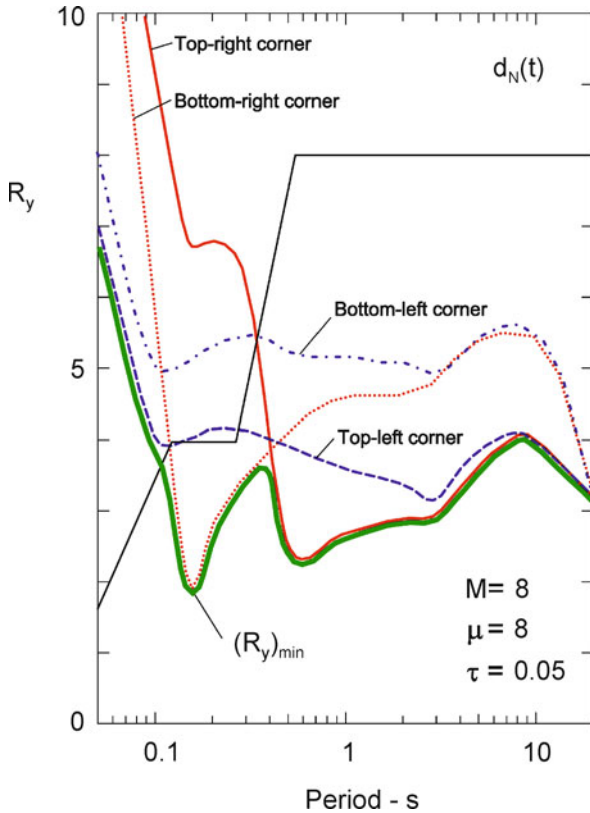
$$u_{g_2}(t) = u_{g_1}(t - \tau); \quad v_{g_2}(t) = v_{g_1}(t - \tau);$$

$$\theta_{g_2}(t) = \theta_{g_1}(t - \tau); \quad \tau = L/C_x,$$

with  $\tau$  being the time delay between the motions at the two piers and  $C_x$  the horizontal phase velocity of the incident waves. The functional forms of  $u_{g_i}$ ,  $v_{g_i}$ , and  $\theta_{g_i}$  are defined by the near-source ground motions [21], and the rocking component of the ground motion is approximated by [28]  $\theta_{g_i}(t) = -\dot{v}_{g_i}(t)/C_x$ , where  $\dot{v}_{g_i}(t)$  is the vertical velocity of the ground motion at the  $i$ th column. Of course, in a more accurate modeling, the ratio of the  $v_{g_i}$  to  $u_{g_i}$  amplitudes will depend upon the incident angle and the character of incident waves, while the associated rocking  $\theta_{g_i}$  will be described by a superposition of the rocking angles associated with incident body and dispersed surface waves [28].

The yield-strength reduction factor for the system subjected to synchronous ground motion is  $R_y = f_0/f_y = u_0/u_y$ , where all of the quantities are defined in Fig. 6. In this example, for the assumed model and because of the differential ground motions and rotation of the beams, the relative rotation for the two columns at their top and bottom will be different. Therefore, it is necessary to define the  $R$ -factor and ductility for each corner of the system, instead of one factor for the entire system. In all calculations here, we consider the actions of the horizontal, vertical, and rocking components of the ground motion, the effects of gravity force, dynamic instability, and geometric nonlinearity. For the structure in Fig. 7, we calculate maximum linear and nonlinear relative rotations at four corners of the system under downward ( $-v_{g_i}$ ), radial, and rocking, and upward ( $+v_{g_i}$ ), radial and rocking near-source differential ground motions corresponding to a given earthquake magnitude, ductility  $\mu$ , and for different time delays,  $\tau$ . Then we plot  $R_y$  versus  $T_n$  for the four corners of the system.

Figure 8 illustrates typical results for  $R_y$  versus the oscillator period for near-source, fault-parallel displacement  $d_N(t) = A_N(1 - e^{-t/\tau_N})/2$  [21], with downward vertical ground displacement, magnitude  $M = 8$ , for a ductility ratio of 8 and a time delay of  $\tau = 0.05$  s. It shows the results for the top-left, top-right, bottom-left, and bottom-right corners of the system, assuming wave propagation from left to right (see Fig. 7). For reference and easier comparison with the previously published results, we also plot one of the oldest estimates of  $R_y$  versus period, using piecewise straight lines [21]. The curve  $(R_y)_{\min}$  shows the minimum values of  $R_y$  for  $d_N(t)$  motion with  $-v_{g_i}$ , and for  $M = 8$ ,  $\mu = 8$ , and  $\tau = 0.05$  s.



Earthquake Engineering, Non-linear Problems in, Figure 8  
 Example of the effects of the differential ground motion on the strength-reduction factors  $R_y$  at the four corners of the structure in Fig. 7, subjected to horizontal, vertical, and rocking components of the fault-parallel displacement, for downward vertical motion ( $-v_{gj}$ ) for earthquake magnitude  $M = 8$ , ductility  $\mu = 8$ , and delay at the right support  $\tau = 0.05$  s. The amplitudes of the piecewise straight representation of the classical  $R_y$  are shown for comparison [21].  $(R_y)_{min}$  shows the smallest values of the  $R$ -factors, which for the set of conditions in this example are determined by the response at the top left corner (for periods shorter than 0.1 s), at the bottom right corner (for periods between 0.1 and 0.35 s), and at the top right corner (for periods longer than 0.35 s)

For periods longer than 5 to 10 s,  $R_y$  curves approach “collapse boundaries” [21]. This is implied in Fig. 8 by the rapid decrease of  $R_y$  versus period for periods longer than about 7 s. At or beyond these boundaries, the nonlinear system collapses due to the action of gravity loads and dynamic instability.

The complex results illustrated in Fig. 8 can be simplified by keeping only  $(R_y)_{min}$ , since it is only the minimum value of  $R_y$  that is needed for engineering design. By mapping  $(R_y)_{min}$  versus period of the oscillator for different earthquake magnitudes,  $M$ , different ductilities,  $\mu$ , and different delay times,  $\tau$ , design criteria can be formulated

for design of simple structures to withstand near-fault differential ground motions [21]. Nevertheless, the above shows how complicated the response becomes even for as simple a structure as the one shown by the model in Fig. 7, when differential ground motion with all of the components of motion is considered. In this example, this complexity results from simultaneous consideration of material and geometric nonlinearities, dynamic instability, and kinematic boundary conditions.

**Response in Terms of Wave Propagation – An Example**

The vibrational representation of the solution of response of a multi-degree-of-freedom system subjected to earthquake shaking is frequently simplified by considering only the fundamental and, occasionally, a few of the lowest frequencies of the system. Doing so is analogous to low-pass filtering of the complete solution [56,57], but it can work well when the excitation amplitudes are small and the motions are associated with long waves. However, during strong earthquakes, the ground motion contains large displacement pulses, the duration of which can be shorter than the fundamental period of the structure. For this type of excitation, the vibrational representation of response and the response spectrum superposition method cease to be suitable and should be replaced by a solution in terms of propagating waves. For short impulsive ground motions, the damage can occur before the wave entering the structure completes its travel up and down the structure, and well before the wave interference can occur—that is, well before the physical conditions can lead to the interference of waves and creation of the mode shapes.

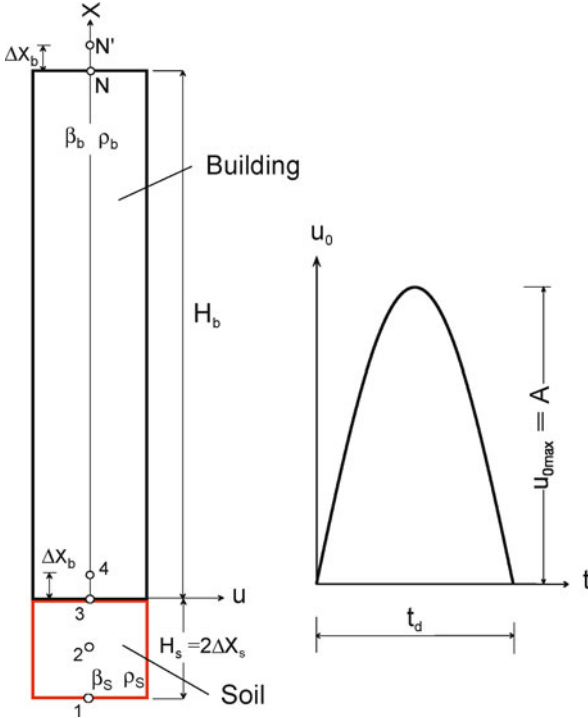
To illustrate the phenomena that can occur during nonlinear wave propagation in a building, we describe horizontal motions,  $u$ , in a one-dimensional shear beam, supported by one-dimensional half space and excited by a vertically propagating shear wave described by a half-sine-pulse (Fig. 9). A finite-difference scheme for solution of this problem with accuracy,  $O(\Delta t^2, \Delta x^2)$ , where  $\Delta x$  and  $\Delta t$  are the space and time increments, leads to the exact solution for  $\beta \Delta t / \Delta x = 1$ , where  $\beta$  is the velocity of shear waves. For simplicity, the incident displacement in the soil is chosen to be a sinusoidal pulse with the characteristics shown on Fig. 9.

A mesh with different spatial intervals in the soil and in the building will be used. The equation of motion is

$$v_t = (\sigma)_x / \rho, \tag{7a}$$

and the relation between the derivative of the strain and the velocity is





Earthquake Engineering, Non-linear Problems in, Figure 9  
Shear beam (building) (left) and incoming strong-motion displacement pulse (right) in the soil

$$\varepsilon_t = v_x, \quad (7b)$$

where  $v$ ,  $\rho$ ,  $\sigma$ , and  $\varepsilon$  are particle velocity, density, shear stress, and shear strain, respectively, and the subscripts  $t$  and  $x$  represent derivatives with regard to time and space.

The domain consists of two materials (Fig. 9): (1)  $-2\Delta x_s \leq x < 0$  with physical properties  $\rho_s$  and  $\mu_s$ , representing foundation soil, and (2)  $0 < x \leq H_b$  with physical properties  $\rho_b$  and  $\mu_b$  for linear response, where  $\rho_i$  is the density and  $\mu_i$  is the shear modulus in the soil ( $i = s$ ) or in the building ( $i = b$ ).  $v = \partial u / \partial t$  and  $\varepsilon = \partial u / \partial x$  are the velocity and the strain of a particle, and  $u$  is out-of-plane displacement of a particle perpendicular to the propagation ray.

It is assumed that the incoming wave is known and that its displacement as a function of time is prescribed at the point 1 in the soil ( $x = -2\Delta x_s$ ). Also, it is assumed that the soil is always in the linear elastic state. The finite difference method for a set of simultaneous equations is used to solve the problem, and spatial intervals are defined by  $\Delta x_i = \beta_i \cdot \Delta t$ , where  $\beta_i$  is the velocity of shear waves in the soil ( $i = s$ ) or in the building ( $i = b$ ) and  $\Delta t$  is the time step. The transparent boundary adopted for this study, which is described in Fujino and

Hakuno [8], is a perfect, transparent boundary for one-dimensional waves when  $\beta \Delta x / \Delta t = 1$ . Point 1 is where the prescribed displacement is applied, and we assume that this displacement travels upward in each time step. Point 2 is the boundary point of the model, where the quantities of motion are updated in each time step, and point 3 is the first spatial point, where the motion is computed using finite differences.

For the linear case at the contact (see point 3 in Fig. 9), one part of the incoming wave is transmitted into the other medium and one is reflected back into the same medium. The corresponding coefficients are obtained from the boundary conditions of continuity of the displacements and stresses at the contact. For a transmitted wave from medium B to medium A, the transmission coefficient is equal to  $k_{trB \rightarrow A} = 2 / [1 + \rho_a \beta_a / (\rho_b \beta_b)]$ . For a reflected wave from medium A back into medium B, this coefficient is  $k_{refB \rightarrow B} = [1 - \rho_a \beta_a / (\rho_b \beta_b)] / [1 + \rho_a \beta_a / (\rho_b \beta_b)]$ . For the opposite direction of propagation, the numerators and the denominators in these fractions exchange places.

### Numerical Examples

We consider a shear beam supported by elastic soil, as shown in Fig. 9. The densities of the soil and of the beam are assumed to be the same:  $\rho_b = \rho_s = \rho = 2000 \text{ kg/m}^3$ . The velocity of the shear waves in the soil is taken as  $\beta_s = 250 \text{ m/s}$ , and in the building as  $\beta_b = 100 \text{ m/s}$ .

To describe nonlinear response and the development of permanent deformations in the beam, we introduce two dimensionless parameters: (1) dimensionless amplitude  $\alpha = A / (H_b \varepsilon_{yb})$ , where  $A$  is the amplitude of the pulse (Fig. 9),  $H_b$  is the height of the building, and  $\varepsilon_{yb}$  is the yielding strain in the building, and (2) dimensionless frequency  $\eta = H_b / (\beta_b t_d)$ , where  $\beta_b t_d$  is one half of the wavelength of the wave in the building,  $\beta_b$  is the shear-wave velocity in the building, and  $t_d$  is the duration of the half-sine pulse.

To understand the development of the permanent strain in the nonlinear beam, we describe first the solution for the linear beam. The displacement and the strain for the linear beam are:

$$\begin{aligned} u(x, t) &= A \sum_{j=1}^{\infty} k_j \left\{ \sin \frac{\pi}{t_d} \left( t - t_{j-1} - \frac{x}{\beta_b} \right) \left[ H \left( t - t_{j-1} - \frac{x}{\beta_b} \right) \right. \right. \\ &\quad \left. \left. - H \left( t - t_{j-1} - \frac{x}{\beta_b} - t_d \right) \right] + \sin \frac{\pi}{t_d} \left( t - t_j + \frac{x}{\beta_b} \right) \right. \\ &\quad \left. \cdot \left[ H \left( t - t_j + \frac{x}{\beta_b} \right) - H \left( t - t_j + \frac{x}{\beta_b} - t_d \right) \right] \right\} \end{aligned} \quad (8)$$

and

$$\begin{aligned} \varepsilon(x, t) = & A \frac{\pi}{\beta_b t_d} \sum_{j=1}^{\infty} k_j \left\{ -\cos \frac{\pi}{t_d} \left( t - t_{j-1} - \frac{x}{\beta_b} \right) \right. \\ & \cdot \left[ H \left( t - t_{j-1} - \frac{x}{\beta_b} \right) - H \left( t - t_{j-1} - \frac{x}{\beta_b} - t_d \right) \right] \\ & + \cos \frac{\pi}{t_d} \left( t - t_j + \frac{x}{\beta_b} \right) \\ & \cdot \left. \left[ H \left( t - t_j + \frac{x}{\beta_b} \right) - H \left( t - t_j + \frac{x}{\beta_b} - t_d \right) \right] \right\} \quad (9) \end{aligned}$$

where  $j$  is the order number of the passage of the wave on the path bottom-top-bottom in the building,  $t_j = 2jH_b/\beta_b$  ( $j = 0, 1, 2, 3, \dots$ ), is the time required for the wave to pass  $j$  times over the path bottom-top-bottom (two heights),  $k_j = k_t k_r^{j-1}$  is the amplitude factor of the pulse in the soil in its  $j$ th passage along the path bottom-top-bottom through the building, and  $k_t$  and  $k_r$  are coefficients defined by  $k_{trB \rightarrow A}$  and  $k_{refB \rightarrow B}$  above.

The odd terms in Eq. (8) and Eq. (9) describe the response to the pulse coming from below, while the even terms describe the response to the pulse arriving from above. For the shear-wave velocities in our example,  $k_t = 10/7$  and  $k_r = -3/7$ . In Eq. (8) the displacement is positive for odd passages and negative for even passages. The displacement and velocity change sign after reflection from the soil-building interface and do not change sign after reflection from the top of the building. The strain changes sign after reflection from the top of the building and does not change sign after reflection from the building-soil interface. The constant that multiplies the series in Eq. (8) in terms of dimensionless amplitude and dimensionless frequency is  $A\pi/(\beta_b t_d) = A_\varepsilon = \pi\alpha\eta\varepsilon_{yb}$ .

To describe the occurrence of permanent strain, we consider two characteristic points in the building: (1) Point B ( $x = 0$ ) at the soil-building interface (point 3 in the grid, see Fig. 9), and (2) point T ( $x = H_b - \beta_b t_b/2$ ), where the amplitudes of the strain with the same sign meet after reflection from the top of the building. The location of this point is dependent upon the duration (wavelength) of the pulse. The first term in Eq. (8) is one if the argument of the cosine function is equal to  $t_d(t - t_0 - x/\beta_b = t_d)$ , and the second term is one if the argument of the second cosine function is equal to 0 ( $t - t_1 + x/\beta_b = 0$ ). The position of point T, where the strain amplitude is two times larger than the strain entering the beam, is at  $x = H_b - \beta_b t_d/2$ , and the time when this occurs is  $t = H_b/\beta_b + t_d/2$ . From Eq. (9) in the first passage of the pulse,  $t < 2H_b/\beta_b$ , and only the first term in the series exists. The strain at point B

reaches its absolute maximum at the very beginning, during the entrance of the pulse into the building, and its value is  $|\varepsilon_{B\max}^1| = \pi\alpha\eta\varepsilon_{yb}k_t$ . If this strain is greater than the yielding strain in the building,  $\varepsilon_{yb}$ , a permanent strain at the interface will develop, and the condition for occurrence of permanent strain at this point is  $|\varepsilon_{B\max}^1| > \varepsilon_{yb}$ , or, in terms of the dimensionless parameters,

$$\alpha\eta > (\pi k_t)^{-1} = (\beta_b + \beta_s)/(2\pi\beta_s) = C_B. \quad (10B)$$

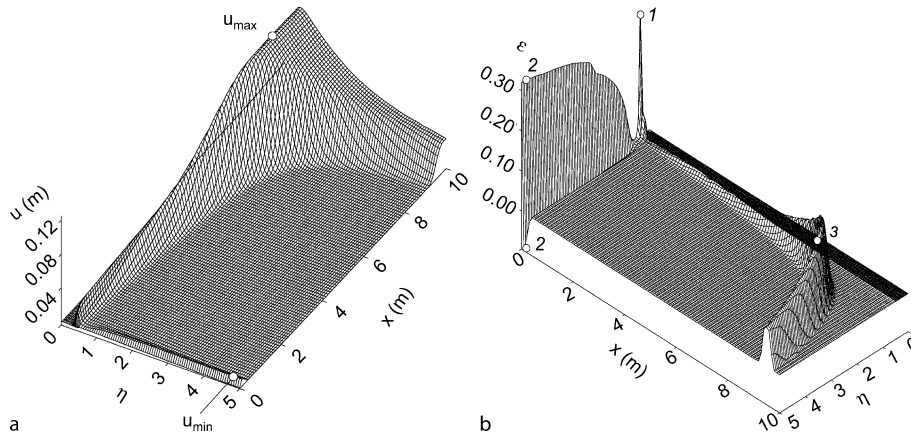
At point T (this point does not exist if  $t_d > 2H_b/\beta_b$ , and it coincides with point B if  $t_d = 2H_b/\beta_b$ ), from Eq. (9), the maximum strain during the first passage occurs at  $t = H_b/\beta_b + t_d/2$ , and its amplitude is  $2A_\varepsilon \cdot k_t$ . The condition for occurrence of the permanent strain is

$$\alpha\eta > (2\pi k_t)^{-1} = (\beta_b + \beta_s)/(4\pi\beta_s) = C_B/2 = C_T. \quad (10T)$$

For the shear-wave velocities in our example  $C_B = 0.2228$  and  $C_T = 0.1114$ .

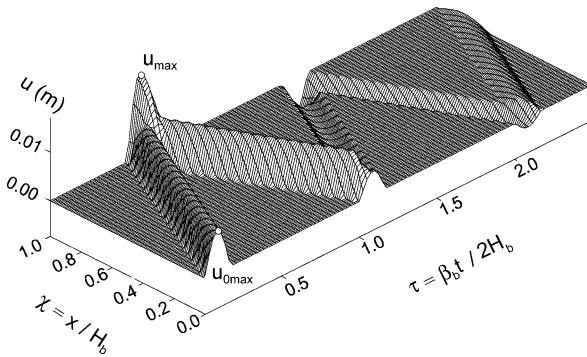
For the above simple model, the occurrence, development, and amplitudes of permanent strains and displacements have been studied by Gicev and Trifunac [10,11]. They found that for large ground-displacement pulses (large  $\alpha$ ) the maximum permanent strains occur mainly at the interface of the building with the soil, while for smaller amplitudes of pulses permanent strains occur closer to the top of the building. They distinguished three zones of the permanently deformed beam: (1) a permanently deformed zone at the bottom; (2) an intermediate zone, which is not deformed at its bottom part and is deformed in the top part; and (3) a non-deformed zone at the top of the beam. The occurrence and development of these zones depends upon the dimensionless excitation amplitudes and the dimensionless frequencies, and in particular on the conditions that lead to the occurrence of the first permanent strain (see Eqs. (10B) and (10T)). For large and long strong-motion pulses ( $\eta \leq 0.5$ ; first, the condition in Eq. (10B) is relevant), only zones 1 and 3 are present in the beam. For large amplitudes and short strong-motion pulses, all three zones develop and are present. For smaller excitation amplitudes (when the condition in Eq. (10B) cannot be satisfied for long pulses, and when the condition in Eq. (10T) is satisfied), only zones 2 and 3 exist in the beam. For larger values of  $\eta$  (when the condition in Eq. (10B) is satisfied) all three zones exist.

Gicev and Trifunac [10,11] found a similar situation for the occurrence of the maximum strains. For large and long pulses, maximum strain is located at the bottom of the building, and, as the pulses become shorter, peak strains occur at higher positions in the building. For some high frequencies of excitation, the maximum strain again appears at the bottom of the building because the loss of



Earthquake Engineering, Non-linear Problems in, Figure 10

Permanent displacements ( $u_{\max} = 0.126$  m) (left), and permanent strains ( $\epsilon_1 = 0.31$ ,  $\epsilon_2 = 0.32$ ,  $\epsilon_3 = 0.20$ ) (right), along the building versus dimensionless frequency  $\eta$  and for dimensionless amplitude  $\alpha = 0.3$



Earthquake Engineering, Non-linear Problems in, Figure 11

Linear displacements along the normalized length of the beam,  $\chi = x/H_b$ , versus normalized time  $\tau = \beta_b t / 2H_b$ , for dimensionless pulse amplitude  $\alpha = 0.03$  and dimensionless frequency  $\eta = 3$

energy due to the development of the permanent strain at the bottom overcomes the effects of the wave reflections from the top of the building (Fig. 10).

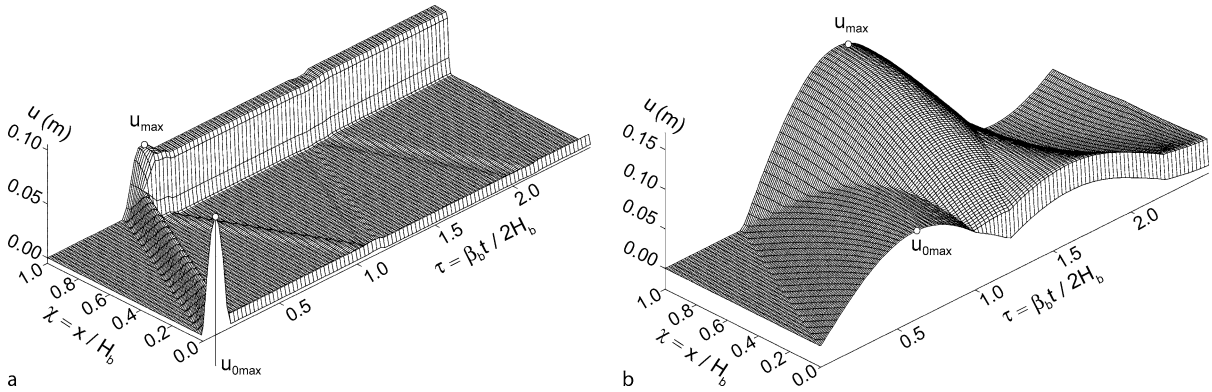
Creation of large permanent deformation zones in the building by the incident waves absorbs some or most of the incident wave energy and can reduce or eliminate further wave propagation and the associated energy transport (Figs. 11 and 12). To the extent that the locations of the plastic deformation zones can be controlled by the design process, absorption of the incident-wave energy by structural members may become a new and powerful tool for performance-based design. To take advantage of such possibilities, the governing differential equations must be solved by the wave-propagation method.

Examples illustrated here show that for excitation of structures by large, near-field displacement pulses failure

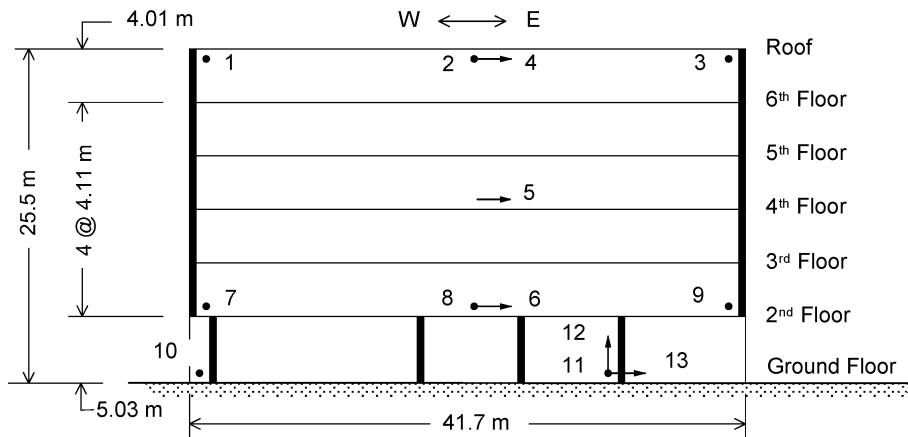
can occur anywhere in the building before the incident wave has completed its first travel from the foundation to the top of the building and back to the foundation ( $2H_b/\beta_b$ ). Because this travel time is shorter (by 1/2) than the natural period of the structure on the fixed base, it is seen that the common response spectrum method of analysis (based on the vibrational formulation of the solution) cannot provide the required details for the design of structures for such excitation. The complexity of the outcome increases with amplitudes of excitation and depends upon the pulse duration. Because actual strong ground motion in the near field has at least several strong pulses, it can be seen that the complexity in real structures responding to strong earthquake motions will be even greater. In engineering approximation based on the vibrational solution of the problem and on the SDOF models, where the location of ductile response is predetermined by the simple modeling assumptions, this complexity cannot be included because of the modeling constraints. The outcome is that it is virtually impossible for simplified models to identify or to predict the location of damage. In contrast, for properly chosen wave propagation models, prediction and identification of damage is a natural and logical outcome of interaction between excitation and model properties. A good example of this can be found in Gicev and Trifunac [12], who showed how a simple wave-propagation model can predict the actually observed location of damage.

### Observations of Nonlinear Response

Invaluable for understanding and proper treatment of the actual nonlinear response, and for validation of vibration



Earthquake Engineering, Non-linear Problems in, Figure 12  
 Nonlinear displacements along the normalized length of the beam,  $\chi = x/H_b$ , versus normalized time  $\tau = \beta_b t/2H_b$  for dimensionless pulse amplitude  $\alpha = 0.3$  and dimensionless frequencies  $\eta = 3$  (left) and  $\eta = 0.41$  (right)



Earthquake Engineering, Non-linear Problems in, Figure 13  
 Layout of the seismic monitoring array in the ICS building (dots, without arrows, show the NS recording channels)

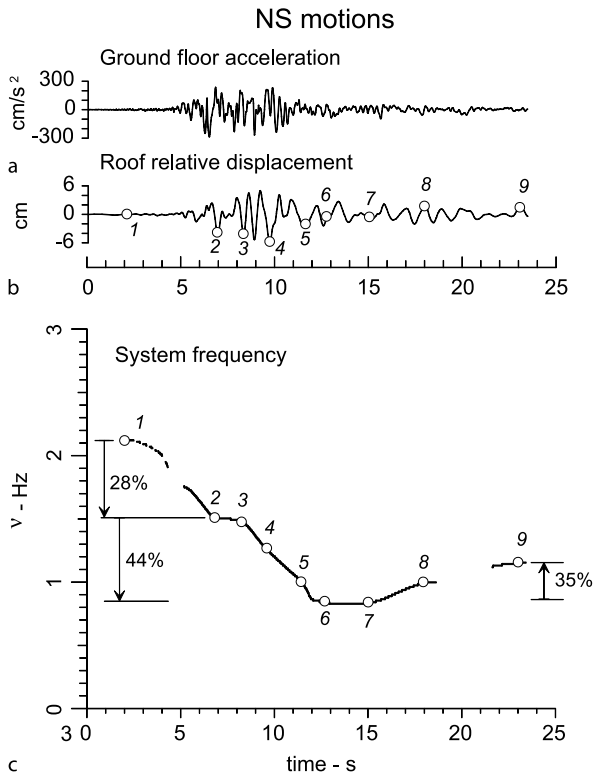
monitoring and analysis methods for real-life problems, are earthquake response data from well-instrumented, full-scale structures that have been damaged by an earthquake. Such data are rare and are not always freely available. An example of an instrumented building that has been damaged by an earthquake, and for which information about the damage and strong-motion data on the causative earthquake are available, is the former Imperial County Services Building in El Centro, California, which was severely damaged by the magnitude 6.6 Imperial Valley earthquake of October 15, 1979, and later demolished [23,51]. Its transverse (NS) response was recorded by three vertical arrays (recording channels 1, 3, 7, 9, 10, and 11; see Fig. 13), and its longitudinal (EW) response was recorded by one vertical array (recording channels 4, 5, 6, and 13, also shown in Fig. 13).

For a simplified soil-structure interaction model of a building supported by a rigid foundation, the difference

between the roof and base horizontal displacements during earthquake shaking is the sum of the horizontal displacements due to (1) horizontal deformation of the soil, (2) rigid-body rocking of the foundation, and (3) deformation of the structure. The estimated frequency from such data is referred to as system or “apparent” frequency, which differs from the fixed-base frequency of the building. While the fixed-base frequency depends only upon the properties of the structure, the apparent frequency depends also upon the stiffness of the foundation soil. The following relationship holds:

$$\frac{1}{\omega_{sys}^2} = \frac{1}{\omega_1^2} + \frac{1}{\omega_H^2} + \frac{1}{\omega_R^2}, \tag{13}$$

where  $\omega_{sys} = 2\pi\nu_{sys}$  is the soil-structure system frequency,  $\omega_1$  is the fundamental fixed-base frequency of the structure, and  $\omega_H$  and  $\omega_R$  are the horizontal and rocking



Earthquake Engineering, Non-linear Problems in, Figure 14  
**Time-frequency analysis for the NS response of the ICS building: a ground acceleration, b relative roof response, and c system frequency versus time**

frequencies, respectively, of a rigid structure on flexible soil [33].

Figure 14c shows that during earthquake shaking (Fig. 14a) the NS frequency of relative system response (Fig. 14b) dropped from  $\nu \approx 2.12$  Hz in the early stage of response (at  $t \approx 2$  s) to  $\nu \approx 1.52$  Hz at  $t \approx 6.8$  s ( $\Delta\nu \approx 0.6$  Hz,  $\Delta\nu/\nu \approx 28\%$ ), that it was constant during the interval  $t \approx 6.8 - 8.5$  s, and that it dropped further to  $\nu \approx 0.85$  Hz at  $t \approx 12$  s ( $\Delta\nu \approx 0.67$  Hz,  $\Delta\nu/\nu \approx 44\%$ ). Then, toward the end of the recorded shaking, the frequency increased to  $\nu \approx 1.15$  Hz ( $\Delta\nu \approx 0.3$  Hz;  $\Delta\nu/\nu \approx 35\%$ ). Early in the response ( $t < 7$  s), the amplitudes of the first story drifts in the building were relatively small ( $< 0.5\%$ ), and the observed decrease of system frequency is believed to be due to changes in the soil and bonding between the soil and foundation. This was followed by a further decrease in the system frequency of about 44% (between 8 and 12 s). The first-story drifts in the building were large when this occurred ( $> 0.5\%$  for NS), and the principal cause for this change is believed to be the damage, with the most severe damage occurring between 8 and 12 s after

trigger. Near the end of the shaking, a 35% increase in system frequency was observed, suggesting system hardening, which is believed to be due to changes in the soil [51].

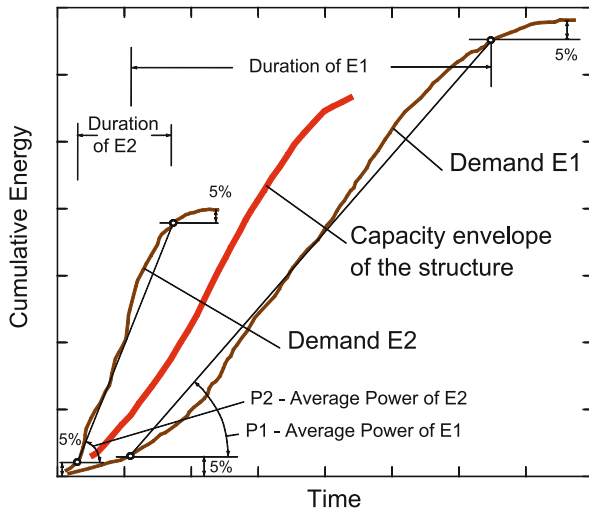
Changes similar to what is shown in Fig. 14c were first observed following the San Fernando earthquake in California in 1971 [66] and then during many subsequent earthquakes. It is known at present that many different factors can contribute to fluctuations of the system frequency, including rainfall, temperature fluctuations, changes in occupancy, remodeling and strengthening of buildings, wind, and earthquakes [49]. The simultaneous action of some of these factors and the associated time-dependent changes in the physical model contribute to complex and evolving system changes that make predictions of the dynamic response difficult.

### Future Directions

Well-designed structures are expected to have ductile behavior during the largest credible shaking, and a large energy reserve to at least delay failure if it cannot be avoided. As the structure finally enters large nonlinear levels of response, it absorbs the excess of the input energy through ductile deformation of its components. Thus, it is logical to formulate future earthquake-resistant design procedures in terms of the energy driving this process. From the mechanics point of view, this introduces nothing new, because the energy equations can be derived directly from the dynamic equilibrium equations. The advantage of using energy is that the duration of strong motion, the number of cycles to failure, and dynamic instability all can be addressed directly and explicitly. This, of course, requires scaling of the earthquake source and of the attenuation of strong motion described in terms of its wave energy. Trifunac et al. [65] reviewed the seismological aspects of empirical scaling of seismic wave energy,  $E_s$ , and showed how the radiated energy can be represented by the functionals of strong ground motion [53,54,55]. They described the energy propagation and attenuation with distance and illustrated it for the three-dimensional geological structure of the Los Angeles basin during the 1994 Northridge, CA earthquake, then they described the seismic energy flow through the response of soil-foundation-structure systems, analyzed the energy available to excite the structure, and finally examined the relative response of the structure.

### Power Design

Figure 15 illustrates the cumulative wave energies recorded at a building site during two hypothetical earthquakes, E1 and E2, and presents a conceptual framework



Earthquake Engineering, Non-linear Problems in, Figure 15  
**Schematic comparison of strong-motion power demands E1 and E2 with an envelope of structural power capacity**

that can be used for development of the power design method. E1 results in a larger total shaking energy at the site and has a long duration of shaking, leading to relatively small average power, P1. E2 leads to smaller total shaking energy at the site but has short duration and thus greater power, P2. The power capacity of a structure cannot be described by one unique cumulative curve, as this depends upon the time history of shaking. For the purposes of this illustration, the line labeled “capacity envelope of the structure” can be thought of as an envelope of all possible cumulative energy paths for the response of this structure. Figure 15 implies that E1 will not damage this structure, but E2 will. Hence, *for a given structure, it is not the total energy of an earthquake event (and the equivalent energy-compatible relative velocity spectrum) but the rate with which this energy arrives and shakes the structure that is essential for the design of the required power capacity of the structure to withstand this shaking and to control the level of damage.*

Trifunac [57] outlined the elementary aspects of such design based on the power of the incident wave pulses. He showed how this power can be compared with the capacity of the structure to absorb the incident wave energy and described the advantages of using the computed power of incident strong motion for design. Power (amplitude and duration) of the strong near-field pulses will determine whether the wave entering the structure will continue to propagate through the structure as a linear wave or will begin to create nonlinear zones (at first near the top and/or near the base of the structure; Gicev and Tri-

funac [10,11,12]). For high-frequency pulses, the nonlinear zone, with permanent strains, can be created before the wave motion reaches the top of the structure—that is, before the interference of waves has even started to occur and lead to formation of mode shapes. Overall duration of strong motion [60] will determine the number of times the structure may be able to complete full cycles of response and the associated number of “minor” excursions into the nonlinear response range when the response is weakly non-linear [13], while the presence of powerful pulses of strong motion will determine the extent to which the one-directional quarter period responses [57] may lead to excessive ductility demand, leading to dynamic instability and failure, precipitated by the gravity loads [20]. All of these possibilities can be examined and quantified deterministically by computation of the associated power capacities and power demands for different scenarios, for given recorded or synthesized strong-motion accelerograms, or probabilistically by using the methods developed for Uniform Hazard Analysis [52].

## Bibliography

1. Beltrami E (1987) Mathematics for Dynamic Modeling. Wiley, New York
2. Biot MA (1932) Vibrations of buildings during earthquakes. In: Transient Oscillations in Elastic System. Ph D Thesis No. 259, Chapter II. Aeronautics Department, California Institute of Technology, Pasadena
3. Biot MA (1933) Theory of elastic systems vibrating under transient impulse with an application to earthquake-proof buildings. Proc Natl Acad Sci 19(2):262–268
4. Biot MA (1934) Theory of vibration of buildings during earthquakes. Z Angew Math Mech 14(4):213–223
5. Biot MA (2006) Influence of foundation on motion of blocks. Soil Dyn Earthq Eng 26(6–7):486–490
6. Bycroft GN (1980) Soil-foundation interaction and differential ground motions. Earthq Eng Struct Dyn 8(5):397–404
7. Crutchfield JP (1992) Knowledge and meaning. In: Lam L, Naroditsky V (eds) Modeling Complex Phenomena. Springer, New York, pp 66–101
8. Fujino Y, Hakuno M (1978) Characteristics of elasto-plastic ground motion during an earthquake. Bull Earthq Res Inst Tokyo Univ 53:359–378
9. Gicev V (2005) Investigation of soil-flexible foundation-structure interaction for incident plane SH waves. Ph D Dissertation, Department of Civil Engineering, University Southern California, Los Angeles
10. Gicev V, Tifunac MD (2006) Rotations in the transient response of nonlinear shear beam. Department of Civil Engineering Report, CE 06–02. University Southern California, Los Angeles
11. Gicev V, Tifunac MD (2006) Non-linear earthquake waves in seven-story reinforced concrete hotel. Dept. of Civil Engineering, Report, CE 06–03. University Southern California, Los Angeles

12. Gicev V, Trifunac MD (2007) Permanent deformations and strains in a shear building excited by a strong motion pulse. *Soil Dyn Earthq Eng* 27(8):774–792
13. Gupta ID, Trifunac MD (1996) Investigation of nonstationarity in stochastic seismic response of structures. Dept. of Civil Eng. Report, CE 96–01. University of Southern California, Los Angeles
14. Gwinn EG, Westervelt RM (1985) Intermittent chaos and low-frequency noise in the driven damped pendulum. *Phys Rev Lett* 54(15):1613–1616
15. Hackett K, Holmes PJ (1985) Josephson Junction, annulus maps, Birkhoff Attractors, horsehoes and rotation sets. Center for Applied Math Report, Cornell University, Ithaca
16. Holmes PJ (1979) A nonlinear oscillator with a strange attractor. *Philos Trans R Soc London A* 292:419–448
17. Holmes PJ (1982) The dynamics of repeated impacts with a sinusoidally vibrating table. *J Sound Vib* 84:173–189
18. Holmes PJ (1985) Dynamics of a nonlinear oscillator with feedback control. *J Dyn Syst Meas Control* 107:159–165
19. Holmes PJ, Moon FC (1983) Strange Attractors and Chaos in Nonlinear Mechanics. *J Appl Mech* 50:1021–1032
20. Husid R (1967) Gravity effects on the earthquake response of yielding structures. Ph.D Thesis, California Institute of Technology, Pasadena
21. Jalali R, Trifunac MD (2007) Strength-reduction factors for structures subjected to differential near-source ground motion. *Indian Soc Earthq Technol J* 44(1):285–304
22. Kapitaniak T (1991) *Chaotic Oscillations in Mechanical Systems*. Manchester University Press, Manchester
23. Kojic S, Trifunac MD, Anderson JC (1984) A post-earthquake response analysis of the Imperial County Services building in El Centro. Report, CE 84–02. University of Southern California, Department of Civil Engineering, Los Angeles
24. Kuhn T (1962) *The Structure of Scientific Revolutions*. The University of Chicago Press, Chicago
25. Lee VW (1979) Investigation of three-dimensional soil-structure interaction. Department of Civil Engineering Report, CE 79–11. University of Southern California, Los Angeles
26. Lee VW, Trifunac MD (1982) Body wave excitation of embedded hemisphere. *ASCE, EMD* 108(3):546–563
27. Lee VW, Trifunac MD (1985) Torsional accelerograms. *Int J Soil Dyn Earthq Eng* 4(3):132–139
28. Lee VW, Trifunac MD (1987) Rocking strong earthquake accelerations. *Int J Soil Dyn Earthq Eng* 6(2):75–89
29. Levin PW, Koch BP (1981) Chaotic behavior of a parametrically excited damped pendulum. *Phys Lett A* 86(2):71–74
30. Lichtenberg AJ, Leiberman MA (1983) *Regular and Stochastic Motion*. Springer, New York
31. Lighthill J (1994) Chaos: A historical perspective. In: Newman WI, Gabrielov A, Turcotte D (eds) *Nonlinear Dynamics and Predictability of Geophysical Phenomena*. Geophysical Monograph 83, vol 18. IUGG, American Geophysical Union, Washington DC, pp 1–5
32. Lomnitz C, Castanos H (2006) Earthquake hazard in the valley of Mexico: entropy, structure, complexity, Chapter 27. In: Teisseyre R, Takeo M, Majewski E (eds) *Earthquake Source Asymmetry, Structural Media and Rotation Effects*. Springer, Heidelberg
33. Luco JE, Wong HL, Trifunac MD (1986) Soil-structure interaction effects on forced vibration tests. Department of Civil Engineering Report, No. 86–05. University of Southern California, Los Angeles
34. McLaughlin JB (1981) Period-doubling bifurcations and chaotic motion for a parametrically forced pendulum. *J Stat Phys* 24(2):375–388
35. Miles J (1984) Resonant motion of spherical pendulum. *Physica* 11D:309–323
36. Miles J (1984) Resonantly forced motion of two quadratically coupled oscillators. *Physica* 13D:247–260
37. Moon FC (1980) Experiments on chaotic motions of a forced nonlinear oscillator: Strange attractors. *ASME J Appl Mech* 47:638–644
38. Moon FC (1980) Experimental models for strange attractor vibration in elastic systems. In: Holmes PJ (ed) *New Approaches to Nonlinear Problems in Dynamics*. SIAM, Philadelphia, pp 487–495
39. Moon FC, Holmes PJ (1979) A magnetoelastic strange attractor. *J Sound Vib* 65(2):275–296; A magneto-elastic strange attractor. *J Sound Vib* 69(2):339
40. Moon FC, Holmes WT (1985) Double Poincare sections of a quasi-periodically forced, chaotic attractor. *Phys Lett A* 111(4):157–160
41. Moon FC, Shaw SW (1983) Chaotic vibration of beams with nonlinear boundary conditions. *J Nonlinear Mech* 18:465–477
42. Poddar B, Moon FC, Mukherjee S (1986) Chaotic motion of an elastic-plastic beam. *J Appl Mech ASME* 55(1):185–189
43. Rasband SN (1990) *Chaotic Dynamics of Nonlinear Systems*. Wiley, New York
44. Reitherman R (2006) The effects of the 1906 earthquake in California on research and education. *Earthq Spectr* S2(22):S207–S236
45. Richter PH, Scholtz HJ (1984) Chaos and classical mechanics: The double pendulum. In: Schuster P (ed) *Stochastic Phenomena and Chaotic Behavior in Complex Systems*. Springer, Berlin, pp 86–97
46. Shaw SW (1985) The dynamics of a harmonically excited system having rigid amplitude constraints, parts 1, 2. *J Appl Mech* 52(2):453–464
47. Shaw S, Holmes PJ (1983) A periodically forced piecewise linear oscillator. *J Sound Vib* 90(1):129–155
48. Sorrentino L (2007) The early entrance of dynamics in earthquake engineering: Arturo Danusso's contribution. *ISET J* 44(1):1–24
49. Todorovska MI, Al Rjoub Y (2006) Effects of rainfall on soil-structure system frequency: Examples based on poroelasticity and comparison with full-scale measurements. *Soil Dyn Earthq Eng* 26(6–7):708–717
50. Todorovska MI, Trifunac MD (1990) Note on excitation of long structures by ground waves. *ASCE, EMD* 116(4):952–964 (Errata in 116:1671)
51. Todorovska MI, Trifunac MD (2007) Earthquake Damage Detection in the Imperial County Services Building I: the Data and Time-Frequency Analysis. *Soil Dyn Earthq Eng* 27(6):564–576
52. Todorovska MI, Gupta ID, Gupta VK, Lee VW, Trifunac MD (1995) Selected topics in probabilistic seismic hazard analysis. Department of Civil Engineering Report, No. CE 95–08. University of Southern California, Los Angeles
53. Trifunac MD (1989) Dependence of Fourier spectrum amplitudes of recorded strong earthquake accelerations on magni-

- tude, local soil conditions and on depth of sediments. *Earthq Eng Struct Dyn* 18(7):999–1016
54. Trifunac MD (1993) Long-period Fourier amplitude spectra of strong motion acceleration. *Soil Dyn Earthq Eng* 12(6):363–382
  55. Trifunac MD (1994) Q and High-Frequency Strong-Motion Spectra. *Soil Dyn Earthq Eng* 13(3):149–161
  56. Trifunac MD (2003) 70th Anniversary of Biot Spectrum, 23rd Annual ISET Lecture. *Indian Soc Earthq Technol* 1(40):19–50
  57. Trifunac MD (2005) Power design method. *Proc. of Earthquake Engineering in the 21st Century to Mark 40th Anniversary of IZIS-Skopje*, 28 Aug–1 Sept. Skopje and Ohrid, Macedonia
  58. Trifunac MD (2006) Effects of torsional and rocking excitations on the response of structures, Ch 39. In: Teisseyre R, Takeo M, Majewski E (eds) *Earthquake Source Asymmetry, Structural Media, and Rotation Effects*. Springer, Heidelberg
  59. Trifunac MD, Gicev V (2006) Response Spectra for Differential Motion of Columns, Paper II: Out-of-Plane Response. *Soil Dyn Earthq Eng* 26(12):1149–1160
  60. Trifunac MD, Novikova EI (1994) State-of-the-art review on strong motion duration. 10th European Conf on Earthquake Eng, Vienna, 28 Aug – 2 Sep 1994 vol I. AA Balkema, Rotterdam, pp 131–140
  61. Trifunac MD, Todorovska MI (1997) Response spectra and differential motion of columns. *Earthq Eng Struct Dyn* 26(2): 251–268
  62. Trifunac MD, Ivanovic SS, Todorovska MI (2001) Apparent periods of a building I: Fourier analysis. *J Struct Eng ASCE* 127(5):517–526
  63. Trifunac MD, Ivanovic SS, Todorovska MI (2001) Apparent periods of a building II: Time-frequency analysis. *J Struct Eng ASCE* 127(5):527–537
  64. Trifunac MD, Hao TY, Todorovska MI (2001) Response of a 14-story reinforced concrete structure to nine earthquakes: 61 years of observation in the hollywood storage building. Department of Civil Engineering Report, CE 01–02. University of Southern California, Los Angeles
  65. Trifunac MD, Hao TY, Todorovska MI (2001) On energy flow in earthquake response, Department of Civil Engineering Report, No. CE 01–03. University of Southern California, Los Angeles
  66. Udwadia FE, MD Trifunac (1974) Time and amplitude dependent response of structures. *Earthq Eng Struct Dyn* 2:359–378
  67. Ueda Y (1980) Steady motions exhibited by Duffing's equation. In: Holmes PJ (ed) *A picture book of regular and chaotic motions. New Approaches to Nonlinear Problems in Dynamics*. SIAM, Philadelphia
  68. Veletsos AS, Newmark NM (1960) Effect of inelastic behavior on the response of simple systems to earthquake motions. *Proc 2nd World Conf on Earthquake Engineering*, Jul 1960, vol II. Science Council of Japan, Tokyo, pp 859–912
  69. Veletsos AS, Newmark NM (1964) Response spectra for single-degree-of-freedom elastic and inelastic systems. Report No. RTD-TDR-63-3096, vol III. Air Force Weapons Lab, Albuquerque
  70. Wong HL, Trifunac MD (1979) Generation of artificial strong motion accelerograms. *Int J Earthq Eng Struct Dyn* 7(6): 509–527



## Earthquake Forecasting and Verification

JAMES R. HOLLIDAY, JOHN B. RUNDLE,  
DONALD L. TURCOTTE  
University of California, Davis, USA

### Article Outline

Glossary  
Definition of the Subject  
Introduction  
Earthquake Forecasting  
Forecast Verification  
Future Directions  
Bibliography

### Glossary

**Binary forecast** A type of forecast where earthquakes are forecast to occur in certain regions and forecast not to occur in other regions.

**Continuum forecast** A type of forecast where the likelihood of an earthquake throughout an entire region is specified.

**Failure to predict** Earthquake event that occurs where no earthquake are forecasted to occur.

**False alarm** Earthquake event that is forecasted to occur at a specific location at a specific time but does not occur.

**PDF** Probability Density Function – A probability density function is any function  $f(x)$  that describes the probability density in terms of the input variable  $x$  such that  $f(x)$  is greater than or equal to zero for all values of  $x$  and the total area of the function is 1.

### Definition of the Subject

Forecasts of likely future events are used in almost every field: from forecasting tomorrow's weather to modeling the rise and fall of financial indices to predicting the growth of cancerous cells in human tissue. Generally, these forecasts are created under the belief that having a forecast – regardless of the level of complexity of the underlying models – is more desirable than not having a forecast. That is, human nature prefers the foreseeable over the unexpected. It is therefore important to verify the forecast and measure its skill, or “goodness”, and its value, or “usefulness”. The process of testing a given forecast with past trend data is the study of forecast verification. Forecast verification allows for a precise and repeatable (as op-

posed to relative or subjective) judgment of a forecasting model.

### Introduction

Earthquakes are the most feared of natural hazards because they generally occur without warning. Hurricanes can be tracked, floods develop gradually, tornados are caused by measurable atmospheric conditions, and volcanic eruptions are preceded by a variety of precursory phenomena. Earthquakes, however, occur suddenly and often without precursory indicators. There have been a wide variety of approaches applied to the forecasting of earthquakes [30,31,37,40,56,65]. These approaches can be divided into two general classes. The first approach is based on empirical observations of precursory changes. Examples include precursory seismic activity, precursory ground motions, and many others. The second approach is based on statistical patterns of seismicity. Neither approach has been able to provide reliable short-term forecasts (days to months) on a consistent basis.

Although short-term predictions are not available (see Table 1), long-term seismic-hazard assessments can be made. It is also possible to assess the long-term probability of having an earthquake of a given magnitude in a given region. These assessments are primarily based on the hypothesis that future earthquakes will occur in regions where past earthquakes have occurred [14,35]. Specifically, the rate of occurrence of small earthquakes in a region can be analyzed to assess the probability of occurrence of much larger earthquakes. While some earthquakes occur in plate interiors – a specific example is the three large (magnitude  $\sim 7.7$ ) earthquakes that occurred near New Madrid, Missouri in 1810 and 1811 – the large majority of all earthquakes occur in the vicinity of plate boundaries. A number of large cities are located very close to plate boundaries. Examples include Tokyo, Los Angeles, San Francisco, Seattle, Lima, Jakarta, and Santiago. Much of China is a diffuse plate boundary, and major earthquakes have caused devastating losses of life throughout this region. A recent example was in the 1976 Tangshan earthquake with some 500,000 deaths.

A major goal for earthquake forecasting is to quantify the risk of occurrence of an earthquake of a specified magnitude, in a specified area, and in a specified time window. This is routinely done and results in the creation of hazard maps. Another goal is to specifically forecast or predict earthquakes. The fundamental question is whether forecasts of the time and location of future earthquakes can be accurately made. It is accepted that long term hazard maps of the expected rate of occurrence of earthquakes are rea-

Earthquake Forecasting and Verification, Table 1

Warning times, scientific bases, and scientific feasibility for various types of earthquake predictions and estimates of long-term potential [62]

Term	Warning Time	Scientific Basis	Feasibility
Immediate alert	0 to 20 seconds	Speed of electro -magnetic waves $\gg$ speed of seismic waves	Good
Short-term prediction	Hours to weeks	Accelerating aseismic slip, foreshocks for some events	Unknown
Mid-term prediction	1 month to 10 years	Changes in seismicity, strain, chemistry, and fluid pressure	Fair
Long-term prediction	10 to 30 years	Time remaining in cycle of large shocks, increase in regional shocks	Good
Long-term potential	> 30 years	Long-term rate of activity, plate tectonic setting	Very good

sonably accurate. But is it possible to do better? Are there precursory phenomena that will allow earthquakes to be forecast?

### Earthquake Forecasting

#### Chaos and Forecasting

One of the reasons earthquakes are difficult to accurately forecast is the underlying complexity of the fault system. Earthquakes are caused by displacements on preexisting faults. Most earthquakes occur at or near the boundaries between the near-rigid plates of plate tectonics. Earthquakes in California are associated with the relative motion between the Pacific plate and the North American plate. Much of this motion is taken up by displacements on the San Andreas fault, but deformation and earthquakes extend from the Rocky Mountains on the east into the Pacific Ocean on the west. Clearly this deformation and the associated earthquakes are extremely complex.

It is now generally accepted that earthquakes are examples of deterministic chaos [66]. Some authors [16,17] have argued that this chaotic behavior precludes the prediction of earthquakes. Weather systems, however, are also chaotic, yet short-term forecasts are routinely made. Weather forecasts are probabilistic in the sense that weather cannot be predicted exactly. One such example is the track of a hurricane. Probabilistic forecasts of hurricane tracks are made every year; sometimes they are extremely accurate while at other times they are not. Another example of weather forecasting is the forecast of El Niño events. Forecasting techniques based on pattern recognition and principle components of the sea surface temperature fluctuation time series have been developed that are quite successful in forecasting future El Niños, but again they are probabilistic in nature [11]. It has also been argued [62] that chaotic behavior does not preclude the probabilistic forecasting of future earthquakes. The belief is that the chaos and nonlinearity in earthquakes arise mainly during unstable sliding in large events. Thus, predictions are possible before large earthquakes, but take a fi-

nite amount of time for the system to recover after large earthquakes.

#### Unobservable Dynamics

Another reason earthquakes are difficult to accurately forecast is that the true dynamics driving the system are simply unobservable and unmeasurable. As discussed above, earthquake faults occur in topologically complex, multi-scale networks that are driven to failure by external forces arising from plate tectonic motions [66]. The basic problem is that the details of the true space-time, force-displacement dynamics are in general unobservable, except in a few selected locations such as deep drill holes [52] or in a very crude, time-averaged sense such as the World Stress Map [81]. In order to completely describe the system, the true dynamics would have to be observable for all space and at all times. In fault systems these unobservable dynamics are usually encoded [59] in the time evolution of the Coulomb failure function,  $CFF(x, t)$ :

$$CFF(x, t) = \tau(x, t) - \mu_s \sigma_N(x, t), \tag{1}$$

where  $\tau(x, t)$  is the shear stress at point  $x$  and at time  $t$ ,  $\mu_s$  is the coefficient of static friction, and  $\sigma_N(x, t)$  is normal stress at point  $x$  and at time  $t$ . The space-time patterns associated with the time, location, and magnitude of the earthquakes, however, are observable. This leads to a focus on understanding the observable, multi-scale, apparent dynamics [52] of earthquakes in an attempt to infer the underlying dynamics.

#### Empirical Approaches

Empirical approaches to earthquake prediction rely on local observations of precursory phenomena in the vicinity of the earthquake to be predicted. It has been suggested that one or more of the following phenomena may indicate a future earthquake [30,31,37,40,56,65]:

1. precursory increase or decrease in seismicity in the vicinity of the origin of a future earthquake rupture,

2. precursory fault slip that leads to surface tilt and/or displacements,
3. electromagnetic signals,
4. chemical emissions, and
5. changes in animal behavior.

Examples of successful near-term predictions of future earthquakes based solely on empirical observations have been rare. A notable exception was the prediction of the  $M = 7.3$  Haicheng earthquake in northeast China that occurred on 4 February 1975. This prediction led to the evacuation of the city which undoubtedly saved many lives. The Chinese reported that the successful prediction was based on foreshocks, groundwater anomalies, and animal behavior. Unfortunately, a similar prediction was not made prior to the magnitude  $M = 7.8$  Tangshan earthquake that occurred on 28 July 1976 [68]. Official reports placed the death toll in this earthquake at 242,000, although unofficial reports placed it as high as 655,000.

In order to thoroughly test for the occurrence of direct precursors the United States Geological Survey (USGS) initiated the Parkfield (California) Earthquake Prediction Experiment in 1985 [1,30]. Earthquakes on this section of the San Andreas had occurred in 1857, 1881, 1901, 1922, 1934, and 1966. It was expected that the next earthquake in this sequence would occur by the early 1990s, and an extensive range of instrumentation was installed. The next earthquake in the sequence finally occurred on 28 September 2004. No precursory phenomena were observed that were significantly above the background noise level. Although the use of empirical precursors cannot be ruled out, the future of those approaches does not appear to be promising at this time.

### Statistical Approaches

A variety of studies have utilized variations in seismicity over relatively large distances to forecast future earthquakes. The distances are large relative to the rupture dimension of the subsequent earthquake. These approaches are based on the concept that the earth's crust is an activated, or driven, thermodynamic system [52]. Among the evidence for this behavior is the continuous level of background seismicity in all seismographic areas. About a million magnitude two earthquakes occur each year on our planet. In southern California about a thousand magnitude two earthquakes occur each year. Except for the aftershocks of large earthquakes, such as the 1992  $M = 7.3$  Landers earthquake, this seismic activity is essentially constant over time. If the level of background seismicity varied systematically with the occurrence of large earth-

quakes, earthquake forecasting would be relatively easy. This, however, is not the case.

While there is yet no indication of a universal earthquake indicator, there is increasing evidence that there are systematic precursory variations in some aspects of regional seismicity at least some of the time. For example, it has been observed that there is a systematic variation in the number of magnitude  $M = 3$  and larger earthquakes prior to at least some magnitude  $M = 5$  and larger earthquakes, and a systematic variation in the number of magnitude  $M = 5$  and larger earthquakes prior to some magnitude  $M = 7$  and larger earthquakes. The spatial regions associated with this phenomena tend to be relatively large, suggesting that an earthquake may resemble a phase change with an increase in the "correlation length" prior to an earthquake [5,26]. A specific example is the sequence of earthquakes that preceded the 1906 San Francisco earthquake [61]. This seismic activation has been quantified as a power law increase in seismicity prior to earthquakes [4,5,6,7,8,9,10,26,34,38,46,55,79]. There have also been reports of anomalous quiescence in the source region prior to a large earthquake, a pattern that is often called a "Mogi Donut" [30,40,76,77]. Unfortunately, these studies have all been performed retrospectively and their successes have depended on knowing the location of the subsequent earthquake.

There are two fundamentally different approaches to assessing the probabilistic risk of earthquake occurrence using statistical methods. The first of these is fault based, where the statistical occurrence of earthquakes is determined for mapped faults. The applicable models are known as renewal models and a tectonic loading of faults is included. The second approach is seismicity based, where the risk of future earthquakes is based on the past seismicity in the region. These are also known as cluster models and include the epidemic type aftershock sequence (ETAS) model and the branching aftershock sequence (BASS) model.

**Fault Based Models** Fault based models consider the earthquakes that occur on recognized (i. e., previously known) active faults. These models are also known as renewal models. Renewal models assume that the stress on an individual fault is "renewed" by the tectonic drive of plate tectonics. The simplest renewal model would be that of a single planar strike-slip fault subjected to a uniform rate of strain accumulation (plate motion). In this case, "characteristic" earthquakes would occur periodically. Clearly the earth's crust is much more complex with faults present at all scales and orientations. This complexity leads to chaotic behavior and statistical variability.

An important question is whether the concept of quasi-periodic “characteristic” earthquakes is applicable to tectonically active areas. There is extensive evidence that characteristic earthquakes do occur quasi-periodically on major faults. Many studies have been carried out to quantify the recurrence time statistics of these characteristic earthquakes [43,45,67]. Recurrence time statistics can be characterized by a mean value,  $\mu$ , and a coefficient of variation,  $C_v$ . The coefficient of variation is the ratio of the standard deviation to the mean. Mathematically,  $C_v = 0$  for periodic characteristic earthquakes and  $C_v = 1$  for a random distribution of recurrence times. Ellsworth et al. [13] reviewed many examples of recurrence time statistics and concluded that  $C_v \approx 0.5$  for characteristic earthquakes. Many probability distribution functions have been proposed for recurrence times, including the Weibull, lognormal, Brownian passage time, and gamma distributions.

Two major renewal simulation models have been developed. The first is “Virtual California” [49,50,53]. This is a geometrically realistic numerical simulation of earthquakes occurring on the San Andreas fault system and includes all major strike-slip faults in California. The second model is the “Standard Physical Earth Model” (SPEM) developed by Ward [69] and applied to characteristic earthquakes associated with subduction at the Middle American trench. This model was further developed and applied to the entire San Andreas fault system by Goes and Ward [18], to the San Andreas system in southern California by Ward [70], and to the San Andreas system in northern California by Ward [71].

Both simulation models utilize backslip, with the accumulation of a slip deficit on each fault segment prescribed using available data. The backslip represents the tectonic drive. Both models “tune” the prescribed static friction to give recurrence times that are consistent with available data. In both models fault segments are treated as dislocations when characteristic earthquakes occur, and all fault segments interact with each other elastically utilizing dislocation theory. These chaotic interactions result in statistical distributions of recurrence times on each fault. The resulting coefficients of variation are measures of this interaction.

Yakovlev et al. [78] utilized the Virtual California model to test alternative distributions of recurrence times. They concluded that the Weibull distribution is preferable and based its use on its scale invariance. The hazard rate is the probability that a characteristic earthquake will occur at a given time after the last characteristic earthquake. The Weibull distribution is the only distribution that has a power-law (scale-invariant) hazard function. In

the same study, Yakovlev et al. [78] found that the coefficient of variation of the recurrence times of 4606 simulated great earthquakes on the northern San Andreas fault is  $C_v = 0.528$ . Goes and Ward [18] using the SPEM simulator found that  $C_v = 0.50 - 0.55$  on this fault. The two simulations are quite different, so the statistical variability appears to be a robust feature of characteristic earthquakes. A similar simulation model for New Zealand has been given by Robinson and Benites [47,48].

Renewal models have also formed the basis for three formal assessments of future earthquake probabilities in California. These assessments were carried out by the United States Geological Survey [72,73,74,75]. A major problem with renewal models is that large earthquakes in nature often occur on faults that were not previously recognized. Recent examples in California include the 1952 Kern County earthquake, the 1971 San Fernando Valley earthquake, the 1992 Landers earthquake, the 1994 Northridge earthquake, and the 1999 Hector Mine earthquake. At the times when these earthquakes occurred, the associated faults were either not mapped or were considered too small to have such large earthquakes. To compensate for this problem, renewal models often include a random level of background seismicity unrelated to recognized faults.

**Seismicity Based Models** An alternative approach to probabilistic seismic hazard assessment and earthquake forecasting is to use observed seismicity. The universal applicability of Gutenberg–Richter frequency-magnitude scaling allows the rate of occurrence of small earthquakes to be extrapolated to estimate the rate of occurrence and location of large earthquakes. This type of extrapolation played an important role in creating the national seismic hazard map for the United States [15].

A more formalistic application of this extrapolation methodology is known as a relative intensity (RI) forecast. This type of forecast was made on a world wide basis by Kossobokov et al. [35] and to California by Holliday et al. [23]. A related forecasting methodology is the pattern informatics (PI) method [22,24,25,51,63,64]. This method was used by Rundle et al. [51] to forecast  $m = 5$  and larger earthquakes in California for the time period 2000–2010. This forecast successfully predicted the locations of 16 of the 18 large earthquakes that have subsequently occurred.

Keilis-Borok [31,33] and colleagues utilized patterns of seismicity to make formal intermediate term earthquake predictions. The most widely used algorithm, M8, has been moderately successful in predicting the times and locations of large earthquakes. More recently, this group has used chains of premonitory earthquakes to make interme-

mediate term predictions [32,58]. Again, moderate success was achieved.

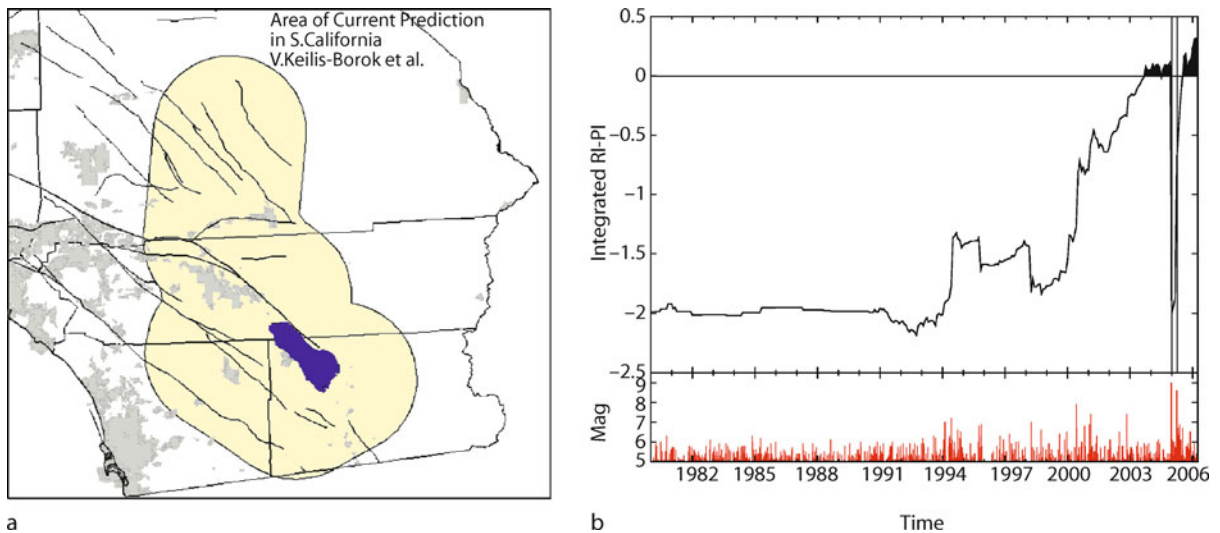
It has also been proposed that there is an increase in the number of intermediate sized earthquakes prior to a large earthquake [26]. This phenomenon is known as accelerating moment release (AMR) and is due primarily to an increase in the number of intermediate-size events that occur within a characteristic distance of the main shock and that scale with magnitude. AMR is characterized by a decrease in the rate of regional seismicity followed by a rapid rebound back to historic levels. Sammis and Bowman [54] have proposed a number of physical models to explain AMR. These include:

1. an analogy with critical phase transitions where the correlation length of the stress field rapidly increases as the system nears the critical point,
2. an erosion of a stress shadow from some previous, large event, and
3. a slow, silent earthquake propagating upward on a ductile extension loading the seismogenic crust above.

The existence of such a seismicity pattern does, however, appear to require a certain regional fault system structure

and density. Simulation models using a hierarchical distribution of fault sizes match this pattern well, but other types of fault distributions may also support AMR [26]. Conversely, some real-world fault distributions may not support AMR as a predictive tool. The AMR approach has shown considerable success retrospectively [5,10,55] but has not evolved into a successful prediction algorithm as of yet.

Seismicity based models are often referred to as clustering models. That is, clusters of small earthquakes indicate the future occurrence of larger earthquakes. The RI, PI, and AMR models clearly belong to this class. Other approaches in this class are the epidemic type aftershock sequence (ETAS) model [21,29,42,44] and the branching aftershock sequence (BASS) model. These are statistical models based on applicable scaling laws: Gutenberg–Richter scaling relation and the modified Båth’s law for the scaling relation of magnitudes, Omori’s law for the distribution of earthquake times, and a modified form of Omori’s law for the distribution of earthquake locations. Clustering by definition is not a random process. A rationale for the application of clustering models is that the clustering is related to families of foreshocks, main shocks, and aftershocks.



Earthquake Forecasting and Verification, Figure 1

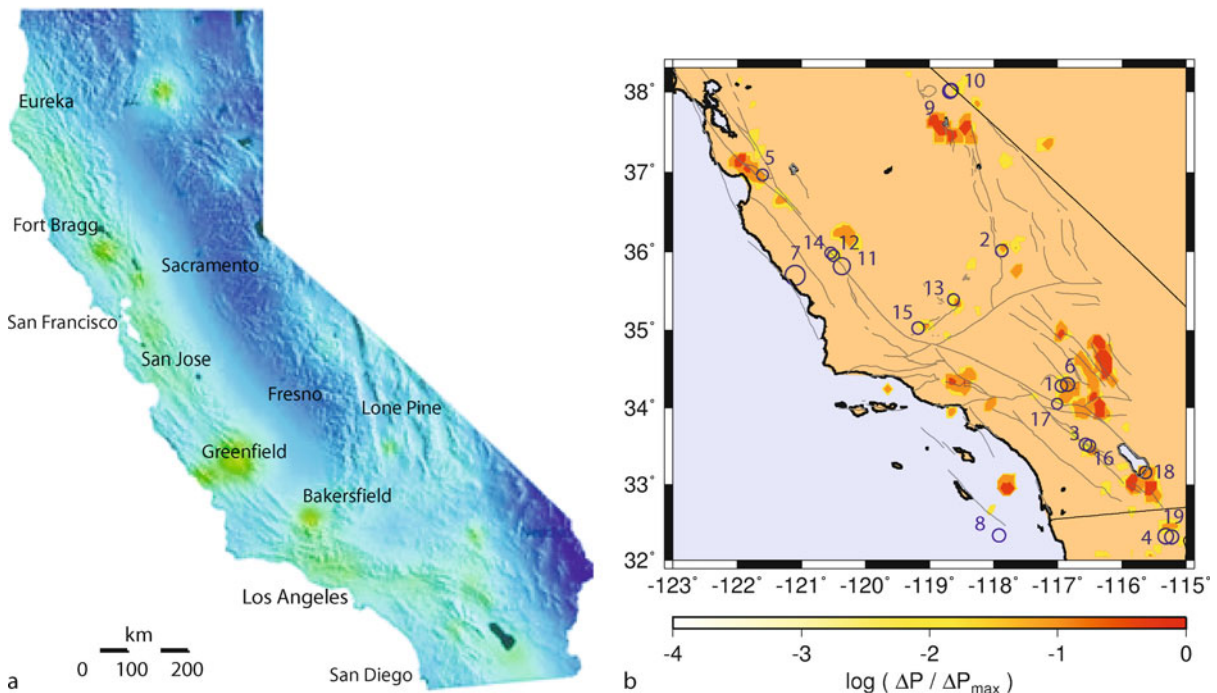
Plots of sample binary earthquake forecasts. **a** Prediction for a magnitude 6.4 or greater earthquake to occur between 5 January 2004 and 4 September 2004, within a 12,440 sq. miles area of southern California using the M8 algorithm (image courtesy Dr. Vladimir Keilis-Borok). This is a binary forecast since it forecasts an earthquake to occur within the *shaded* area during the time period and not to occur in the *non-shaded* region. Ultimately no magnitude 6.4 or greater earthquake occurred in the test region during the forecast interval. **b** Retrospective prediction for a magnitude 8.0 or greater earthquake to occur in the Sumatra region between 1 June 2003 and 1 June 2005 (image courtesy Dr. James Holliday). This is a time-dependent binary forecast since large earthquakes are forecasted to occur in the test region within a two year window once the time series becomes positive. Note that the magnitude 9.0 Sumatra-Andaman earthquake occurred 18 months after the time series became positive

## Types of Forecasts

**Binary Forecasts** The simplest type of earthquake forecast is a binary forecast. An earthquake is forecast to occur in a certain regions and forecast not to occur in other regions. This is analogous to the issuance of tornado warnings. Examples of two binary forecast maps are presented in Fig. 1. The plot on the left is for a prediction for a magnitude 6.4 or greater earthquake to occur between 5 January 2004 and 4 September 2004, within a 12,440 sq. miles area of southern California using the M8 algorithm. This map is a binary forecast since it forecasts an earthquake to occur within the shaded area during the time period and not to occur in the non-shaded region. The plot on the right is a retrospective prediction for a magnitude 8.0 or greater earthquake to occur in the Sumatra region between 1 June 2003 and 1 June 2005. This is a time-dependent binary forecast since large earthquakes are forecasted to occur in the test region within a two year window once the time series becomes positive and not to occur when the time series is negative.

**Continuum Forecasts** The alternative to binary forecasts is a continuum forecast. The likelihood of an earthquake throughout the entire region is specified. This would be analogous to temperature forecasts in the atmospheric sciences. Examples of two continuum forecast maps are presented in Fig. 2. The plot on the left is a time-dependent map produced in real time by the USGS Earthquake Hazards Program giving the probability of strong shaking at any location in California within a given 24-hour period. The plot on the right is a forecast map giving the probability for large (magnitude greater than 5.0) earthquakes in southern California using the Pattern Informatics method.

Any continuum forecast can be converted into a binary forecast through the use of a hard threshold. Spatial regions where the likelihood value is greater than the threshold are taken to be regions where earthquakes are forecasted to occur. Spatial regions where the likelihood value is less than the threshold value are taken to be regions where earthquakes are forecasted not to occur.



Earthquake Forecasting and Verification, Figure 2

Plots of sample continuum earthquake forecasts. **a** Time-dependent map giving the probability of strong shaking at any location in California within a given 24-hour period (image courtesy USGS Earthquake Hazards Program). **b** Forecast map giving the probability for large ( $m > 5$ ) earthquakes in southern California using the Pattern Informatics method (image courtesy Dr. Kristy Tiampo). Circles mark the locations of large earthquakes which occurred after the forecast creation. Both of these are continuum forecasts since they present a continuous likelihood for earthquakes to occur throughout the entire test region

## Forecast Verification

### Continuum Forecasts

**Likelihood Tests** The standard approach for testing the hypothesis that a probability measure can forecast future earthquakes is the maximum likelihood test [2,19,24,28,51,57,63]. The likelihood  $\mathbf{L}$  is a probability measure that can be used to assess the quality of one forecast measure over another. Typically, one computes the log-likelihood  $\mathcal{L} \equiv \log(\mathbf{L})$  for the proposed forecast measure  $\mathbf{L}$ . Models with higher (less negative) log-likelihood values are said to perform better than models with lower (more negative) log-likelihood values. In these types of likelihood tests, a probability density function (PDF) is required. Two different PDFs are commonly used: a global, Gaussian model and a local, Poissonian model.

Tiampo et al. [63] calculated likelihood values by defining  $P[\mathbf{x}]$  to be the union of a set of  $N$  Gaussian density functions  $p_G(|\mathbf{x} - \mathbf{x}_i|)$  [2] centered at each location  $\mathbf{x}_i$ . Each individual Gaussian density has a standard deviation  $\sigma$  equal to the width of their coarse-grained lattice cell and a peak value equal to the calculated probability divided by  $\sigma^2$ .  $P[\mathbf{x}(e_j)]$  was then interpreted as a probability measure that a future large event  $e_j$  would occur at location  $\mathbf{x}(e_j)$ :

$$P[\mathbf{x}(e_j)] = \sum_i \frac{P_i}{\sigma^2} e^{-\frac{|\mathbf{x}(e_j) - \mathbf{x}_i|^2}{\sigma^2}}. \quad (2)$$

If there are  $J$  future events, the normalized likelihood  $\mathbf{L}$  that all  $J$  events are forecast is

$$\mathbf{L} = \prod_j \frac{P[\mathbf{x}(e_j)]}{\sum_i P[\mathbf{x}_i]}. \quad (3)$$

Furthermore, the log-likelihood value  $\mathcal{L}$  for a given calculation can be calculated and used in ratio comparison tests:

$$\mathcal{L} = \sum_j \log \frac{P[\mathbf{x}(e_j)]}{\sum_i P[\mathbf{x}_i]}. \quad (4)$$

The second model commonly used is based on work performed by the Regional Earthquake Likelihood Models (RELM) group [57]. For each coarse-grained lattice cell  $i$  an expectation value  $\lambda_i$  is calculated by scaling the local probability value  $P_i$  by the number of earthquakes that occurred over all space during the forecasted time period:

$$\lambda_i = n \cdot P_i, \quad (5)$$

where  $n$  is the number of future events. Note that for any future time interval  $(t_2, t_3)$ ,  $n$  could in principle be estimated by using the Gutenberg–Richter relation. For each

bin an observation value  $\omega_i$  is also calculated such that  $\omega_i$  contains the number of future earthquakes that actually occurred in cell  $i$ . Note that  $\sum_i \omega_i = n$ . For the RELM model, it is assumed that earthquakes are independent of each other. Thus, the probability of observing  $\omega_i$  events in cell  $i$  with expectation  $\lambda_i$  is the Poissonian probability

$$p_i(\omega_i|\lambda_i) = \frac{\lambda_i^{\omega_i}}{\omega_i!} e^{-\lambda_i}. \quad (6)$$

The log-likelihood  $\mathcal{L}$  for observing  $\omega$  earthquakes at a given expectation  $\lambda$  is defined as the logarithm of the probability  $p(\omega|\lambda)$ , thus

$$\mathcal{L}(\omega|\lambda) = \log p(\omega|\lambda) = -\lambda + \omega \log \lambda - \log(\omega!). \quad (7)$$

Since the joint probability is the product of the individual cell probabilities, the log-likelihood value for a given calculation is the sum of  $\mathcal{L}(\omega_i|\lambda_i)$  over all cells  $i$ :

$$\mathcal{L} = \sum_i \mathcal{L}(\omega_i|\lambda_i) = \sum_i (-\lambda_i + \omega_i \log \lambda_i - \log(\omega_i!)). \quad (8)$$

Most tests of earthquake forecasts have emphasized the likelihood test [24,28,51,64]. These tests have the significant disadvantage that they are overly sensitive to the least probable events. For example, consider two forecasts. The first perfectly forecasts 99 out of 100 events but assigns zero probability to the last event. The second assigns zero probability to all 100 events. Under a log-likelihood test, both forecasts will have the same skill score of  $-\infty$ . Furthermore, a naive forecast that assigns uniform probability to all possible sites will always score higher than a forecast that misses only a single event but is otherwise superior. For this reason, likelihood tests are more subject to unconscious bias. Other methods of evaluating earthquake forecasts are suggested by Vere-Jones [20] and Holliday et al. [25].

**Information Metrics** One such alternative is the use of information metrics. Using methods from information theory [12], it is possible to calculate the entropy,  $H$ , of a forecast map. Entropy can be considered a measure of disorder (e.g. randomness) or “surprise”, hence maps with lower entropy contain more useful information than maps with higher entropy. We define entropy as

$$H(z) = - \sum_{i=1}^N p(\mathbf{x}_i; z) \log p(\mathbf{x}_i; z), \quad (9)$$

where

$$p(\mathbf{x}_i; z) = \begin{cases} P(\mathbf{x}_i) & P(\mathbf{x}_i) \geq z \\ 0 & P(\mathbf{x}_i) < z \end{cases}, \quad (10)$$

and the probabilities are scaled such that  $\sum_{i=1}^N p(x_i) = 1$ . This definition allows a measurement of entropy as a function of some lower threshold  $z$ . A small, non-zero value for  $z$  allows for the measurement of the entropy above and relative to background noise.

## Binary Forecasts

**ROC Analysis** The standard approach to the evaluation of a binary forecast is the use of a relative operating characteristic (ROC) diagram [39,60]. This method evaluates the performance of the forecast method relative to random guessing by constructing a plot of the fraction of failures to predict (events that occur where no event is forecast) against the fraction of false alarms (events that are forecast to occur at a location but do not occur) for an ensemble of forecasts. Molchan [41] has used a modification of this method to evaluate the success of intermediate term earthquake forecasts.

The binary approach has a long history, over 100 years, in the verification of tornado forecasts [39]. These forecasts take the form of a tornado forecast for a specific location and time interval, each forecast having a binary set of possible outcomes. For example, during a given time window of several hours duration, a forecast is issued in which a list of counties is given with a statement that one or more tornadoes will or will not occur. A  $2 \times 2$  contingency table is then constructed, the top row contains the counties in which tornadoes are forecast to occur and the bottom row contains counties in which tornadoes are forecast to not occur. Similarly, the left column represents counties in which tornadoes were actually observed, and the right column represents counties in which no tornadoes were observed.

With respect to earthquakes, binary forecasts take exactly this form. A time window is proposed during which the forecast of large earthquakes having a magnitude above some minimum threshold is considered valid. An example might be a forecast of earthquakes larger than  $M = 5$  during a period of five or ten years duration. A map of the seismically active region is then completely covered ("tiled") with boxes of two types: boxes in which the epicenters of at least one large earthquake are forecast to occur and boxes in which large earthquakes are forecast to not occur. In other types of forecasts, large earthquakes are given some continuous probability of occurrence from 0% to 100% in each box [28]. These forecasts can be converted to the binary type by the application of a threshold value. Boxes having a probability below the threshold are assigned a forecast rating of non-occurrence during the time window, while boxes having a probability above

Earthquake Forecasting and Verification, Table 2  
Schematic contingency table for categorical forecasts of a binary event

Forecast	Observed		
	Yes	No	Total
Yes	$a$	$b$	$a + b$
No	$c$	$d$	$c + d$
Total	$a + c$	$b + d$	$a + b + c + d = N$

the threshold are assigned a forecast rating of occurrence. A high threshold value may lead to many failures to predict, but few false alarms. The level at which the threshold is set is then a matter of public policy specified by emergency planners, representing a balance between the prevalence of failures to predict and false alarms.

**Contingency Tables** An extensive review on forecast verification in the atmospheric sciences has been given by Jolliffe and Stephenson [27]. The wide variety of approaches that they consider are directly applicable to earthquake forecasts as well. Verification of earthquake forecasts proceeds in exactly the same way as for, say, tornado forecasts when using these approaches. For a given forecast, the contingency table (see Table 2) is constructed. Values for the table elements  $a$  (Forecast=yes, Observed=yes),  $b$  (Forecast=yes, Observed=no),  $c$  (Forecast=no, Observed=yes), and  $d$  (Forecast=no, Observed=no) are obtained from the forecast map. The fraction of alarm space, also called the probability of forecast of occurrence, is  $r = (a + b)/N$ , where the total number of boxes is  $N = a + b + c + d$ . The hit rate is  $H = a/(a + c)$  and the false alarm rate is  $F = b/(b + d)$ . From these quantities a number of descriptive, performance, and skill measures can be constructed [27]. Table 3 lists a few possible measures.

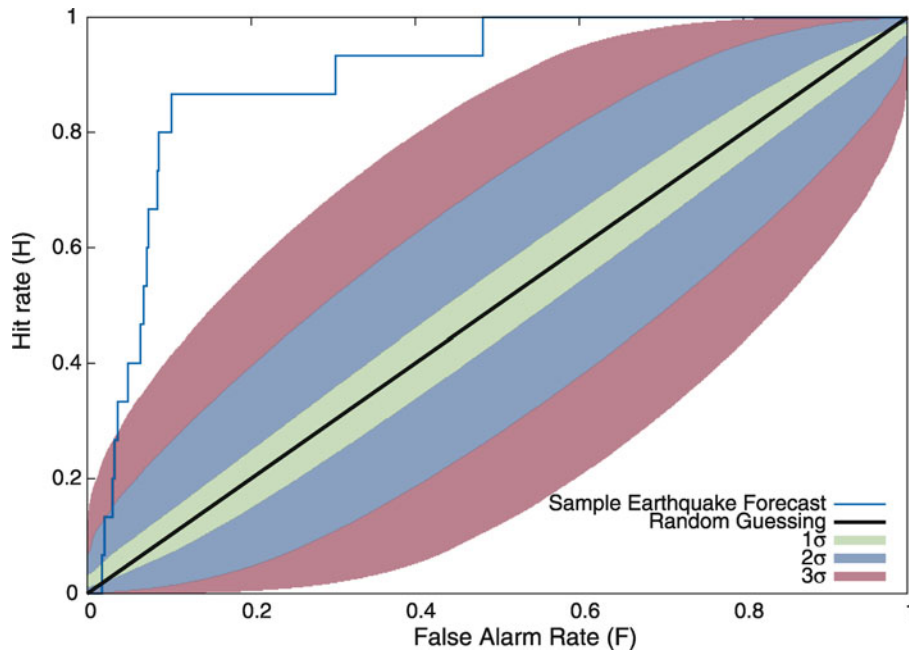
**ROC Curves** The standard ROC diagram [23,27] is a plot of the points  $\{H, F\}$  calculated for a binary forecast (see Fig. 3). If the forecast was converted from continuum map,  $H$  and  $F$  are plotted as the lower (conversion) threshold is varied. A perfect forecast of occurrence (perfect order, no fluctuations) would consist of two line segments, the first connecting the points  $(H, F) = (0, 0)$  to  $(H, F) = (1, 0)$ , and the second connecting  $(H, F) = (1, 0)$  to  $(H, F) = (1, 1)$ . A curve of this type can be described as maximum possible hits ( $H = 1$ ) with minimum possible false alarms ( $F = 0$ ). Another type of perfect forecast consists of two lines connecting the points  $(0, 0)$  to  $(0, 1)$  and  $(0, 1)$  to  $(1, 1)$ , a perfect forecast of non-occurrence.



Earthquake Forecasting and Verification, Table 3

Table of various descriptive and performance measures that can be calculated from the binary contingency table for an earthquake forecast

Name	Definition	Definition ( $H, F,$ and $\tau$ )	Range
Fraction of alarm space $\tau$	$\tau = (a + c)/N$	$\tau$	[0, 1]
Hit rate $H$	$H = a/(a + c)$	$H$	[0, 1]
False alarm rate $F$	$F = b/(b + d)$	$F$	[0, 1]
False alarm ratio $FAR$	$FAR = b/(a + b)$	$FAR = (1 + \frac{\tau}{1-\tau} \frac{H}{F})^{-1}$	[0, 1]
Miss rate $\nu$	$\nu = c/(a + c)$	$\nu = 1 - H$	[0, 1]
Peirce's skill score $PSS$	$PSS = \frac{ad-bc}{(b+d)(a+c)}$	$PSS = H - F$	[-1, 1]
Yule's $Q$	$Q = \frac{ad-bc}{ad+bc}$	$Q = \frac{H-F}{H(1-F)+F(1-H)}$	[-1, 1]
Peirce Area $A$	$A = \int PSS$	$A = \int HdF - \frac{1}{2}$	$[-\frac{1}{2}, \frac{1}{2}]$



Earthquake Forecasting and Verification, Figure 3

Sample relative operating characteristic (ROC) diagram. Shown is a plot of hit rates,  $H$ , as a function of false alarm rates,  $F$ , for a sample earthquake forecast (blue) and random guessing (black). Confidence intervals for the one-, two- and three- $\sigma$  levels are shown as well [23,80]

The line  $H = F$  occupies a special status, and corresponds to a completely random forecast [23,27] (maximum disorder, maximum fluctuations) where the false alarm rate is the same as the hit rate and no information is produced by the forecast. Points above this line are said to have performed better than simple random guessing. If competing forecasts are plotted on the same graph, the forecast whose  $H-F$  curves lies the highest is said to outperform the others. Often, however, competing forecasts will have intersecting curves. In this case, forecasts

are said outperform each other only in specific ranges and only for specific choices of the lower (conversion) threshold value.

**$\nu$ - $\tau$  Curves** An alternative diagram [41] is a plot of the points  $\{\nu, \tau\}$  for a binary forecast map. In this case a perfect forecast of occurrence would consist of two line segments, the first connecting the points  $(\nu, \tau) = (0, 0)$  to  $(\nu, \tau) = (0, 1)$ , and the second connecting  $(\nu, \tau) = (0, 1)$  to  $(\nu, \tau) = (1, 1)$ . A curve of this type can be described as

minimum possible missed events ( $\nu = 0$ ) with minimum possible alarm space ( $\tau = 0$ ).

As with the  $H$ - $F$  curve, the line  $\nu = \tau$  corresponds to a completely random forecast. Points below this line are said to have performed better than simple random guessing. If competing forecasts are plotted on the same graph, the forecast whose  $\nu$ - $\tau$  curves lies the lowest is said to outperform the others. As can be verified from Table 3,  $\nu$ - $\tau$  curves offer the same information as  $H$ - $F$  curves and are identical in the range  $a \ll d$ .

### Future Directions

It is actually quite surprising that immediate local precursory phenomena are not seen. Prior to a volcanic eruption, increases in regional seismicity and surface movements are generally observed. For a fault system, the stress gradually increases until it reaches the frictional strength of the fault and a rupture is initiated. It is certainly reasonable to hypothesize that the stress increase would cause increases in background seismicity and aseismic slip. In order to test this hypothesis the Parkfield Earthquake Prediction Experiment was initiated in 1985. The expected Parkfield earthquake occurred beneath the heavily instrumented region on 28 September 2004. No local precursory changes were observed [3,36]. In the absence of local precursory signals, the next question is whether broader anomalies develop, and in particular whether there is anomalous seismic activity.

Assuming precursors do exist and can be exploited to create earthquake forecasts, testing and verification of the usefulness of the forecasts is the necessary next step. A forecast method that predicts all earthquakes but carries a high false alarm rate is likely to be useless as a public warning tool. Similarly, a forecast method that issues warnings for only a small fraction of actual earthquakes but never issues false alarms is likely to be a poor tool for catastrophe preparation. Forecast verification techniques can be used to find the middle ground that is most useful.

As a final warning, researchers must be careful not to create artificial skill in their forecasts. Since nature provides us only with one earthquake record (the actual history of a given test region), the quality of a new forecast system is often assessed on the same data set used to create it. This can potentially lead to an optimistic bias in any skill scores. This is a particular problem if the score itself is used to calibrate the method, either directly or indirectly. Development of realistic earthquake simulators and cross-testing against test regions with similar fault structures can help protect against this.

## Bibliography

### Primary Literature

1. Bakun WH, Lindh AG (1985) The Parkfield, California, earthquake prediction experiment. *Science* 229:619–624
2. Bevington PR, Robinson DK (1992) *Data Reduction and Error Analysis for the Physical Sciences*. McGraw-Hill, New York
3. Borchardt RD, Johnston MJS, Glassmoyer G, Dietel C (2006) Recordings of the 2004 parkfield earthquake on the general earthquake observation system array: Implications for earthquake precursors, fault rupture, and coseismic strain changes. *Bull Seismol Soc Am* 96(4b):73–89
4. Bowman DD, King GCP (2001) Accelerating seismicity and stress accumulation before large earthquakes. *Geophys Res Lett* 28:4039–4042
5. Bowman DD, Ouillon G, Sammis CG, Sornette A, Sornette D (1998) An observational test of the critical earthquake concept. *J Geophys Res* 103:24359–24372
6. Bowman DD, Sammis CG (2004) Intermittent criticality and the Gutenberg–Richter distribution. *Pure Appl Geophys* 161:1945–1956
7. Brehm DJ, Braile LW (1998) Intermediate-term earthquake prediction using precursory events in the New Madrid seismic zone. *Bull Seismol Soc Am* 88:564–580
8. Brehm DJ, Braile LW (1999) Intermediate-term earthquake prediction using the modified time-to-failure method in southern California. *Bull Seismol Soc Am* 89:275–293
9. Buffe CG, Nishenko SP, Varnes DJ (1994) Seismicity trends and potential for large earthquakes in the Alaska–Aleutian region. *Pure Appl Geophys* 142:83–99
10. Buffe CG, Varnes DJ (1993) Predictive modeling of the seismic cycle of the greater San Francisco Bay region. *J Geophys Res* 98:9871–9883
11. Chen D, Cane MA, Kaplan A, Zebian SE, Huang D (2004) Predictability of El Niño in the past 148 years. *Nature* 428:733–736
12. Cover TM, Thomas JA (1991) *Elements of Information Theory*. Wiley-Interscience, New York
13. Ellsworth WL, Mathews MV, Nadeau RM, Nishenko SP, Reasenberg PA, Simpson RW (1999) A physically-based earthquake recurrence model for estimation of long-term earthquake probabilities. Open-File Report 99-522, US Geological Survey
14. Frankel AF (1995) Mapping seismic hazard in the central and eastern United States. *Seismol Res Lett* 60:8–21
15. Frankel AF, Mueller C, Barnhard T, Perkins D, Leyendecker EV, Dickman N, Hanson S, Hopper M (1996) National seismic hazard maps. Open-File Report 96-532, US Geological Survey
16. Geller RJ (1997) Earthquake prediction: A critical review. *Geophys J Int* 131:425–450
17. Geller RJ, Jackson DD, Kagen YY, Mulargia F (1997) Earthquakes cannot be predicted. *Science* 275:1616–1617
18. Goes SDB, Ward SN (1994) Synthetic seismicity for the San Andreas fault. *Annali Geofis* 37:1495–1513
19. Gross S, Rundle JB (1998) A systematic test of time-to-failure analysis. *Geophys J Int* 133:57–64
20. Harte D, Vere-Jones D (2005) The entropy score and its uses in earthquake forecasting. *Pure Appl Geophys* 162:1229–1253. doi:10.1007/s00024-004-2667-2
21. Helmstetter A Is earthquake triggering driven by small earthquakes? *Phys Rev Lett* 91:0585014

22. Holliday JR, Chen CC, Tiampo KF, Rundle JB, Turcotte DL, Donnellan A (2007) A RELM earthquake forecast based on pattern informatics. *Seis Res Lett* 78(1):87–93
23. Holliday JR, Nanjo KZ, Tiampo KF, Rundle JB, Turcotte DL (2005) Earthquake forecasting and its verification. *Nonlinear Process Geophys* 12:965–977
24. Holliday JR, Rundle JB, Tiampo KF, Klein W, Donnellan A (2006) Modification of the pattern informatics method for forecasting large earthquake events using complex eigenvectors. *Tectonophysics* 413:87–91. doi:10.1016/j.tecto.2005.10.008
25. Holliday JR, Rundle JB, Tiampo KF, Klein W, Donnellan A (2006) Systematic procedural and sensitivity analysis of the pattern informatics method for forecasting large ( $M \geq 5$ ) earthquake events in southern California. *Pure Appl Geophys* 163:2433–2454. doi:10.1007/s00024-006-0131-1
26. Jaumé SC, Sykes LR (1999) Evolving towards a critical point: A review of accelerating seismic moment/energy release prior to large and great earthquakes. *Pure Appl Geophys* 155:279–306
27. Jolliffe IT, Stephenson DB (2003) *Forecast Verification*. Wiley, Chichester
28. Kagan YY, Jackson DD (2000) Probabilistic forecasting of earthquakes. *Geophys J Int* 143:438–453
29. Kagan YY, Knopoff L (1981) Stochastic synthesis of earthquake catalogs. *J Geophys Res* 86(4):2853–2862
30. Kanamori H (2003) Earthquake prediction: An overview. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake & Engineering Seismology*. Academic Press, Amsterdam, pp 1205–1216
31. Keilis-Borok V (2002) Earthquake predictions: State-of-the-art and emerging possibilities. *Ann Rev Earth Planet Sci* 30:1–33
32. Keilis-Borok V, Shebalin P, Gabrielov A, Turcotte D (2004) Reverse tracing of short-term earthquake precursors. *Phys Earth Planet Int* 145:75–85
33. Keilis-Borok VI (1990) The lithosphere of the earth as a nonlinear system with implications for earthquake prediction. *Rev Geophys* 28:19–34
34. King GCP, Bowman DD (2003) The evolution of regional seismicity between large earthquakes. *J Geophys Res* 108:2096
35. Kossobokov VG, Keilis-Borok VI, Turcotte DL, Malamud BD (2000) Implications of a statistical physics approach for earthquake hazard assessment and forecasting. *Pure Appl Geophys* 157:2323–2349
36. Lindh AG (2005) Success and failure at Parkfield. *Seis Res Lett* 76:3–6
37. Lomnitz C (1994) *Fundamentals of Earthquake Prediction*. Wiley, New York
38. Main IG (1999) Applicability of time-to-failure analysis to accelerated strain before earthquakes and volcanic eruptions. *Geophys J Int* 139:F1–F6
39. Mason IB (2003) Binary events. In: Jolliffe IT, Stephenson DB (eds) *Forecast Verification*. Wiley, Chichester, pp 37–76
40. Mogi K (1985) *Earthquake Prediction*. Academic Press, Tokyo
41. Molchan GM (1997) Earthquake predictions as a decision-making problem. *Pure Appl Geophys* 149:233–247
42. Ogata Y (1988) Statistical models for earthquake occurrences and residual analysis for point processes. *J Am Stat Assoc* 83:9–27
43. Ogata Y (1999) Seismicity analysis through point-process modeling: a review. *Pure Appl Geophys* 155:471–507
44. Ogata Y, Zhuang J (2006) Space-time ETAS models and an improved extension. *Tectonophysics* 413:13–23. doi:10.1016/j.tecto.2005.10.016
45. Rikitake T (1982) *Earthquake forecasting and warning*. D. Reidel Publishing Co, Dordrecht
46. Robinson R (2000) A test of the precursory accelerating moment release model on some recent New Zealand earthquakes. *Geophys J Int* 140:568–576
47. Robinson R, Benites R (1995) Synthetic seismicity models of multiple interacting faults. *J Geophys Res* 100:18229–18238
48. Robinson R, Benites R (1996) Synthetic seismicity models for the Wellington Region, New Zealand: implications for the temporal distribution of large events. *J Geophys Res* 101:27833–27844
49. Rundle JB, Rundle PB, Donnellan A (2005) A simulation-based approach to forecasting the next great San Francisco earthquake. *Proc Natl Acad Sci* 102(43):15363–15367
50. Rundle JB, Rundle PB, Donnellan A, Fox G (2004) Gutenberg–Richter statistics in topologically realistic system-level earthquake stress-evolution simulations. *Earth Planets Space* 55(8):761–771
51. Rundle JB, Tiampo KF, Klein W, Martins JSS (2002) Self-organization in leaky threshold systems: The influence of near-mean field dynamics and its implications for earthquakes, neurobiology, and forecasting. *Proc Natl Acad Sci USA* 99(Suppl 1):2514–2521
52. Rundle JB, Turcotte DL, Shcherbakov R, Klein W, Sammis C (2003) Statistical physics approach to understanding the multiscale dynamics of earthquake fault systems. *Rev Geophys* 41(4):1019. doi:10.1029/2003RG000135
53. Rundle PB, Rundle JB, Tiampo KF, Donnellan A, Turcotte DL (2006) Virtual California: fault model, frictional parameters, applications. *Pure Appl Geophys* 163:1819–1846
54. Sammis CG, Bowman DD (2006) Competing models for accelerating moment release before large earthquakes. 5th Annual ACES International Workshop, Maui, Hawaii, USA
55. Sammis CG, Bowman DD, King G (2004) Anomalous seismicity and accelerating moment release preceding the 2001–2002 earthquakes in northern Baha California, Mexico. *Pure Appl Geophys* 161:2369–2378
56. Scholz CH (2002) *The Mechanics of Earthquakes & Faulting*, 2nd edn. Cambridge University Press, Cambridge
57. Schorlemmer D, Jackson DD, Gerstenberger M (2003) Earthquake likelihood model testing. <http://moho.ess.ucla.edu/~kagan/sjg.pdf>. Accessed 8 Oct 2004
58. Shebalin P, Keilis-Borok V, Zaliapin I, Uyeda S, Nagao T, Tsybin N (2004) Advance short-term prediction of the large Tokachi-oki earthquake, September 25,  $M = 8.1$ : A case history. *Earth Planets Space* 56:715–724
59. Stein RS (1999) The role of stress transfer in earthquake occurrence. *Nature* 402:605–609
60. Swets JA (1973) The relative operating characteristic in psychology. *Science* 182:990–1000
61. Sykes LR, Jaumé SC (1990) Seismic activity on neighboring faults as a long-term precursor to large earthquakes in the San Francisco Bay area. *Nature* 348:595–599
62. Sykes LR, Shaw BE, Scholz CH (1999) Rethinking earthquake prediction. *Pure Appl Geophys* 155:207–232
63. Tiampo KF, Rundle JB, McGinnis S, Gross SJ, Klein W (2002) Eigenpatterns in southern California seismicity. *J Geophys Res* 107(B12):2354. doi:10.1029/2001JB000562

64. Tiampo KF, Rundle JB, McGinnis S, Gross SJ, Klein W (2002) Pattern dynamics and forecast methods in seismically active regions. *Pure Appl Geophys* 159:2429–2467
65. Turcotte DL (1991) Earthquake prediction. *Ann Rev Earth Planet Sci* 19:263–281
66. Turcotte DL (1997) *Fractals & Chaos in Geology & Geophysics*, 2nd edn. Cambridge University Press, Cambridge
67. Utsu T (1984) Estimation of parameters for recurrence models of earthquakes. *Earthquake Res Insti-Univ Tokyo* 59:53–66
68. Utsu T (2003) A list of deadly earthquakes in the world: 1500–2000. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake & Engineering Seismology*. Academic Press, Amsterdam, pp 691–717
69. Ward SN (1992) An application of synthetic seismicity in earthquake statistics: the Middle America trench. *J Geophys Res* 97(B5):6675–6682
70. Ward SN (1996) A synthetic seismicity model for southern California: cycles, probabilities, and hazard. *J Geophys Res* 101(B10):22393–22418
71. Ward SN (2000) San Francisco Bay Area earthquake simulations: a step toward a standard physical earthquake model. *Bull Seismo Soc Am* 90(2):370–386
72. Working Group on California Earthquake Probabilities (1988) Probabilities of large earthquakes occurring in California on the San Andreas fault. Open-File Report 88-398, US Geological Survey
73. Working Group on California Earthquake Probabilities (1990) Probabilities of large earthquakes in the San Francisco Bay region, California. Circular 1053, US Geological Survey
74. Working Group on California Earthquake Probabilities (1995) Seismic hazards in southern California: probable earthquakes, 1994–2024. *Seis Soc Am Bull* 85:379–439
75. Working Group on California Earthquake Probabilities (2003) Earthquake probabilities in the San Francisco Bay Region, 2002–2031. Open-File Report 2003-214, US Geological Survey
76. Wyss M (1997) Nomination of precursory seismic quiescence as a significant precursor. *Pure Appl Geophys* 149:79–114
77. Wyss M, Habermann RE (1988) Precursory seismic quiescence. *Pure Appl Geophys* 126:319–332
78. Yakovlev G, Turcotte DL, Rundle JB, Rundle PB (2006) Simulation-based distributions of earthquake recurrence times on the San Andreas fault system. *Bull Seis Soc Am* 96:1995–2007
79. Yang W, Vere-Jones D, Li M (2001) A proposed method for locating the critical region of a future earthquake using the critical earthquake concept. *J Geophys Res* 106:4151–4128
80. Zechar JD, Jordan TH (2005) Evaluation techniques for alarm-based forecasts. EOS Trans. AGU. Fall meeting
81. Zoback ML (1992) First- and second-order patterns of stress in the lithosphere: The world stress map project. *J Geophys Res* 97:11703–11728

### Books and Reviews

- Jolliffe IT, Stephenson DB (2003) *Forecast Verification*. Wiley, Chichester
- Turcotte DL, Schubert G (2002) *Geodynamics*. Cambridge University Press, Cambridge
- Wilks DS (1995) *Statistical Methods in the Atmospheric Sciences*. Academic Press, San Diego

## Earthquake Location, Direct, Global-Search Methods

ANTHONY LOMAX<sup>1</sup>, ALBERTO MICHELINI<sup>2</sup>,  
ANDREW CURTIS<sup>3</sup>

<sup>1</sup> ALomax Scientific, Mouans-Sartoux, France

<sup>2</sup> Istituto Nazionale di Geofisica e Vulcanologia,  
Roma, Italy

<sup>3</sup> ECOSSE (Edinburgh Collaborative  
of Subsurface Science and Engineering),  
Grant Institute of GeoSciences, The University  
of Edinburgh, Edinburgh, United Kingdom

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[The Earthquake Location Problem](#)

[Location Methods](#)

[Illustrative Examples](#)

[Future Directions](#)

[Bibliography](#)

### Glossary

**Arrival time** The time of the first measurable energy of a seismic phase on a seismogram.

**Centroid** The coordinates of the spatial or temporal average of some characteristic of an earthquake, such as surface shaking intensity or moment release.

**Data space** If the data are described by a vector  $\mathbf{d}$ , then the data space  $D$  is the set of all possible values of  $\mathbf{d}$ .

**Direct search** A search or inversion technique that does not explicitly use derivatives.

**Earthquake early-warning** The goal of earthquake early-warning is to estimate the shaking hazard of a large earthquake at a nearby population center or other critical site before destructive  $S$  and surface waves have reached the site. This requires that useful, probabilistic constraint on the location and size of an earthquake is obtained very rapidly.

**Earthquake location** An earthquake location specifies a spatial position and time of occurrence for an earthquake. The location may refer to the earthquake hypocenter and corresponding origin time, a mean or centroid of some spatial or temporal characteristic of the earthquake, or another property of the earthquake that can be spatially and temporally localized. This term also refers to the process of locating an earthquake.

**Epicenter** The point on the Earth's surface directly above a hypocenter.

**Error** A specified variation in the value assumed by a variable. See also *uncertainty*.

**Global search** A search or inversion that samples throughout the prior *pdf* of the unknown parameters.

**Hypocenter** The point in three-dimensional space of initial energy release of an earthquake rupture or other seismic event.

**Importance sampling** A sampling procedure that draws samples following the posterior *pdf* of an inverse, optimization or other search problem. Since these problems involve initially unknown, posterior *pdf* functions, importance sampling can only be performed approximately, usually through some adaptive or learning procedure as sampling progresses.

**Inverse problem, inversion** The problem of determining the parameters of a physical system given some data. The solution of an inverse problem requires measurements of observable quantities of the physical system, and the mathematical expression (the forward problem) that relates the parameters defining the physical system (model space) to the data (data space). In inverse problems, estimates of the unknown parameters in the model space and of their uncertainties are sought from the combination of the available information on the model parameters (prior *pdf*), the data and the forward problem.

**Likelihood function** A non-normalized *pdf*.

**Misfit function** A function that quantifies the disagreement between observed and calculated values of one or more quantities. See *objective function*.

**Model space** If the model parameters are described by a vector  $\mathbf{m}$ , then model space  $M$  is the set of all possible values of  $\mathbf{m}$ .

**Objective function** A function expressing the quality of any point in the model space. Inversion and optimization procedures use an objective function to rank and select models. Usually objective functions are defined in terms of misfit functions, and for probabilistic inversion the objective function must be a *pdf* or likelihood function.

**Origin time** The time of occurrence of initial energy release of an earthquake rupture or other seismic event.

**Prior pdf** A *pdf* that expresses the information on the unknown parameters available before an inverse problem is solved. For an earthquake location, the prior *pdf* is often a simple function (e. g., boxcar) of three spatial dimensions and time. See also *Inverse problem*.

**Probability density function – pdf** A function in one or more dimensional space  $X$  that (i) when integrated

over some interval  $\Delta \mathbf{x}$  in  $X$  gives a probability of occurrence of any event within  $\Delta \mathbf{x}$ , and (ii) has unit integral over space  $X$ , where  $X$  represents a space of possible events. An earthquake location *pdf* is often a 3-dimensional probability density function over all possible spatial locations or a 4-dimensional probability density function over all possible spatial locations and times of occurrence.

**Posterior pdf** A *pdf* that expresses the information about the unknown parameters available after inversion. The posterior *pdf* for an earthquake location is often a function of the three spatial dimensions and the origin time of the hypocenter parameters; this function may be complicated. See also *Inverse problem*.

**Ray path** A local minimum-time path between a source and receiver of idealized, infinite frequency wave energy of a specified wave type (e. g.,  $P$  or  $S$ ).

**Receiver or station** Synonyms for an observation point where ground motion is detected and a seismogram recorded.

**Seismic phase** A distinct packet of energy from a seismic source. Usually refers to a specified wave type (e. g.  $P$  or  $S$ ) satisfying a particular physics of wave propagation.

**Seismicity** The distribution in space and time of seismic event locations.

**Seismogram** An analogue or digital recording of the ground motion at a point (receiver or station) in the Earth. Also called a waveform.

**Source** A general term referring to an earthquake, explosion or other release of seismic energy as a physical phenomenon localized in space and time.

**Station** See *receiver*.

**Travel time** The time that a signal, e. g. elastic wave energy of a seismic phase, takes to propagate along a ray path between two points in a medium.

**Uncertainty** Random variation in the values assumed by a variable. See also *error*.

## Definition of the Subject

An earthquake location specifies the place and time of occurrence of energy release from a seismic event. A location together with a measure of size forms a concise description of the most important characteristics of an earthquake. The location may refer to the earthquake's epicenter, hypocenter, or centroid, or to another observed or calculated property of the earthquake that can be spatially and temporally localized. A location is called *absolute* if it is determined or specified within a fixed, geographic coordinate system and a fixed time base (e. g., Coordinated

Universal Time, UTC); a location is called *relative* if it is determined or specified with respect to another spatio-temporal object (e. g., an earthquake or explosion) which may have unknown or uncertain absolute location.

For rapid hazard assessment and emergency response, an earthquake location provides information such as the locality of potential damage or the source region of a possible tsunami, and a location is required to calculate most measures of the size of an earthquake, such as magnitude or moment. Locations are required for further analysis and characterization of the event, for studies of general patterns of seismicity, to calculate distributions of stress and strain changes around the earthquake, for assessing future earthquake hazard, and for basic and applied seismological research.

Since earthquakes occur deep in the Earth, their source locations must be inferred indirectly from distant observations, and earthquake location is thus a remote-sensing problem. Most commonly an earthquake location is determined by the match or misfit between observed arrival times of seismic wave-energy at seismic stations, and predictions of these arrival times for different source locations using a given elastic-wave speed model; this is an inverse problem. Essentially, many potential locations (place and time) are examined and those for which some measure of misfit between predicted and measured arrival times is smallest are retained as best estimates of the true location.

Many numerical location methods involve linearization of the equations relating the predicted arrival times to the location through Taylor expansion involving partial derivatives; these are called *linearized* methods. Methods that do not involve linearization are called *nonlinearized* or *direct-search* methods. The term *nonlinear* is used ambiguously in geophysics to refer to linearized-iterated and to nonlinearized methods. In this chapter we focus on nonlinearized, direct-search methods, and to avoid ambiguity we identify them with the term *direct-search*.

Direct-search location can be performed through graphical analysis, regular or stochastic searches over a space of possible locations, and other algorithms. Direct-search earthquake location is important because, relative to linearized methods, it is easy to apply with realistic earth models which may have abrupt and complicated velocity variations in three-dimensions, it places little restriction on the form of the measure of misfit, it is stable (i. e., does not suffer numerical convergence problems) when the observations are insufficient to fully constrain the spatial location or origin time, and it can produce comprehensive, probabilistic solutions which indicate the full location uncertainty, often a complex function of space and time.

Conversely, the primary advantage of linearized location methods is that they are much less demanding computationally than direct-search methods.

## Introduction

Most commonly, an earthquake location is determined using observed seismic-phase arrival-times and associated uncertainties, and predicted travel times in a given wave-speed model. Ideally, the location procedure will determine a 4-dimensional, posterior probability density function, or location *pdf*, over all possible solutions (spatial locations and origin times). This location *pdf* quantifies the agreement between predicted and observed arrival times in relation to all uncertainties, and forms a complete, probabilistic solution. In practice, however, an earthquake location is often specified as some optimal solution (a point in space and time) with associated uncertainties.

The earliest, formal earthquake locations using seismic-phase arrival-time observations employed direct-search procedures such as graphical methods (e. g., [37]) or simple grid searches (e. g., [52]). The advent of digital computers in the 1960's lead to the use of iterated linearized approaches based mainly on Geiger's method [17]. Since the 1980's, the increasing power of digital computers has made large-scale, grid and stochastic direct searches practical for routine earthquake location. Direct-search methods are now used routinely in research and earthquake monitoring (e. g., [22,23,31,32,46,48,61]).

In principle, direct-search methods can be applied to locate the relative positions of ensembles of events, and for joint epicentral determination (e. g., [47]) to simultaneously determine multiple earthquake locations and station corrections related to errors in the velocity model. However, the high-dimensionality of such problems makes direct-search solution difficult and computationally demanding; at the present time these problems are usually performed through large scale, linearized procedures. For these reasons, we mainly consider here absolute location of individual events.

In this article we describe the earthquake location problem and direct-search methods used to perform this location, and we present a number of examples of direct-search location. We do not compare different direct-search location methods or compare direct-search to linearized algorithms, but instead focus on illustrating important features and complexity in earthquake location results. For this reason we emphasize direct, global-search, probabilistic location, which produces general and complete solutions that best illuminate these features and complexity.

## The Earthquake Location Problem

### An Inherently Nonlinear Problem

In a homogeneous medium with wave speed  $v$  and *slowness* defined to be  $u = 1/v$ , the arrival time,  $t_{\text{obs}}$ , at an observation point  $x_{\text{obs}}, y_{\text{obs}}, z_{\text{obs}}$  of a signal emitted at origin time  $t_0$  from a source location at  $x_0, y_0, z_0$  is,

$$t_{\text{obs}} = t_0 + u \left[ (x_{\text{obs}} - x_0)^2 + (y_{\text{obs}} - y_0)^2 + (z_{\text{obs}} - z_0)^2 \right]^{1/2}, \quad (1)$$

This expression shows that a change in the spatial position of the source introduces a nonlinear change in  $t_{\text{obs}}$ , even in the simplest possible medium. When the speed  $v$  and hence slowness  $u$  are inhomogeneous in space, the arrival time at the observation point becomes,

$$t_{\text{obs}} = t_0 + \int_{\mathbf{r}_0(s)} u(\mathbf{r}_0) ds, \quad (2)$$

where  $\mathbf{r}_0(s)$  denotes a point at distance  $s$  along ray path  $\mathbf{r}_0$  between source and receiver locations. Equation (1) is a special case of (2) that has straight source-receiver ray paths. Equation (2) is nonlinear since a change in the source location changes the ray path over which the integral is calculated. Thus, earthquake location, which maps arrival times into spatial location and origin time, is inherently a nonlinear problem.

### The Observed Data

Data used to constrain earthquake locations are usually derived from seismograms recorded at seismic stations distributed around the earthquake source area, usually at or near the surface of the Earth. The derived data for earthquake location include arrival times, polarization angles, or array slownesses and azimuths. For earthquake location there are three important aspects of this data determination:

- 1) choosing locations for the stations (before data have been collected),
- 2) deriving data and associated uncertainties from the seismograms, and
- 3) association of the derived data into subsets of data corresponding to unique events.

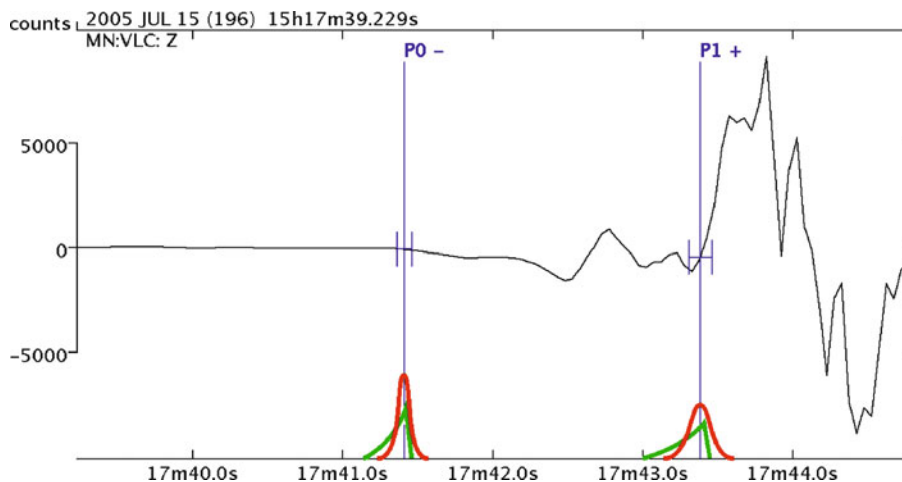
The first important aspect of data determination is choosing station locations with the goal of constraining as tightly as possible event locations for a given source area; this is classified as a problem of "experimental design" in the field of statistics. The design problem must be resolved prior to

data collection and so is posed in terms of *expected* data, and *expected* location results. We describe experimental design techniques in more detail later, after introducing and discussing the location solution on which such designs depend.

Once stations are installed and have recorded seismograms from earthquakes of interest, a data set must be extracted that is sensitive to the event source location, and which we can associate with some physics (e.g., of  $P$  or  $S$  waves) and paths of wave propagation. Most commonly for earthquake location the data set will be phase arrival times and associated uncertainties picked manually or automatically from seismograms (Fig. 1). It is often easy to detect and pick arrivals manually since the human eye can identify a change in amplitude or frequency in the signal even in the presence of significant noise. The picking of the  $S$  phase is sometimes more difficult because it arrives in the  $P$  coda and can be preceded by  $S$  to  $P$  or other converted phases; this is a common problem with recordings at local (e.g., up to about 100 km) and near-regional (e.g., up to about 300 km) distances, especially if horizontal component seismograms are not available. The automatic detection, identification and picking of  $P$  and  $S$  arrivals is much more difficult, especially in the presence of high noise levels. However, automatic detection and picking is faster and, for the case of initial  $P$  phases or other phases with characteristic forms, can produce a more con-

sistent data set than manual processing. Automatic arrival detection relies on identifying temporal variations in energy, frequency content, polarization or other characteristics of the signal which are anomalous relative to their background or noise level. Often the detection and picking algorithms are applied to filtered and processed time-series in order either to reduce noise, or to augment the signal in pre-set or dynamically-determined frequency bands or polarization directions. See [18] for an approach that exploits neural networks for phase identification and picking, and [82] for a review and systematic comparison of several approaches to automatic detection and picking.

The data used for earthquake location (e.g., arrival times) must have associated uncertainty estimates otherwise the location uncertainty and a probabilistic solution (i.e., location *pdf*) can not be calculated. Most generally, a vector  $\mathbf{d}$  that describes the data takes values from a data space  $\mathbf{D}$ , and  $p(\mathbf{d})$  denotes the *pdf* representing uncertainty in  $\mathbf{d}$ . The uncertainty in arrival time data should include not only an estimate of the uncertainty in the picked phase arrival time, but also uncertainty in which phase (e.g.,  $P$  or  $S$ ) is associated with the pick. If there are multiple expected phase arrivals close to the picked arrival time of a phase, then ideally these should all be taken as candidate phase types for the arrival. Also, the pick uncertainty of each phase may be best described by a *pdf* that is asymmetric in time, since usually a latest-possible time



Earthquake Location, Direct, Global-Search Methods, Figure 1

A short waveform segment ( $\sim 5$  sec) showing the first  $P$  wave arrivals from a small earthquake in Northern Italy recorded on a vertical component seismogram at a nearby station. Automatic arrival pick times (vertical blue lines) and uncertainty estimates (blue error bars) are shown for two phases, a first arriving  $P$  phase ( $P_0$ ) and secondary  $P$  arrival ( $P_1$ ). The red curves show the data *pdf* functions representing these arrival pick times and uncertainties for an event location procedure where the data  $P(\mathbf{d})$  is approximated by a normal distribution. The green curves show irregular, asymmetric *pdf* functions that may more accurately represent the uncertainty in the phase arrival times; if such *pdf* functions were routinely estimated during arrival picking, they could be used for direct-search location without major difficulty



for a pick is much easier to define than the earliest time (Fig. 1). True data uncertainty *pdf*'s are therefore generally multi-modal, and can be quite complex to calculate and parametrize. In practice, an enumerated quality indication or, at best, a simple normal distribution (Gaussian uncertainty) is used to describe the picking error, and the phase association is usually fixed (e. g. to *P* or *S*) so corresponding uncertainties are ignored. In many cases these simplified data uncertainty estimates will lead to bias or increased error in the resulting event locations.

The third important aspect of data determination is the association of the derived data into sets of data for unique events. For example, this association may entail the assignment of each observed arrival time within a specified time window to a unique event, forming the minimum possible number of events and corresponding arrival time sets required to explain observed data. This association procedure can be very difficult, especially with automatic systems and when there are signals from multiple seismic events that are close or overlapping in time (e. g., [24]), and we do not address this issue further here. In the following, except for an examination of outlier data, we implicitly assume that location is performed with a data set that is already associated to a unique event.

### The Velocity or Slowness Model

The velocity or slowness model specifies seismic wave-speeds in the region of the Earth containing the sources, the receivers and the ray paths between the sources and receivers. Equation (2) is nonlinear with respect to source location, but also with respect to slowness since a change in the slowness distribution of the medium changes the ray path. The velocity structure is sometimes estimated through coupled, simultaneous inversions for velocity structure and event locations (commonly called seismic tomography), but these are very large inverse problems solved mainly with linearized methods. Usually for earthquake location the velocity model is taken as known and fixed.

Often, for computational convenience or due to lack of information, the velocity model is parametrized with velocity varying only with depth. This is commonly called a laterally homogeneous or 1-dimensional (1D) model. Such a model may consist of one or more layers of constant or vertical-gradient wave-speeds. For work at a local or near-regional scale the layers may be horizontal and flat; for larger, regional or global scale problems the layers should be spherically symmetric shells to represent the curvature of the Earth. When more information on the velocity structure is available, a 3D model may be used

in order to increase the accuracy of the ray paths and travel times, and hence of the locations, relative to a 1D model. All models, whether 1D or 3D, are described by a limited number of parameters and include some form of spatial averaging or interpolation with respect to the true Earth. Although 3D models can potentially represent velocity variations in the Earth more accurately than 1D models, in practice the velocities in 3D models can locally be poorly constrained and have large errors. It is therefore often important to consider several different possible 1D and 3D velocity structures in a location study, either to test the sensitivity of the locations to errors in velocity, or to better estimate the travel-time uncertainties and produce a more meaningful location *pdf*. In principle the use of diverse velocity models poses no difficulties with direct-search location methods.

### The Travel-Time Calculation

The theoretical seismic wave travel-times through a given velocity model between any particular source and receiver locations are required by most location algorithms. The calculation of travel times is commonly referred to as forward modeling, because inverse theory need not be invoked. There are three basic classes of methods to calculate the travel times: full-waveform methods, ray methods, and Huygens wavefront or eikonal methods.

Full-waveform methods (e. g., [1]) produce complete synthetic seismograms from which predicted travel times can be extracted. These methods include frequency-wavenumber or modal-summation techniques which are valid for a broad range of frequencies and can produce exact waveforms, but which are only applicable for relatively simple velocity structures. Numerical techniques such as finite elements and finite differences can accurately model full wave phenomena in complicated structures, but these methods typically require large computing resources and computing time. Currently, full-waveform methods are rarely used to determine predicted travel times for earthquake location because these times can be obtained directly and more efficiently with ray and eikonal methods.

Ray methods (e. g., [1,7,75]) provide travel times and the path, or ray, traveled by high-frequency waves, and can be applied to complicated and 3D velocity structures. With simple model parametrizations such as flat layers with constant or gradient velocity, ray paths and travel times can be determined very rapidly with analytical or semi-numerical algorithms. For these and more complicated models, shooting, or ray tracing techniques generate rays by iteratively solving of a set of ray-tracing equations starting in a specified direction at the source or receiver lo-

cation. The ray that passes through a specified end point is found by a search over the direction at the starting point; this search can be time consuming or unstable. In addition, shooting methods do not produce diffracted arrivals (e. g., “head waves” from the Mohorovičić discontinuity) which are often the first arriving signal at near-regional distances and are thus critical for earthquake location. Two-point, ray bending and perturbation techniques rely on Fermat’s principle of least time: an initial guess at the ray between two points is perturbed repeatedly to attain a minimum travel time and corresponding ray between the points. These techniques perform best with smooth models, but do produce diffracted arrivals. In general, except for analytical or semi-numerical algorithms in simple models, ray methods are too computationally expensive for direct-search location, which usually requires evaluation of travel times between a very large number of source and receiver positions. However, some ray bending methods (e. g., [39,77]) are efficient enough to be used in direct-search location when a relatively small number of source and receiver positions need to be examined.

Wavefront, eikonal and graph-based methods [75] provide travel-times of the first arriving, high frequency waves including diffracted arrivals, and are efficient and applicable with complicated, 3D velocity structures. In effect, these methods propagate wavefronts through a velocity model with repeated application of Huygen’s principle, by considering a large number of virtual sources (*Huygens sources*) along each wavefront. At time  $t$  these sources emit circular wavelets which expand for a small time  $\Delta t$  through the (constant) local, medium velocity. The locus of the first arriving circular wavelets defines the new wavefront location at time  $t + \Delta t$ . The synthetic travel time of the first-arriving energy at the receiver is the time at which a wavefront first touches the receiver. In practice, this problem is solved on a computer either by replicating this “wavefront marching” process (e. g., [50,66]), or by finding a numerical solution to the eikonal equation (e. g., [44,79]), or by graphical analysis (e. g., [38]). Though wavefront, eikonal and graph-based methods produce directly only the travel time of the first-arriving signal, information about the path traveled by the signal can be derived numerically from the travel-time field or from ray-tracing, and travel-times of secondary arrivals can be obtained through multi-stage calculations (e. g., [44,51]).

Wavefront, eikonal and graph-based methods can efficiently generate the travel-times from one point in a gridded velocity model to all other points in the model. This makes these methods particularly useful for direct-search location, which may test a large number of possible source positions widely distributed in space. For this purpose, the

travel times from each seismic station to all points in the model can be pre-calculated and stored in computer disk files or in memory; obtaining the travel time from a station to any other point then reduces to a simple lookup (e. g., [35,38]).

### A Complete Solution – Probabilistic Location

Consider a vector  $\mathbf{d}_{\text{obs}}$  of observed data (e. g., arrival times) that takes values in a data space  $\mathbf{D}$ , and let  $p(\mathbf{d})$  be the *pdf* over  $\mathbf{D}$  describing the data uncertainty in  $\mathbf{d}_{\text{obs}}$  due to measurement and processing uncertainties. Similarly, let  $\mathbf{m}$  denote the vector of source location parameters (spatial coordinates and origin time) which take values from parameter space  $\mathbf{M}$ . Let  $p(\mathbf{m})$  be the prior *pdf* representing all information available about the location before (prior to) using the data  $\mathbf{d}_{\text{obs}}$ ;  $p(\mathbf{m})$  might include knowledge of the known, active fault zones in the area, or might specify the bounds of a region within which we know the event occurred from damage reports, or of a region containing the network of stations that recorded the event. Also consider the forward problem (e. g., travel time calculation) relating  $\mathbf{m}$  to a vector of predicted data (e. g., arrival times),  $\mathbf{d}_{\text{calc}}$ . In general the forward problem may also be uncertain, for example due to uncertainties in velocity structure, so we use  $F(\mathbf{d}, \mathbf{m})$  to denote the *pdf* of the relationship between  $\mathbf{d}_{\text{calc}}$  and  $\mathbf{m}$  as constrained by the forward problem.

As an example of  $F$ , it is commonly assumed that for each particular  $\mathbf{m}$ , the corresponding predicted data  $\mathbf{d}_{\text{calc}}$  are given by a function  $\mathbf{f}(\mathbf{m})$  with negligible errors. Then the conditional *pdf*  $F(\mathbf{d}|\mathbf{m})$  (the probability distribution of  $\mathbf{d}$  when  $\mathbf{m}$  is fixed at a particular value) is described by  $F(\mathbf{d}|\mathbf{m}) = \delta[\mathbf{d} - \mathbf{f}(\mathbf{m})]$  where  $\delta$  is the Dirac delta-function. Also, the forward problem is often assumed to place minimum possible constraint on parameters  $\mathbf{m}$ ; the *pdf* describing this state of information about  $\mathbf{m}$  is called the homogeneous distribution, represented by  $\mu(\mathbf{m})$ . No *pdf* exists that describes zero information, but *some* information about  $\mathbf{m}$  always exists in practice (the positivity of parameter values, for example);  $\mu(\mathbf{m})$  describes that minimum state of information. In that case the forward problem is given by [73,74],

$$F(\mathbf{d}, \mathbf{m}) = \delta[\mathbf{d} - \mathbf{f}(\mathbf{m})] \mu(\mathbf{m}) . \quad (3)$$

A solution to the earthquake location problem is found by combining the information in the observed data,  $p(\mathbf{d})$ , the prior *pdf*,  $p(\mathbf{m})$ , and the ability of the forward problem to predict the observed data,  $F(\mathbf{d}, \mathbf{m})$  [73,74]. This is achieved in a probabilistic framework by constructing a *pdf*  $Q$  describing the state of posterior (post-experimental) infor-

mation by:

$$Q(\mathbf{d}, \mathbf{m}) = k \frac{p(\mathbf{d})F(\mathbf{d}, \mathbf{m})p(\mathbf{m})}{\mu(\mathbf{d}, \mathbf{m})}, \quad (4)$$

where the constant  $k$  normalizes  $Q$  to unit integral over  $\mathbf{D} \times \mathbf{M}$  and  $\mu(\mathbf{d}, \mathbf{m})$  is the homogeneous distribution over data  $\mathbf{d}$  and parameters  $\mathbf{m}$ . Equation (4) contains all information (from the prior knowledge, data and physics) that could have a bearing on location  $\mathbf{m}$ , and defines a joint *pdf* between parameters  $\mathbf{m}$  and data  $\mathbf{d}$ . The final, posterior state of information about location parameters  $\mathbf{m}$  is given by integrating over the data  $\mathbf{d}$  to obtain the marginal posterior *pdf*,

$$Q(\mathbf{m}) = k p(\mathbf{m}) \int_{\mathbf{D}} \frac{p(\mathbf{d})F(\mathbf{d}, \mathbf{m})}{\mu(\mathbf{d}, \mathbf{m})} d\mathbf{d}. \quad (5)$$

Equation (5) is the general, probabilistic solution to the inverse problem of event location from the available data since it describes the uncertainty in event location  $\mathbf{m}$  given all available information. It is usual to call the integral in (5) the *likelihood* function  $L(\mathbf{m})$ , which gives a (non-normalized) measure of how good any model  $\mathbf{m}$  is in explaining the observed data  $p(\mathbf{d})$ .

As mentioned earlier it is often the case that  $p(\mathbf{d})$  for the observed data is approximated by a Gaussian distribution, described by mean  $\mathbf{d}_0$  and covariance matrix  $\mathbf{C}_d$ . Assuming that the uncertainties in the forward problem  $F$  relating  $\mathbf{d}$  and  $\mathbf{m}$  are negligible results in the form of  $F$  in Eq. (3). It is also usually assumed that  $\mathbf{d}$  and  $\mathbf{m}$  are independent and hence that  $\mu(\mathbf{d}, \mathbf{m})$  can be written  $\mu(\mathbf{d})\mu(\mathbf{m})$ ;  $\mu(\mathbf{d})$  is usually taken to be constant. With these simplifications, used by many current direct-search location procedures, the (non-normalized) likelihood function is given by,

$$L(\mathbf{m}) = \exp \left\{ -\frac{1}{2} [\mathbf{d}_0 - \mathbf{f}(\mathbf{m})]^T \mathbf{C}_d^{-1} [\mathbf{d}_0 - \mathbf{f}(\mathbf{m})] \right\}. \quad (6)$$

With the above simplifications a maximum likelihood origin time,  $t_0$ , can be determined analytically from weighted means of the observed arrival times and the predicted travel times (e. g., [74]), and if the observed and predicted times are uncorrelated we arrive at a likelihood function,

$$L(\mathbf{x}) = \exp \left\{ -\frac{1}{2} \sum_i \frac{[T_i^o - T_i^c(\mathbf{x})]^2}{\sigma_i^2} \right\}, \quad (7)$$

where  $\mathbf{x}$  is the spatial part of  $\mathbf{m}$ ,  $T_i^o$  are observed travel times,  $T_i^c$  are the calculated travel times for observation  $i$  (i. e.,  $T_i^c$  represents the travel time, rather than arrival time,

part of  $\mathbf{f}(\mathbf{m})$ ), and  $\sigma_i$  summarizes the associated standard deviation of uncertainty in  $T_i^o$  and  $T_i^c$ .

Though not normalized,  $L(\mathbf{x})$  is sufficient to provide the *relative* probability of any location  $\mathbf{m}$  being the best estimate of the event location given the available data measurements. Since in practice integrating over all of  $\mathbf{D} \times \mathbf{M}$  to find normalizing constant  $k$  in Eq. (5) is often computationally intractable, the product of the prior, spatial location information  $p(\mathbf{x})$  (i. e., the spatial part of  $p(\mathbf{m})$ ) and the non-normalized likelihood  $L(\mathbf{x})$  is usually taken as the objective function for inversion and searching in direct-search location algorithms. If  $L(\mathbf{x})$  is determined through-out the prior *pdf*  $p(\mathbf{x})$  through a global-search, then Eq. (5) can be normalized approximately after location. In the following text and examples, we refer to such an approximately normalized function,  $p(\mathbf{x})L(\mathbf{x})$ , as a location *pdf*.

The likelihood function in Eq. (5) is entirely defined by the probabilistic error processes involved. However, often it is desirable to change the approximations employed in deriving Eqs. (6) and (7) from Eq. (5), in order to remove biases or instability in the solution. The approximation in Eq. (6) uses the exponential of an L2-norm misfit function (the term in braces  $\{\}$  in Eq. (6) or (7)) to represent the *pdf* of the data error variation, but because data used for location often contain outliers it is often considered that an L1 norm or other  $L_p$  norm ( $p < 2.0$ ) is more appropriate (e. g., [69]), where  $L_p$ -norm  $|\mathbf{x}| = \sqrt[p]{\sum |x_i|^p}$ . Earthquake location problems formulated with an  $L_p$  norm (or indeed other kinds of likelihood functions – see Eq. (8) below), can be solved relatively easily with direct-search methods, which, unlike linearized methods, do not require determination of partial derivatives of the likelihood or objective function with respect to event location.

An alternative to  $L_p$ -likelihood functions that is very robust in the presence of outliers is given by the equal differential-time (EDT) formulation [16,31,84]. For the EDT case, the location likelihood is given by,

$$L(\mathbf{x}) = \left[ \sum_{a,b} \frac{1}{\sqrt{\sigma_a^2 + \sigma_b^2}} \cdot \exp \left( -\frac{\{ [T_a^o - T_b^o] - [TT_a^c(\mathbf{x}) - TT_b^c(\mathbf{x})] \}^2}{\sigma_a^2 + \sigma_b^2} \right) \right]^N, \quad (8)$$

where  $\mathbf{x}$  is the spatial part of  $\mathbf{m}$ ,  $T_a^o$  and  $T_b^o$  are the observed arrival times and  $TT_a^c$  and  $TT_b^c$  are the calculated travel times for two observations  $a$  and  $b$ ; the sum is taken over all pairs of observations, and  $N$  is the total number of observations. Standard deviations  $\sigma_a$  and  $\sigma_b$  summarize the

assigned uncertainties on the observed arrival times and calculated travel times, where it is assumed that the observed and the calculated times are uncorrelated.

In Eq. (8), the first and second terms in brackets in the exponent are, respectively, the differences between the observed arrival times and the differences between the calculated travel times. The exponent is the difference between these two terms, and thus the exponential has a maximum value of 1 which occurs at points  $\mathbf{x}$  where the two differences are equal (hence, the name “equal differential time”). Such points  $\mathbf{x}$  best satisfy the two observations  $a$  and  $b$  together, and, in general, the set of  $\mathbf{x}$  where the exponential is nonzero forms a “fat,” curved surface in 3D space. Because the summation over observations is outside the exponential, the EDT location *pdf* has its largest values for those points  $\mathbf{x}$  where the most pairs of observations are satisfied and thus is far less sensitive to outlier data than  $Lp$  norms which seek to best satisfy all of the observations simultaneously. Note that the EDT likelihood function  $L(\mathbf{x})$  does not require calculation of an origin time  $t_0$ ; this reduces the hypocenter search to a purely 3-parameter problem and contributes to the robustness of the EDT method. Nevertheless, a compatible estimate of  $t_0$  can be calculated for any hypocenter point  $\mathbf{x}$ .

Ultimately, the full solution to the probabilistic location problem is a posterior *pdf* which includes as comprehensive as possible uncertainty information over parameters  $\mathbf{m}$ . This may include multiple “locally-optimal” solutions, e. g.,  $Q(\mathbf{m})$  or  $p(\mathbf{x})L(\mathbf{x})$  may have multiple maxima, and may have a highly irregular form. Some studies of seismicity and seismotectonics make explicit use of a probabilistic representation of seismic event locations (e. g., [22,31,46]).

### Experimental Design Methods – Choosing Receiver Locations

As noted earlier, it is important to position stations so as to constrain as tightly as possible the event locations for a given source area. The location inverse problem solution in Eqs. (5), (7) or (8) is constrained by prior information on location  $p(\mathbf{m})$ , by observed data  $p(\mathbf{d})$ , and by forward-problem physics relating  $\mathbf{d}$  and  $\mathbf{m}$ . One way to significantly influence the form of this inverse problem, and hence uncertainty in its solution, is to change the data we record. Thus, we alter both  $p(\mathbf{d})$  and the forward-problem physics,  $F(\mathbf{d}, \mathbf{m})$ .

For seismic location problems we may change the data by employing experimental design methods to choose or change the locations of seismic receivers. The goal of the design procedure is to place receivers such that the loca-

tion information described by solution  $Q(\mathbf{m})$  is expected to be maximized. This is a “macro-optimization” problem where, prior to the occurrence of an earthquake, we optimize the design of the inverse problem that we expect to solve after an earthquake has occurred.

The design is varied such that it maximizes an objective function. This is usually taken to be the expected value of some approximation to the unique measure of information that was discovered by Shannon [67],

$$J(\mathbf{R}) = E_{\mathbf{m}_t} \{I[Q(\mathbf{m}); \mathbf{R}, \mathbf{m}_t]\} \quad (9)$$

where  $\mathbf{R}$  is a vector describing the design (e. g., receiver locations),  $I[Q(\mathbf{m}); \mathbf{R}, \mathbf{m}_t]$  is the information contained in the resulting posterior *pdf*  $Q(\mathbf{m})$  for design  $\mathbf{R}$  when the true parameters (e. g., event location) is  $\mathbf{m}_t$ , and the statistical expectation  $E_{\mathbf{m}_t}$  is taken over all possible  $\mathbf{m}_t$  which (according to our prior knowledge) are expected to be distributed according to the prior distribution  $p(\mathbf{m})$ .  $J(\mathbf{R})$  should be maximized.

Within the expectation in Eq. (9), the design criterion  $J(\mathbf{R})$  takes account of all possible potential true event locations  $\mathbf{m}_t$ , their prior probability of occurrence  $p(\mathbf{m})$ , and the corresponding data (including their uncertainties) that are expected to be recorded for each location (the latter are included within  $Q(\mathbf{m})$ ). To calculate the expectation usually requires integration over a far greater proportion of the model and data spaces,  $\mathbf{M}$  and  $\mathbf{D}$  respectively, than need be considered when solving the inverse problem after a particular event has occurred (since then  $p(\mathbf{d})$  and hence  $Q(\mathbf{m})$  are fixed, and  $p(\mathbf{m})$  is more tightly constrained). Consequently, experimental design is generally far more computationally costly than solving any particular inverse problem post-event.

For this reason, design methods invoking linearized approximations to the model-data relationship  $F(\mathbf{m}, \mathbf{d})$  (e. g., Eq. (10) below) have been employed by necessity in the past [11,13,49,70], or indeed non-probabilistic methods have been employed (e. g., [9,10,36,71]). Truly non-linearized design methods have been developed for location problems only relatively recently [12,78,80]. Historically, however, station network geometry has been defined more by heuristics (rules of thumb) and geographical, logistical and financial constraints, with design theory only recently being deployed.

### Location Methods

Once data have been recorded and prior *pdf*'s defined, a solution such as that of Eqs. (5), (7) or (8) must be evaluated throughout the prior *pdf*,  $p(\mathbf{x})$ , to identify one or more “locally-optimal” solutions, or, preferably, to obtain

a full probabilistic location *pdf*. This evaluation generally requires direct-search optimization and search techniques, which we discuss below. We first digress and summarize linearized location procedures, which typically determine a single optimal hypocenter along with a simplified and approximate representation of the location *pdf* (e. g., a confidence ellipsoidal centered on the estimated hypocenter and origin time).

### Linearized Location Methods

With linearized methods the arrival time expression (2), which is nonlinear with respect to the spatial location  $\mathbf{m} = (x, y, z)$ , is approximated by a Taylor series expansion around some prior estimate  $\mathbf{m}_0 = (x_0, y_0, z_0)$  of the spatial location:

$$f(\mathbf{m}) = f(\mathbf{m}_0) + (\mathbf{m} - \mathbf{m}_0)f'(\mathbf{m}_0) + \frac{(\mathbf{m} - \mathbf{m}_0)^2}{2}f''(\mathbf{m}_0) + O[(\mathbf{m} - \mathbf{m}_0)^3] \quad (10)$$

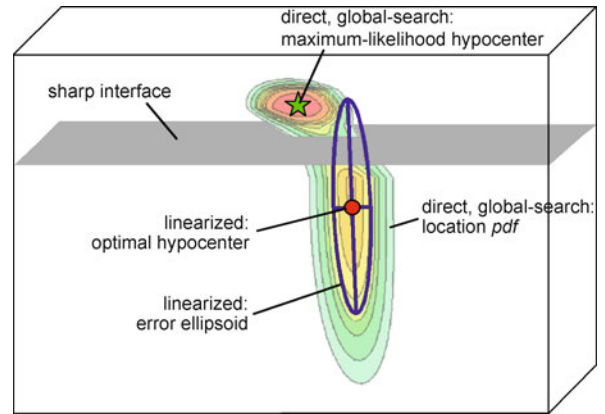
where  $f(\mathbf{m})$  is the forward problem that calculates an arrival time  $d_{\text{calc}}$  given a location  $\mathbf{m}$  (e. g.,  $f(\mathbf{m})$  might represent the right hand side of Eq. (2) directly). A linear vector-matrix inverse problem is obtained if we approximate the forward problem for all  $d_{\text{calc}}$  by using only the first two terms of the Taylor series. The resulting vector-matrix equation may be solved using linear algebraic methods. This process is called *linearized inversion*.

Usually, this linearized inversion is iterated: the prior estimate  $\mathbf{m}_0$  is set equal to the newly-found, best-fit location, the problem is re-linearized around this new estimate using Eq. (10), and the new linear problem solved again. This method may be repeated (iterated) many times, as needed to attain some convergence criteria.

Linearized methods produce a single, best-fit (e. g., maximum likelihood) hypocenter and origin time location, and associated, linearly-estimated uncertainties, such as a multi-dimensional, normal-distribution centered on the best-fit hypocenter and origin time. However, this linearized solution is often a poor representation of the complete solution *pdf* (Fig. 2 and see examples), and it may be unstable when the *pdf* is irregular or has multiple peaks due to insufficient or outlier data, velocity model complexities, and other causes (e. g., [5,34]).

### Direct-Search Location Methods

The earliest, formal earthquake locations from phase arrival time observations used nonlinearized procedures. Milne [37] describes and applies several graphical and algebraic methods to determine earthquake locations. These



Earthquake Location, Direct, Global-Search Methods, Figure 2 Schematic diagram comparing linearized and direct-search locations for the case where the complete location *pdf* is moderately complicated, with two maxima. This example arises from the case of a location at the limits of the recording network and near a sharp, horizontal interface in the velocity model between lower velocities above and higher velocities below. The colored, contoured form shows the true location *pdf*, as should be determined by a complete, probabilistic, direct-search location procedure. A linearized location that iterates from an initial trial location below the sharp interface will find an optimal hypocenter near the secondary, local maximum of the location *pdf*, below the interface. The linearized error ellipsoid, based on the curvature of the misfit function at this optimal hypocenter, reflects the form of this secondary maximum only. The linearized location procedure never identifies or explores the primary maximum of the *pdf* above the sharp interface, and produces incorrect error information above this interface (i. e. the uppermost part of the error ellipsoid). A probabilistic, direct, global-search procedure can determine the complete location *pdf* and identify correctly the maximum likelihood hypocenter located above the sharp interface

include a perpendicular bisector method for the case of 3 or more simultaneous arrival time observations (related to the modern arrival order or bisector method), a method of hyperbolae based on the differences in arrival times at pairs of stations (related to the modern EDT method) and a method using the differences in arrival times of different wave types at individual stations. The latter is a generalization of the method of circles using *S-P* times, in which the distance from a station to the source is, for given *P* and *S* velocity models, a function of the difference of the *S* and *P* arrival times; an epicenter can be constrained with such *S-P* based distances from 3 stations. Reid [52] determined a hypocenter location for the great 1906 California earthquake through a coarse, systematic grid search over velocity, position along the causative fault and depth, solving for the origin time and wave velocity by least-squares at each grid point.

The arrival order or bisector method [2,42] is a non-linear, geometrical approach that uses the constraint that if a phase arrival is earlier at station A than at station B, then the event is closer to A than to B (assuming the velocity model is such that arrival order implies distance order). Applying this constraint to all pairs of stations defines a convex region containing the event location. This method is useful for obtaining some constraint on the location of events far outside of an observing station network, and for rapidly and robustly obtaining starting locations for linearized methods.

Most other modern, direct-search earthquake location methods (excluding graphical methods that are now mainly used for illustrative and educational purposes) are based on deterministic or stochastic searches which may be exhaustive or directed and evolutionary. These searches are used to explore or map likelihood functions such as those given in Eqs. (5), (7) or (8). When these searches gather and retain information globally, throughout the prior *pdf*  $p(\mathbf{x})$ , they can produce a complete, probabilistic location *pdf*. Otherwise, searches may determine a global or local maximum of the location *pdf*, or may explore the neighborhood around these optimal points to locally estimate the *pdf* and obtain uncertainty information.

**Regular, Deterministic Search** Regular and deterministic searches, such as grid-searches, nested grid-searches and stochastic, “crude” Monte-Carlo searches (e. g., [20,62]) use global and well-distributed sampling of the model space and thus can estimate the complete location *pdf*. All of these approaches are computationally demanding for problems with many unknown parameters, large parameter spaces, or time consuming forward calculations, because the number of models that must be tested can be very large. These methods have been successfully applied to the determination of optimal hypocenters (i. e., [14,26,61,69]), and to probabilistic location (i. e., [6,34,38,83]), but their inefficiency may impose unacceptable limitations on the number of events that can be considered, or on the size of the search volume.

**Directed Search** Directed, stochastic search techniques include evolutionary, adaptive global search methods such as the genetic algorithm [19,59] and simulated annealing [28,53,72]. The simplex method is a directed, deterministic search technique that is nonlinearized and can be used for earthquake location (e. g., [48]). Most of these methods were developed for optimization or the identification of some very good solutions, which is equivalent to identifying a global or local maximum of the location *pdf*. In general, these methods do not explore the prior

*pdf*  $p(\mathbf{x})$  in a manner that can produce complete, probabilistic solutions to inverse problems. For example, the genetic algorithm performs global searching and may be one of the most efficient stochastic methods for optimization, but it does not use well distributed sampling (the sampling tends to converge rapidly to the region of a locally optimum solution). Similarly, in the simulated annealing, random-walk method the interaction of its variable “temperature” parameter and step size with the local structure of the misfit function can lead to convergence and stalling near a locally optimum solution, and a sample distribution that is neither well nor globally distributed. Both the genetic algorithm and simulated annealing can be tuned to sample more broadly and in the limit become crude Monte Carlo searches, but this removes the main advantage of these methods – that of rapid stochastic optimization.

Though not directly applicable to complete, probabilistic location, directed search algorithms are useful for direct-search, earthquake hypocenter estimation because of their efficiency (e. g., [4,48,60,61]).

**Importance sampling** The efficiency of a Monte Carlo algorithm used to estimate properties of a target (misfit or likelihood) function can be increased by choosing a sampling density which follows the target function as closely as possible [20,30,45]. Techniques that follow this rule are referred to as importance sampling methods, and were originally developed in physics for fast and accurate numerical integration of multi-dimensional functions. The target function is unknown, however, and consequently the optimum importance sampling distribution cannot be determined a priori. Instead, improved efficiency is attained by adjusting (or adapting or evolving) the sampling by using information gained from previous samples so that the sampling density tends towards the target function as the search progresses [30,40,45,65]. For example, importance sampling to determine an earthquake location *pdf* or likelihood function (e. g., Eqs. (5), (7) or (8)), can be obtained by beginning with a sampling that follows the prior *pdf*,  $p(\mathbf{m})$ , and then adjusting the sampling as the search progresses so that the sampling density approaches the posterior, location *pdf*.

Importance sampling techniques that can be used to find complete, probabilistic solutions to inverse problem include the VEGAS algorithm [30], the Metropolis algorithm [40], the neighborhood algorithm [55] and, for three-dimensional problems, oct-tree [33]. Other importance sampling methods are discussed in Hammersley and Handscomb [20] and in Press et al. [45] in the context of numerical integration.

The VEGAS algorithm [30,45] performs importance sampling by accumulating appropriate sampling distributions independently for each parameter as the sampling proceeds. This method can give very good estimates of an individual or a joint marginal *pdf*, but it loses efficiency if the target function includes strong correlation between parameters or if it is independent of some parameters [45]. In addition, the VEGAS algorithm may be difficult or impossible to implement with prior information, such as smoothness constraints, that introduces correlation between parameters. Consequently, this algorithm may not be appropriate for some geophysical problems, including earthquake location, when the location parameters are often correlated or poorly resolved.

The Metropolis or Metropolis-Hastings algorithm (e. g., [40]) is similar to simulated annealing but with a constant temperature parameter. The Metropolis algorithm performs a random walk in the model space, testing at each step nearby trial samples which are accepted or rejected after evaluation of the forward problem according to a likelihood  $L(\mathbf{m})$ . In [40] it is shown that this algorithm samples from the posterior *pdf* of the problem and is therefore an importance sampling method. They show that, in the limit of a very large number of trials, it will not become permanently “trapped” near local maxima and consequently will produce global sampling. Also, because it is a random walk technique, the Metropolis algorithm can perform well even if the volume of the significant regions of the posterior *pdf* is small relative to the volume of the prior *pdf*. However in practical application, with a finite number of samples, this algorithm can become trapped in strong local maxima of the posterior *pdf* if this function is complicated. The Metropolis algorithm has been applied to earthquake location in 3D structures [34,35].

Another recently developed importance sampling technique used in geophysics is the neighborhood algorithm [55,56,57], applicable to high dimensional model spaces. Given an existing set of samples of the objective function, the neighborhood algorithm forms a conditional *pdf* using an approximate Voronoi cell partition of the space around each sample. The algorithm generates new samples through a uniform random walk within the Voronoi cells of the best fitting models determined so far. This algorithm is applied to the 4D hypocenter location problem in [27,58].

The oct-tree importance-sampling method [33] uses recursive subdivision and sampling of rectangular cells in three-dimensional space to generate a cascade structure of sampled cells, such that the spatial density of sampled cells follows the target *pdf* values. The relative probability

that an earthquake location lies within any given cell  $i$  is approximately,

$$P_i = V_i L(\mathbf{x}_i) , \quad (11)$$

where  $V_i$  is the cell volume and  $\mathbf{x}_i$  is the vector of coordinates of the cell center. Oct-tree importance-sampling is used to determine a location *pdf* by first taking a set of samples on a coarse, regular grid of cells throughout the search volume. This is followed by a recursive process which takes the cell  $k$  that has the highest probability  $P_k$  of containing the event location, and subdividing this cell into 8 child cells (hence the name oct-tree), from which 8 new samples of the *pdf* are obtained. These samples are added to a list of all previous samples, from which the highest probability cell is again identified according to Eq. (11). This recursive process is continued until a predetermined number of samples are obtained, or until another termination criterion is reached.

For most location problems, including those with a complicated location *pdf*, the oct-tree recursive subdivision procedure converges rapidly and robustly, producing an oct-tree structure of cells specifying location *pdf* values in 3D space. This oct-tree structure will have a larger number of smaller cells in areas of higher probability (lower misfit) relative to areas of lower *pdf* value and thus the oct-tree method produces approximate importance-sampling without the need for complex geometrical constructs such as Voronoi cells. Oct-tree sampling can be used with the L2-norm likelihood function in Eq. (7) or the EDT likelihood function in Eq. (8), since both require searching over three-dimensional spatial locations only. Oct-tree sampling has been applied to earthquake location in 3D structures [22,23,31,32]; we use this sampling method to determine locations in the examples presented below. Though limited to determination of the 3D, spatial location, this recursive sampling procedure can be extended to 4D to allow determination of the origin time.

### Illustrative Examples

We illustrate the concepts described in the previous sections using an M3.3 earthquake that occurred in the Garfagnana area of Northern Tuscany, Italy, on March 5, 2007 at 20:16 GMT. The earthquake was recorded by stations of the Italian National Seismic Network (INSN) at distances from less than 10 km to more than 300 km. We use manually picked  $P$  and  $S$  phase arrival times from the INSN bulletin with Gaussian uncertainties (standard deviations from 0.01 to 0.1 s), and a 1-D velocity model similar to the standard model used by INSN for routine

earthquake location in Italy. We perform all event locations with the probabilistic location program NonLin-Loc [31,34,35] (<http://www.alomax.net/nlloc>; NLL hereafter), using the oct-tree sampling algorithm (Sect. “**Location Methods**”) to perform a global-search within a parameter space  $\mathcal{M}$  formed by a rectangular volume 360 km on each side and from the Earth’s surface to 35 km depth (except as noted in figure captions). We use the L2-norm (Eq. (7)) or EDT (Eq. (8)) likelihood functions to obtain location *pdf*’s in 3D space and corresponding maximum likelihood origin times.

In order to describe the location problem and the solution quality for each of the examples presented below we focus on geometrical properties of the location *pdf*, which represents most completely the results of probabilistic, direct, global-search methodologies. We also consider the maximum likelihood hypocenter, defined as the point in space of the maximum value of the location *pdf*, and the corresponding origin time. We examine statistics of the quality of the solutions using the half-lengths of three principal axes of a 68% confidence error ellipsoid approximation to the location *pdf*,  $l_{\text{ell}}$ , the weighted, root-mean-square of the arrival residual (observed – calculated) times, *rms*, and a relative measure of the volume of the high likelihood region of the location *pdf*,  $V_{\text{pdf}}$ , given by,

$$V_{\text{pdf}} = \int_{\mathcal{M}} \frac{\text{pdf}(\mathbf{x})}{\text{pdf}^{\text{max}}} dV, \quad (12)$$

where  $\text{pdf}^{\text{max}}$  is the maximum value of the location *pdf* in  $\mathcal{M}$ . We also make use of standard measures of the experimental design quality (i.e, stations coverage) including the *gap* – the largest angle between the epicenter and two azimuthally adjacent stations used for location, and the distance  $\Delta_0$  from the hypocenter to the closest station. These indicators are summarized in Table 1 for the examples presented here.

These examples are meant to show important features and complexity in earthquake location results, not to compare different direct-search location methods or to compare direct-search to linearized algorithms. However, because linearized earthquake location has been and remains an important and widely used tool, we indicate for each example the location results obtained with a linearized algorithm, Hypoellipse [29]. Hypoellipse uses a least-squares, L2-norm and produces a 68% confidence ellipsoid for the hypocenter location. For well constrained locations this ellipsoid should closely match the *pdf* of our probabilistic, L2-norm locations; we plot the Hypoellipse ellipsoid for cases where it differs notably from the probabilistic location, L2-norm *pdf*.

### Example 1: An Ideal Location

To construct an ideal, reference location and synthetic data set for the 2007 Italian earthquake we first locate the event using the earliest 20 observed *P* or *S* arrival times (Fig. 3; Table 1; Example 1a). Next, we subtract the arrival residuals for this location from the corresponding times for all observations and relocate the event with the earliest 50 of these “corrected” times (Fig. 3; Table 1; Example 1b). This procedure results in an ideal, synthetic data set and a location problem that are equivalent to the case of no “a posteriori” picking error and no travel-time error (i. e., no velocity model error). For this problem the quality of the solution and the shape of the resulting location *pdf* reflect primarily the station geometry and corresponding ray take-off angles about the source.

The reference location (Fig. 3; Table 1; Example 1b) has  $\text{rms} = 0$  s,  $\text{gap} = 63^\circ$ ,  $\Delta_0 \sim 9$  km,  $V_{\text{pdf}} \sim 7.0 \text{ km}^3$  and  $l_{\text{ell}} = 1.05, 1.32$  and  $2.05$  km. The *rms* is necessarily zero because we used residuals as time corrections, while the other indicators and the near-ellipsoidal form of the location *pdf* show a well constrained location. The location is well constrained by the data because stations are available at a wide range of distances and azimuths. In particular, the presence of a station nearly above the event, and of both *P* and *S*-wave arrival times for the closer stations, give good depth constraint.

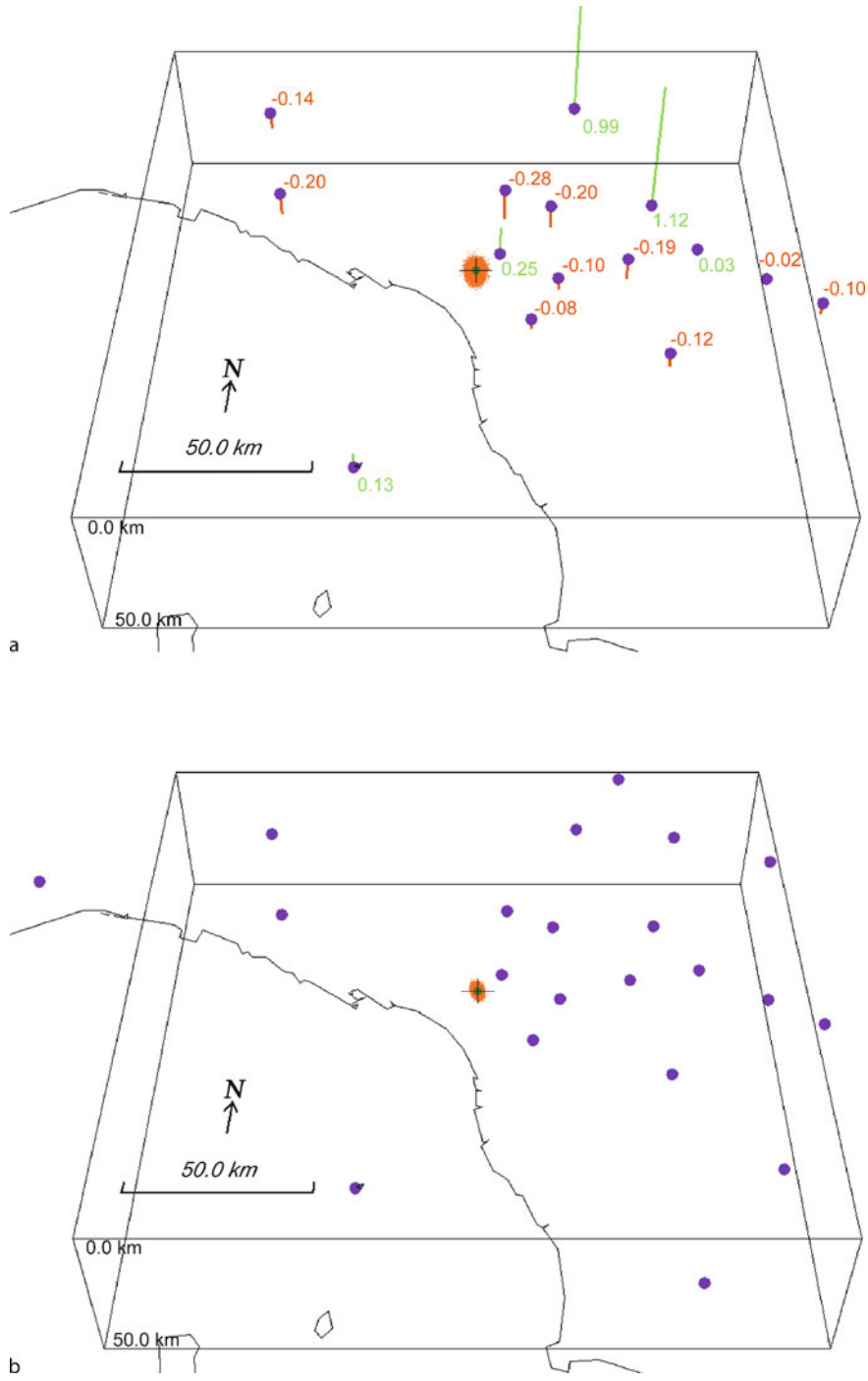
### Examples 2–5: Station Distribution

In the next examples we show locations for three cases with poor station distribution about the source:

- 1) few available stations;
- 2) stations all to one side of the event; and
- 3) no data for stations near or above the source. In addition we illustrate the application of experimental design techniques to improve the station distribution.

**Example 2: Few Available Stations** We first examine relocations of the 2007 Italian earthquake obtained with different numbers of *P* and *S* arrival times selected from the ideal, synthetic data set (Fig. 4; Table 1; Example 2a–d). With only two stations and 2 arrivals (2 *P* phases) the location *pdf* is a fat, near-vertical, planar surface with an elongated, boomerang shape trending perpendicular to the line connecting the two receivers (Fig. 4a). With the addition of *S* arrivals from the same stations (4 arrivals – 2 *P* and 2 *S* phases) the location *pdf* is greatly reduced in volume, and has the form of an annulus oriented roughly perpendicular to the line connecting the two receivers (Fig. 4b). The annular form of this *pdf* results from the intersection





Earthquake Location, Direct, Global-Search Methods, Figure 3

**Example 1: An ideal location.** a Location obtained using the first 20, observed *P* or *S* arrival times; b location obtained using the first 50, *P* or *S* corrected arrival times from the ideal, synthetic data set. The elements shown in these and the following figures are: stations used for location (blue dots, in some cases stations fall outside the plotted region); location pdf (red cloud of points showing an importance sample drawn from the pdf); maximum likelihood hypocenter (green dot); ideal, synthetic location (black cross); *P* arrival residuals at each station: positive (green, up-going bars) and negative (red, down-going bars), numbers indicate residual value in sec. The Hypoellipse linearized locations and ellipsoids do not differ significantly from the direct-search locations shown in this figure

Earthquake Location, Direct, Global-Search Methods, Table 1

Summary of results and quality indicators for the example locations.  $R_{pdf}$  is the radius of a sphere with volume  $V_{pdf}$ ;  $l_{ell}^1, l_{ell}^2, l_{ell}^3$  are the half-lengths of the error ellipsoid axes;  $N_P$  is the number of phases used for the location;  $\Delta_0$  is the distance to the closest station

		Example	Lat (°)	Lon (°)	Depth (km)	rms (s)	gap (°)	$\Delta_0$ (km)	$N_P$	$V_{pdf}$ (km <sup>3</sup> )	$R_{pdf}$ (km)	$l_{ell}^1$ (km)	$l_{ell}^2$ (km)	$l_{ell}^3$ (km)
	Ideal	1a	44.208	10.295	10.98	0.399	89	9.1	20	26	1.8	1.6	2.2	3.2
		1b	44.208	10.295	10.98	0.000	63	9.1	50	7	1.2	1.1	1.3	2.1
Station distribution	Few stations	2a	44.163	12.267	47.44	0.000	335	232	2	742001	82.3	30.9	53.9	223
		2b	44.172	9.565	77.60	0.001	192	25.8	4	21600	25.3	5.9	43.2	81.9
		2c	44.208	10.295	14.40	0.000	173	64.5	3	2011	7.8	4.8	7.3	20.2
		2d	44.208	10.295	10.98	0.000	99	9.1	8	66	2.5	2.1	2.7	4.9
	Side	3	44.207	10.296	11.03	0.004	251	29.0	19	444	4.7	3.4	4.7	11.0
	Far	4a	44.207	10.296	11.03	0.006	103	106	46	234	3.8	2.0	2.4	17.8
		4b	44.208	10.295	10.98	0.000	103	106	50	66	2.5	1.7	2.2	8.1
	Experimental design	5a	44.215	10.290	15.13	0.014	229	9.9	6	1806	13.4	8.3	13.7	59.1
5b		44.208	10.295	7.83	0.007	89	36.3	6	381	6.6	3.1	3.7	11.5	
Outlier	L2-norm	6a	44.207	10.296	11.03	0.007	135	9.00	10	172	3.5	2.9	3.7	6.7
		6b	44.120	10.267	9.02	0.813	156	10.5	10	172	3.5	3.0	4.2	9.0
	EDT	6c	44.221	10.305	9.94	0.006	132	9.4	10	167	3.4	4.5	7.7	15.2
		6d	44.215	10.304	10.02	0.540	133	9.0	10	275	4.0	10.7	30.8	42.9
Early warning	7a	44.139	10.194	24.79	0.012	307	15.6	3	628090	53.1	18.7	81.6	104	
	7b	44.210	10.302	10.57	0.015	250	8.8	4	33908	20.1	18.4	30.0	105	
	7c	44.208	10.295	11.12	0.008	227	9.1	5	1704	7.4	5.3	13.9	30.2	
	7d	44.207	10.296	11.03	0.007	135	9.0	10	172	3.5	2.9	3.7	6.7	
Incorrect velocity model	L2-norm	8a	44.208	10.295	10.98	0.000	63	9.1	50	15	1.5	1.3	1.7	2.9
		8b	44.160	10.244	2.69	0.808	66	11.4	50	17	1.6	1.3	1.7	3.0
	EDT	8c	44.220	10.305	9.98	0.000	63	9.4	50	11	1.4	1.2	1.5	2.6
		8d	44.192	10.280	7.20	0.795	64	9.3	50	167	3.4	2.8	3.9	6.7

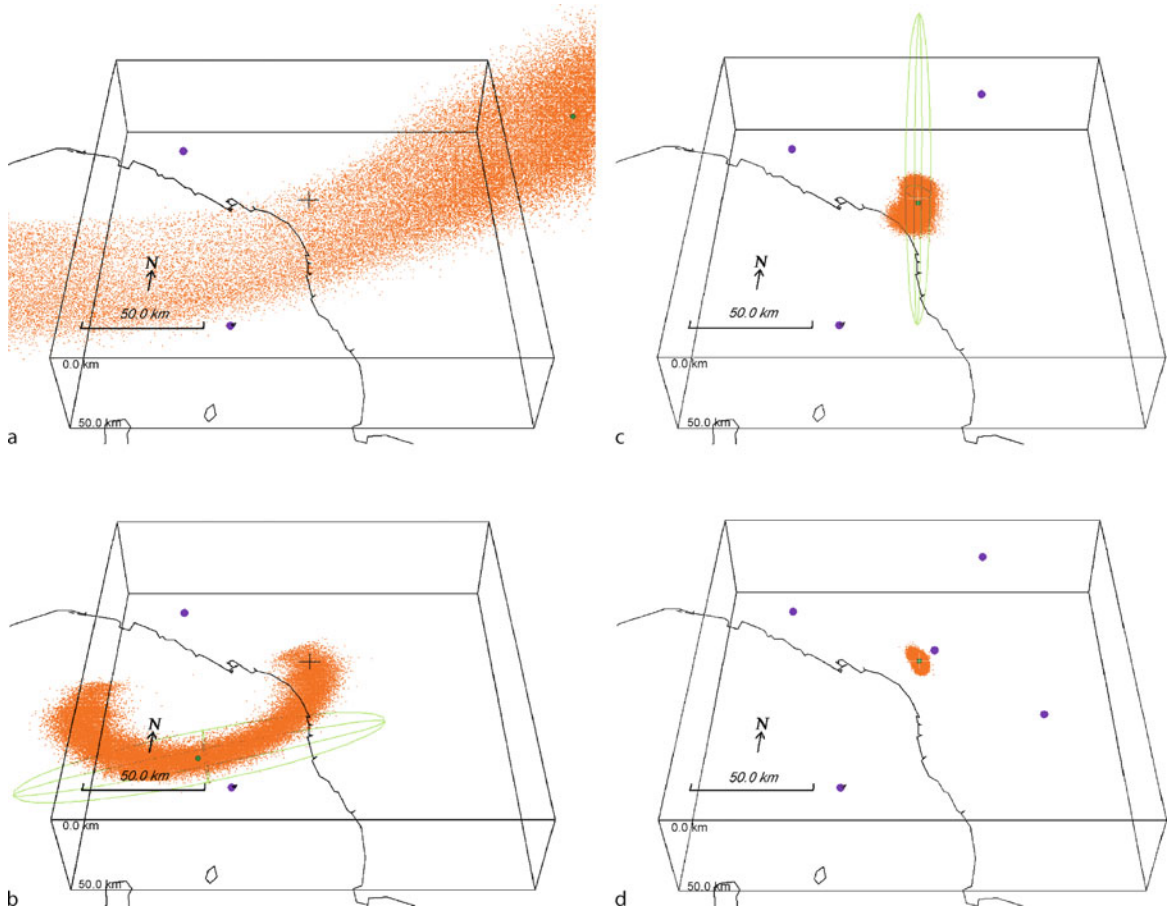
of the boomerang shape *pdf* produced by the 2 *P* phases (Fig. 4a) and two hemispherical *pdf*'s centered on each station. Each of these hemispherical *pdf*'s would be produced by location using only the *P* and the *S* reading from either station; this is the probabilistic analogue to the method of circles using *S-P* times.

With three stations (3 arrivals – 3*P* phases) the location *pdf* forms one mass and its volume is further reduced. This location *pdf* retains an irregular, curved shape resulting from poor constraint of one spatial dimension that trades off with origin time (Fig. 4c). For all of these locations the problem is effectively underdetermined – the data cannot constrain all three hypocentral coordinates and origin time. In these cases a linearized location algorithm may not converge and would be unable to represent properly the effective location uncertainties. As more data are added, the location *pdf* progressively reduces in size and complexity, and with the addition of a station close to and above the source (8 arrivals – 5 *P* and 3 *S* phases), the location *pdf* has a compact, near ellipsoidal form indicating some constraint on all hypocentral coordinates and

origin time (Fig. 4d). This location is similar to that obtained with the complete, ideal data set (Fig. 3b), though the location *pdf* remains much larger than that of the ideal case which has arrival times from many more stations.

**Example 3: Stations to One Side of the Event – Large Gap**

Next, we examine the case of earthquakes occurring outside of the recording network with an example using *P* arrival times from stations only to the southeast of the earthquake (Fig. 5; Table 1; Example 3). The location *pdf* is large and elongated in a northwest-southeast direction oriented towards the centroid of the available stations because the lack of stations to the northwest (and use of *P* times only) allows a strong trade-off between potential hypocenter locations along this direction and origin time. In contrast, there is some constraint of the *pdf* to the northeast and southwest due to the aperture of the available stations. The poor station distribution and potential lack of constraint is clearly indicated by the large gap value for this location, gap = 251°. One or more good quality *S* readings can reduce the elongation of the *pdf*.



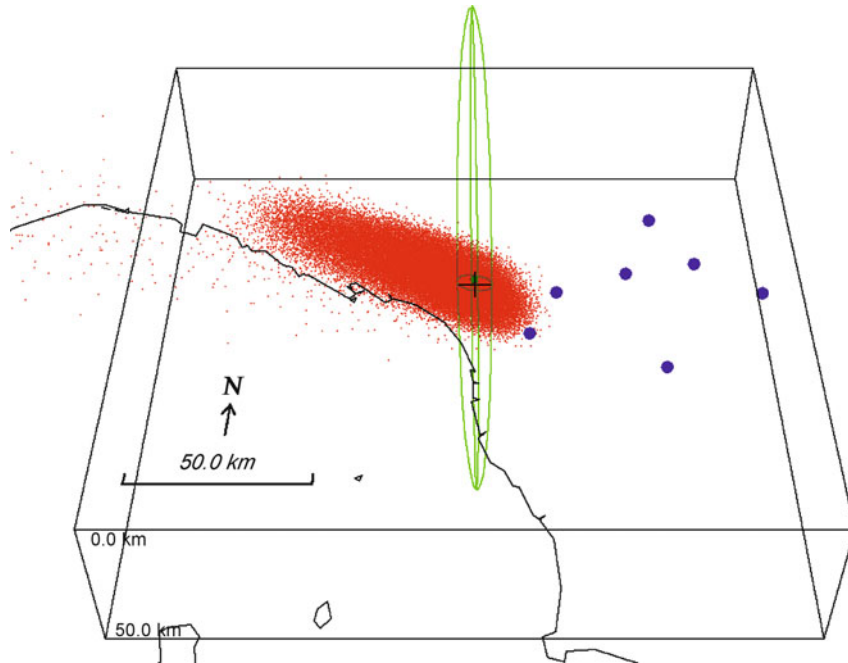
Earthquake Location, Direct, Global-Search Methods, Figure 4

**Example 2: Few available stations.** Locations obtained using progressively (a–d) a larger number of arrival observations. a 2P phases (2 stations); b 2P and 2 S phases (2 stations); c 3P phases (3 stations); d 5P and 3S phases (5 stations). For the locations in a and b the oct-tree search is performed to 100 km depth. In this and the following figures the 68% Hypoellipse ellipsoid is shown with *green lines*. Hypoellipse linearized location: does not converge for the location in panel a; ellipsoid differs markedly from the direct-search location *pdf* in panels b and c; and does not differ markedly from the direct-search location in panel d

**Example 4: Stations Far From the Event – Vertically Elongated PDF** We next show an example where the nearest recording stations are far from the earthquake, relative to its depth, and either *P* arrival times only or both *P* and *S* arrival times are available (Fig. 6; Table 1; Examples 4a–b). With this station geometry the seismic rays leave the source region with approximately the same dip-direction to all stations. Consequently a change in source depth gives about the same change in predicted travel times to all stations. This change in travel time is indistinguishable from a change in origin time (c.f., Eqs. (1) or (2), leading to a strong trade-off between origin time and depth. Consequently the location *pdf* has a vertically elongated shape which, for the case of *P* arrivals only (Fig. 6a),

extends throughout the entire search range in depth indicating no depth constraint. For a linearized location algorithm this location problem can be effectively underdetermined, though most linearized algorithms can fix the hypocenter depth artificially in order to obtain a stable epicentral location. The addition of *S* arrival times (Fig. 6b) improves the depth constraint to some extent, although the location *pdf* remains highly elongated in the vertical direction. The lack of close stations and potential lack of constraint is clearly indicated by the large  $\Delta_0$  value for this location,  $\Delta_0 \approx 106$  km.

This case is common with sparse networks and with shallow sources. Reducing the vertical extent of the *pdf* requires stations at distances of the order of the source depth



Earthquake Location, Direct, Global-Search Methods, Figure 5

**Example 3: Stations to one side of the event.** A location example with  $P$ -wave arrival times at 7 stations only to the southeast of the event. The Hypoellipse ellipsoid differs markedly from the direct-search location  $pdf$  in this figure

or less. The addition of one or more good quality  $S$  readings, especially at the closest stations, would further improve the depth constraint.

#### Example 5: Stations Selection with Experimental Design

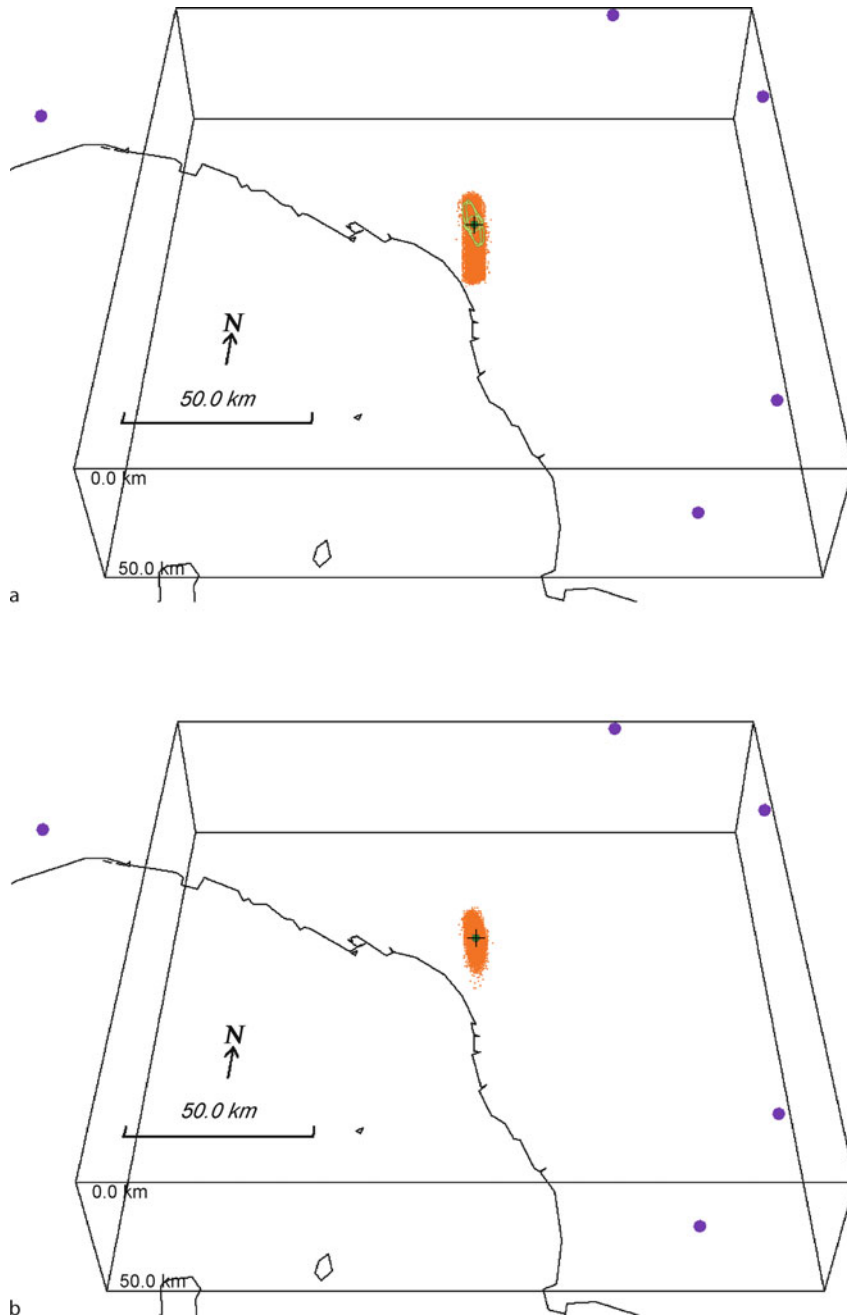
Next, we illustrate the application of experimental design techniques to station selection (Fig. 7; Table 1; Example 5). Considering a case similar to Example 3, which has 6 stations to one side of the event giving poor constraint on the location, we determine an optimal set of 6 stations to best constrain the location. To do this we apply a linearized design method [13] to select an optimal subset of 6 of the available INSN stations to best constrain an event at the (known) location produced by the ideal, synthetic data (Example 1b).

The design procedure does not simply select the 6 closest receivers to the source (i.e. first 6 available arrival times, Fig. 7a), but instead selects receivers distributed around, and to a large distance away from the source (Fig. 7b). This choice can be understood as balancing the distribution of directions (azimuth and inclination) that the rays leave the source to the selected receivers, a direct result of the use of the linearized approximations to the model-data relationship Eq. (10) in the linearized design method [13]. This method is based on selecting sta-

tions based on the similarity between the rows of the location kernel matrix of the linearized problem; the approach does not differ significantly from that of Uhrhammer [76] based on the condition number of the same matrix. The improvement in station distribution in azimuth is indicated by the small gap value for this location,  $\text{gap} = 89^\circ$ . The resulting location  $pdf$  (Fig. 7b) is compact and symmetric relative to the location  $pdf$  obtained from the first 6 stations recording the  $P$  phases (Fig. 7a), and the maximum likelihood hypocenter is close to the ideal location hypocenter.

#### Example 6: Incorrect Picks and Phase Identification – Outlier Data

For a given hypocenter location, an outlier arrival time has a residual that is much greater than its nominal error. Data outliers are common with automatic phase arrival picking algorithms, with  $S$  arrival picks, for small events, distant stations, or other cases where the signal to noise ratio is low, and for early instrumental data where large timing errors are common. In many cases, such as automatic earthquake monitoring and early warning systems, it is important to have robust location procedures that are influenced as little as possible by the presence of outliers. One way to

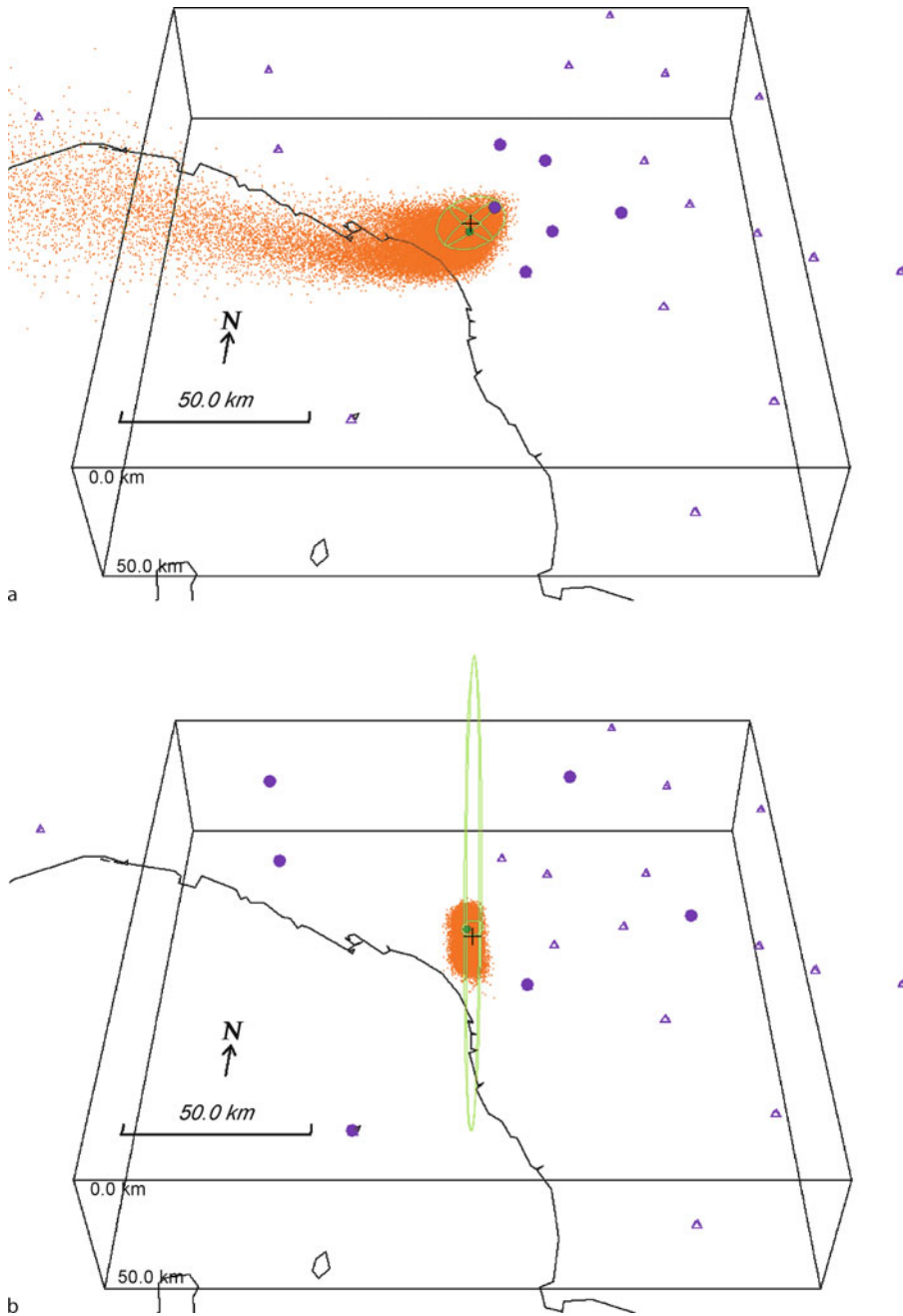


Earthquake Location, Direct, Global-Search Methods, Figure 6

**Example 4: Stations far from the event.** A location example using stations far from the epicenter, with **a**  $P$  arrival times only, **b** both  $P$  and  $S$  arrival times. Hypoellipse linearized location: ellipsoid differs markedly from the direct-search location  $pdf$  in panel **a**; and does not differ markedly from the direct-search location in panel **b**

achieve this is to use robust likelihood functions such as EDT Eq. (8). In the example below, we compare the performance of EDT and the more commonly used L2-norm likelihood functions.

This example uses only stations near the source, and arrival times from ideal, synthetic data sets for both the L2-norm and the EDT likelihood functions. We add 3 s to the  $P$  arrival time at two stations to generate outlier

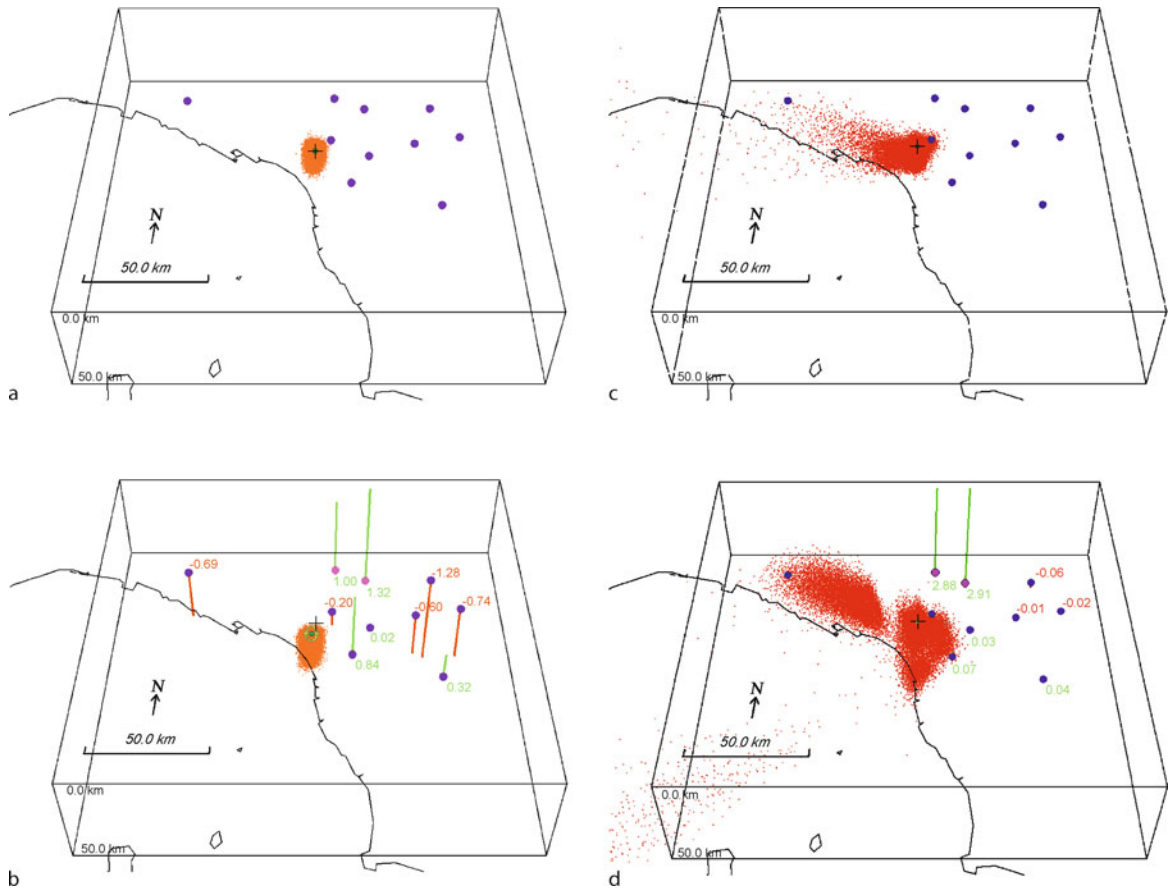


Earthquake Location, Direct, Global-Search Methods, Figure 7

**Example 5: Stations selection with experimental design.** A location example showing a Location using the stations with the first 6 available arrival times, **b** location using an optimal set of 6 stations as determined with a linearized experimental-design method. Available stations not used or selected are shown with *open triangle symbols*. Hypoellipse linearized location: ellipsoids differ markedly from the direct-search location *pdf's* in panels **a** and **b**

data, and examine L2-norm and EDT locations without and with the outlier data (Fig. 8; Table 1; Examples 6a–d). The L2-norm location with the outlier data (Fig. 8b)

does not identify and isolate the two outlier *P*-arrivals but instead mixes information from these arrivals with the other data resulting in relatively large, non-zero residu-



Earthquake Location, Direct, Global-Search Methods, Figure 8

**Example 6: Incorrect picks and phase identification – outlier data.** Locations using ten  $P$ -wave arrival times with L2-norm and a no outliers, **b** two arrival-time outliers, and with EDT and **c** no outliers, **d** two arrival-time outliers. The stations with outlier arrivals are shown with *violet dots*. Note the small  $pdf$  of L2-norm regardless of the outliers and, in contrast, the ability of EDT to detect the outliers (see text). The Hypoellipse ellipsoid differs markedly from the direct-search location  $pdf$  in panel **b**. Hypoellipse not compared to EDT locations in panels **c** and **d**

als for all arrivals. This results in a bias of about 10 km in the maximum likelihood hypocenter location relative to the ideal location hypocenter, while the location  $pdf$  for the L2-norm locations with and without outlier data have about the same size and form, but have little overlap (Figs. 8a and 8b). Thus the L2-norm solution gives no clear indication of the presence of outlier data, or that the solution may be biased. In contrast, the EDT location for the data set containing the outliers (Fig. 8d) correctly identifies the two outlier arrivals (the EDT residuals for these two outlier data are both about 2.9 s) and strongly downweights them (from 1.2 to 0.17 posterior weight), while producing small residuals ( $< 0.08$  s) for the remaining arrival, as would be the case without outlier data. The maximum likelihood hypocenters for the EDT locations with

and without outlier data are almost identical, but the location  $pdf$ 's are very different (Fig. 8c and 8d). With outlier data, the  $pdf$  has an irregular shape and several distinct parts, reflecting the inconsistency of the data set to constrain a unique event location. For the outlier locations, a potential problem with the data set is indicated by the large  $rms$  values with both L2-norm and EDT, and with EDT alone, by the asymmetry in residuals, the irregular  $pdf$  shape, and the large  $V_{pdf}$  and  $I_{ell}$  values.

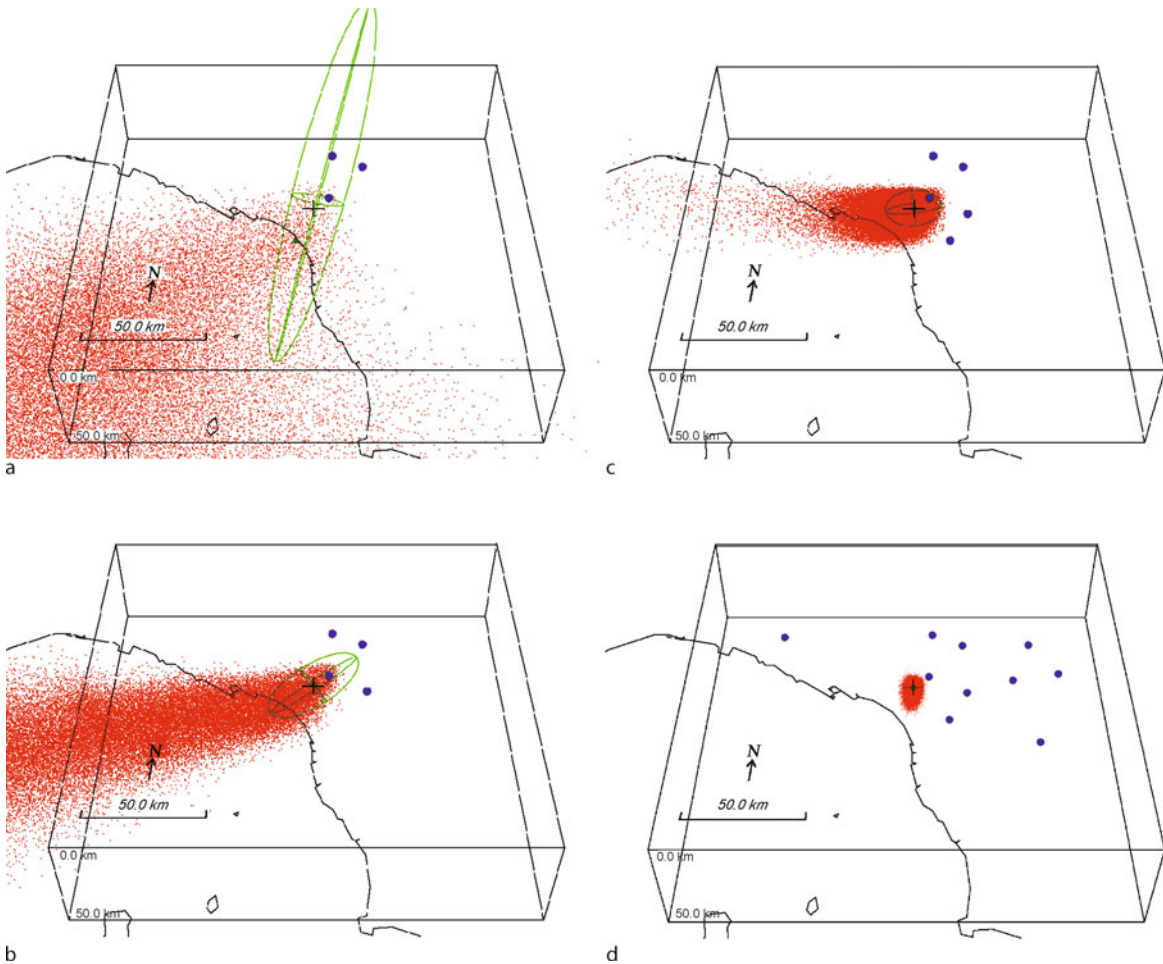
This result shows that location in the presence of outlier data can be remarkably stable with the EDT likelihood function, which is easy to implement with direct-search location techniques. In contrast, the same location with the commonly used, L2-norm likelihood function is biased, while presenting few indicators of this bias.

**Example 7: Earthquake Early-Warning Scenario**

Location for earthquake early-warning must be performed rapidly and in an evolutionary manner starting with the first available phase arrivals. In this example we examine the ability of direct-search location to obtain robust and useful location information using *P* arrivals from the first stations that record the Northern Italian event (Fig. 9; Table 1; Examples 7a–d).

Within about 6 seconds after the origin time,  $t_0$ , three *P* readings are available. Location with these readings produces an extensive location *pdf* that fills the southwest quadrant of the search region (Fig. 9a); this *pdf* does not provide useful constraint on the location, but is robust in that it includes the true location. Progressive addition of

more arrival time data (Fig. 9b and 9c) reduces the size of the location *pdf*. With 5 arrivals, at about 7 s after  $t_0$  (Fig. 9c), the maximum likelihood location is close to that of the ideal, synthetic location and the location *pdf* is well delimited, although elongated towards the west because no arrivals are yet available from stations in that direction. By 13 s after  $t_0$  (Fig. 9d), 10 *P* arrivals are available and the location *pdf* is now compact and symmetrical, primarily because a station to the northwest is included. This *pdf* has small enough  $V_{pdf}$  and  $I_{ell}$  values to provide useful, probabilistic constraint on the location for early-warning purposes at a regional scale, while the maximum likelihood hypocenter is effectively the same as that of the ideal location. In practical application, direct-search location results similar to those illustrated here can be obtained within



Earthquake Location, Direct, Global-Search Methods, Figure 9

**Example 7: Earthquake early-warning scenario.** Progressive location using a 3, b 4, c 5 and d 10 stations. Hypoellipse linearized location: ellipsoid differs markedly from the direct-search location *pdf*'s in panel a, b and c; and does not differ markedly from the direct-search location in panel d



a delay of less than 1 sec after the readings are available (e. g. [64]).

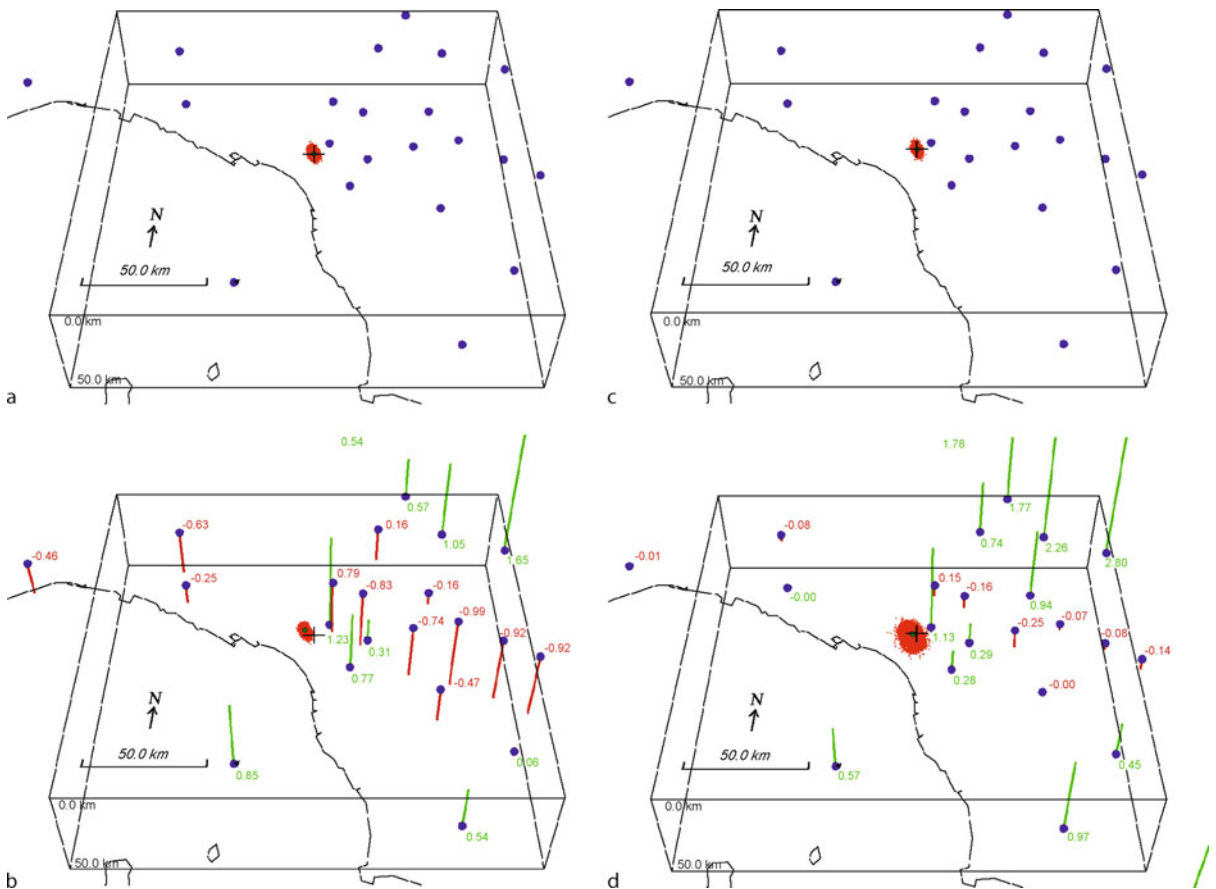
### Example 8: Incorrect Velocity Model

Any velocity model used for earthquake location is an approximation to the true Earth and thus will in general produce erroneous predicted travel times. The magnitude of error in the travel times depends on many factors, but will in general be larger for more distant stations and with increased complexity in the true Earth structure. We examine the effect of an incorrect velocity models by repeating the ideal location (Example 1a and b) with and without the “corrected” times, and using 50  $P$  arrivals (the ideal location was determined using the first 20  $P$  or  $S$  arrivals). We

examine locations using the L2-norm and EDT likelihood functions (Fig. 10; Table 1; Examples 8a–d).

The locations with time corrections (Fig. 10a and 10c) simulate the unrealizable case of perfect knowledge of the velocity structure. With both the L2-norm and EDT the location results show zero residuals, compact location  $pdf$ 's and a maximum likelihood hypocenter that necessarily matches exactly the corresponding ideal location. We note, however, that the L2-norm and EDT “ideal” locations differ slightly because they are derived from noisy, real data, and they use different likelihood functions.

The locations without time corrections (Fig. 10b and 10d) use the true observed data (i. e., travel times through the true Earth) and thus show the effect of an incorrect velocity model (i. e., the 1-D velocity model used



Earthquake Location, Direct, Global-Search Methods, Figure 10

**Example 8: Incorrect velocity model.** Locations using 50  $P$  arrivals with the L2-norm and a time corrections, **b** no time corrections, and with EDT and **c** time corrections, **d** no time corrections. The locations without the ideal time corrections show the effect of an incorrect velocity model. The Hypoellipse linearized locations and ellipsoids do not differ markedly from the direct-search locations shown in this figure

for location). This is shown by the pattern of positive and negative residuals obtained with both the L2-norm and EDT. The L2-norm location without time corrections has a balanced distribution of positive and negative residuals and, relative to the L2-norm location with corrections, a similar size location *pdf* and a biased maximum likelihood hypocenter. In contrast, the EDT location without corrections has more positive than negative residuals and, relative to the EDT location with corrections, a larger location *pdf* and nearly identical, unbiased maximum likelihood hypocenter. For these locations, a potential problem with the velocity model is indicated by the large residuals and *rms* values with both L2-norm and EDT, and, with EDT, by the asymmetry in residuals, the irregular *pdf* shape, and the large  $V_{pdf}$  and  $I_{ell}$  values, as with the outlier data example (Example 6).

In effect, locations with an incorrect velocity model and with outlier data are mathematically similar, though in the former case all or most residuals may be large while in the latter case only a few will be large. It is difficult to distinguish between the two cases with the L2-norm because this algorithm seeks to best satisfy *all* of the observations simultaneously (cf., Eq. (7)) by balancing the distribution of positive and negative residual (cf., Fig. 8b and Fig. 10b). Thus, relative to the residuals corresponding to the correct location, the L2-norm solution damps and hides larger residuals at the expense of increasing small residuals. In contrast, EDT seeks to best satisfy the *most* pairs of observations (cf., Eq. (8)) and imposes no inherent constraint on the distribution of residuals. Thus with EDT the difference in number, magnitude and distribution of large residuals – few and large for the outlier case, many of similar magnitude and spatially correlated for the incorrect velocity model case – allows one, in principle, to distinguish between the two cases (cf., Fig. 8d and Fig. 10d). In addition, the size and complexity of the location *pdf*'s generally increases more rapidly with EDT than with the L2-norm as the solution quality decreases. Thus, with both the outlier and incorrect velocity model cases, the location results with the EDT likelihood function are more informative than with the L2-norm. However, location with the EDT likelihood function can become unstable (e. g. define only a local maximum of the *pdf*) for cases where the outlier data or velocity model errors lead to extreme complexity in the topology of the EDT location *pdf*.

### Future Directions

There are various ways that direct, global-search location methodologies may evolve in the future. For example, the stability and completeness of the location and location *pdf*

could be improved with the use of more complete data uncertainties, expressed as a *pdf*. These *pdf*'s may typically be irregular and asymmetric, and difficult to determine and parametrize. Currently, enumerated quality indications or, at best, simple normal distributions (describing Gaussian uncertainty) are used to describe the picking error.

Similarly, we have shown that earthquake location depends inherently on the velocity model adopted, but that no realistic uncertainties are associated with this model. Differences between the velocity model and the true Earth can result in complicated differences in ray-paths and travel-times, which will depend strongly on the source and receiver positions. These complications, combined with the lack of knowledge about the true Earth, makes estimating true travel time uncertainties effectively impossible. However, it can be assumed that changes become progressively larger with increasing ray-path length. This effect could be accounted for approximately by travel-time uncertainties that increase with the ray-length or travel time. Instead of using a velocity model to generate travel times, another approach is to derive the required times from tables of empirically determined or corrected travel times (e. g. [41,43]). With this approach the travel-time uncertainties are estimated from timing information, with little or no direct use of velocity structures or ray paths.

We have described and illustrated the importance of the source-receiver geometry for locating earthquakes, notably with regards to constraining a compact and symmetric location *pdf*. Thus, improved constraint on event locations can be achieved through prior use of survey design techniques to select station sites. In a related manner, after an event occurs, these techniques could be employed dynamically to weight the available arrival times used for location with respect to the geometry of the available stations around the likely source region.

The demand for rapid, real-time location and earthquake early warning requires improvements in the integration, speed, quality and robustness of the phase arrival picking, phase association and event location procedures. Currently, development is progressing on integrated procedures which are evolutionary and probabilistic, using, for example, robust likelihood functions such as EDT and information from not-yet-triggered stations (e. g., [8,21,54,63,64]).

A current problem in direct-search location is how to describe in a standardized and compact way the sometimes topologically-complex location *pdf*. For example, such a description is needed if the *pdf* is to be included in standard earthquake catalogs and for rapid dissemination

of probabilistic location information for earthquake early-warning. More generally, making full use of the extensive information in direct-search location solutions will require new methods and procedures to store, distribute and analyze the location *pdf*, maximum likelihood hypocenter, arrival residuals and weights, and other statistics and quality indicators of the solutions.

The continuing increase in computer speed will allow application of direct-search inversion methods to relative location of ensembles of events and for joint epicentral determination in the near future. The use of these methods will be important to explore more completely the vast solution space and better determine the error and resolution for such high-dimensional inverse problems.

The continuing increase in computer speed will also make practical earthquake location techniques using waveform recordings directly, without the intermediate stage of extracting phase arrival times. In these techniques, continuous waveform data streams are matched to synthetic Green's functions within a global-search over possible source locations and source parameters. This type of approach is used to locate previously unidentified earthquakes using low amplitude surface waves on off-line, continuous, broadband waveforms [15,68], and for automatic, real-time estimation of moment tensors and location from continuous broadband data streams (e.g., [25]). Waveform methods will likely be applied to earthquake location on local and regional scales as faster computers and more accurate 3D velocity models become available [81]; related applications using simple ray or acoustic theories to generate the Green's functions show promising results (e.g., [3]).

## Bibliography

### Primary Literature

- Aki K, Richards PG (1980) Quantitative Seismology. Freeman, New York
- Anderson K (1981) Epicentral location using arrival time order. *Bull Seism Soc Am* 71:541–545
- Baker T, Granat R, Clayton RW (2005) Real-time Earthquake Location Using Kirchhoff Reconstruction. *Bull Seism Soc Am* 95:699–707
- Billings SD (1994) Simulated annealing for earthquake location. *Geophys J Int* 118:680–692
- Buland R (1976) The mechanics of locating earthquakes. *Bull Seism Soc Am* 66:173–187
- Calvert A, Gomez F, Seber D, Barazangi M, Jabour N, Ibrahima A, Demnati A (1997) An integrated geophysical investigation of recent seismicity in the Al-Hoceima region of North Morocco. *Bull Seism Soc Am* 87:637–651
- Červený V (2001) *Seismic Ray Theory*. Cambridge University Press, Cambridge
- Cua G, Heaton T (2007) The Virtual Seismologist (VS) Method: a Bayesian Approach to Earthquake Early Warning. In: Gasparini P, Gaetano M, Jochen Z (eds) *Earthquake Early Warning Systems*. Springer, Berlin
- Curtis A (1999) Optimal experiment design: Cross-borehole tomographic examples. *Geophys J Int* 136:637–650
- Curtis A (1999) Optimal design of focussed experiments and surveys. *Geophys J Int* 139:205–215
- Curtis A (2004) Theory of model-based geophysical survey and experimental design Part A – Linear Problems. *Lead Edge* 23(10):997–1004
- Curtis A (2004) Theory of model-based geophysical survey and experimental design Part B – Nonlinear Problems. *Lead Edge* 23(10):1112–1117
- Curtis A, Michelini A, Leslie D, Lomax A (2004) A deterministic algorithm for experimental design applied to tomographic and microseismic monitoring surveys. *Geophys J Int* 157:595–606
- Dreger D, Uhrhammer R, Pasyanos M, Frank J, Romanowicz B (1998) Regional and far-regional earthquake locations and source parameters using sparse broadband networks: A test on the Ridgecrest sequence. *Bull Seism Soc Am* 88:1353–1362
- Ekström G (2006) Global Detection and Location of Seismic Sources by Using Surface Waves. *Bull Seism Soc Am* 96:1201–1212. doi:10.1785/0120050175
- Font Y, Kao H, Lallemand S, Liu CS, Chiao LY (2004) Hypocentral determination offshore Eastern Taiwan using the Maximum Intersection method. *Geophys J Int* 158:655–675
- Geiger L (1912) Probability method for the determination of earthquake epicenters from the arrival time only (translated from Geiger's 1910 German article). *Bull St Louis Univ* 8:56–71
- Gentili S, Michelini A (2006) Automatic picking of P and S phases using a neural tree. *J Seism* 10:39–63. doi:10.1007/s10950-006-2296-6
- Goldberg DE (1989) *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, Reading
- Hammersley JM, Handscomb DC (1967) *Monte Carlo Methods*. Methuen, London
- Horiuchi S, Negishi H, Abe K, Kamimura A, Fujinawa Y (2005) An Automatic Processing System for Broadcasting Earthquake Alarms. *Bull Seism Soc Am* 95:708–718
- Husen S, Smith RB (2004) Probabilistic Earthquake Relocation in Three-Dimensional Velocity Models for the Yellowstone National Park Region, Wyoming. *Bull Seism Soc Am* 94:880–896
- Husen S, Kissling E, Deichmann N, Wiemer S, Giardini D, Baer M (2003) Probabilistic earthquake location in complex three-dimensional velocity models: Application to Switzerland. *J Geophys Res* 108:2077–2102
- Johnson CE, Lindh A, Hirshorn B (1994) Robust regional phase association. *US Geol Surv Open-File Rep* pp 94–621
- Kawakatsu H (1998) On the real-time monitoring of the long-period seismic wavefield. *Bull Earthq Res Inst* 73:267–274
- Kennett BLN (1992) Locating oceanic earthquakes – the influence of regional models and location criteria. *Geophys J Int* 108:848–854
- Kennett BLN (2006) Non-linear methods for event location in a global context. *Phys Earth Planet Inter* 158:46–54
- Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. *Science* 220:671–680
- Lahr JC (1999) HYPOELLIPSE: A Computer Program for Determining Local Earthquake Hypocentral Parameters,

- Magnitude, and First-Motion Pattern (Y2K Compliant Version) 1999 Version 1.0. US Geological Survey Open-File Report, pp 99–23. [http://jclahr.com/science/hypoellipse/hypoel/hypoman/hypomst\\_pdf.pdf](http://jclahr.com/science/hypoellipse/hypoel/hypoman/hypomst_pdf.pdf)
30. Lepage GP (1978) A new algorithm for adaptive multi-dimensional integration. *J Comput Phys* 27:192–203
  31. Lomax A (2005) A Reanalysis of the Hypocentral Location and Related Observations for the Great 1906 California Earthquake. *Bull Seism Soc Am* 91:861–877
  32. Lomax A (2008) Location of the Focus and Tectonics of the Focal Region of the California Earthquake of 18 April 1906. *Bull Seism Soc Am* 98:846–860
  33. Lomax A, Curtis A (2001) Fast, probabilistic earthquake location in 3D models using oct-tree importance sampling. *Geophys Res Abstr* 3:955. [www.alomax.net/nlloc/octtree](http://www.alomax.net/nlloc/octtree)
  34. Lomax A, Virieux J, Volant P, Berge C (2000) Probabilistic earthquake location in 3D and layered models: Introduction of a Metropolis-Gibbs method and comparison with linear locations. In: Thurber CH, Rabinowitz N (eds) *Advances in Seismic Event Location*. Kluwer, Amsterdam
  35. Lomax A, Zollo A, Capuano P, Virieux J (2001) Precise, absolute earthquake location under Somma-Vesuvius volcano using a new 3D velocity model. *Geophys J Int* 146:313–331
  36. Maurer HR, Boerner DE (1998) Optimized and robust experimental design. *Geophys J Int* 132:458–468
  37. Milne J (1886) *Earthquakes and Other Earth Movements*. Appleton, New York
  38. Moser TJ, van Eck T, Nolet G (1992) Hypocenter determination in strongly heterogeneous earth models using the shortest path method. *J Geophys Res* 97:6563–6572
  39. Moser TJ, Nolet G, Snieder R (1992) Ray bending revisited. *Bull Seism Soc Am* 82:259–288
  40. Mosegaard K, Tarantola A (1995) Monte Carlo sampling of solutions to inverse problems. *J Geophys Res* 100:12431–12447
  41. Myers SC, Schultz CA (2000) Improving Sparse Network Seismic Location with Bayesian Kriging and Teleseismically Constrained Calibration Events. *Bull Seism Soc Am* 90:199–211
  42. Nicholson T, Gudmundsson Ó, Sambridge M (2004) Constraints on earthquake epicentres independent of seismic velocity models. *Geophys J Int* 156:648–654
  43. Nicholson T, Sambridge M, Gudmundsson Ó (2004) Three-dimensional empirical traveltimes: construction and applications. *Geophys J Int* 156:307–328
  44. Podvin P, Lecomte I (1991) Finite difference computations of traveltimes in very contrasted velocity models: a massively parallel approach and its associated tools. *Geophys J Int* 105:271–284
  45. Press WH, Flannery BP, Saul AT, Vetterling WT (1992) *Numerical Recipes*, 2nd edn. Cambridge Univ Press, New York
  46. Presti D, Troise C, De Natale G (2004) Probabilistic Location of Seismic Sequences in Heterogeneous Media. *Bull Seism Soc Am* 94:2239–2253
  47. Pujol J (2000) Joint event location – The JHD technique and applications to data from local seismic networks. In: Thurber CH, Rabinowitz N (eds) *Advances in Seismic Event Location*. Kluwer, Amsterdam
  48. Rabinowitz N (2000) Hypocenter location using a constrained nonlinear simplex minimization method. In: Thurber CH, Rabinowitz N (eds) *Advances in Seismic Event Location*. Kluwer, Amsterdam
  49. Rabinowitz N, Steinberg DM (2000) A statistical outlook on the problem of seismic network configuration. In: Thurber CH, Rabinowitz N (eds) *Advances in Seismic Event Location*. Kluwer, Amsterdam
  50. Rawlinson N, Sambridge M (2004) Wave front evolution in strongly heterogeneous layered media using the fast marching method. *Geophys J Int* 156:631–647
  51. Rawlinson N, Sambridge M (2004) Multiple reflection and transmission phases in complex layered media using a multi-stage fast marching method. *Geophys* 69:1338–1350
  52. Reid HF (1910) *The Mechanics of the Earthquake*. Vol II of: *The California Earthquake of 18 April 1906*. Report of the State Earthquake Investigation Commission, Lawson AC (Chairman). Carnegie Institution of Washington Publication, vol 87 (reprinted 1969)
  53. Rothman DH (1985) Nonlinear inversion, statistical mechanics, and residual statics estimation. *Geophysics* 50:2784–2796
  54. Rydelek P, Pujol J (2004) Real-Time Seismic Warning with a Two-Station Subarray. *Bull Seism Soc Am* 94:1546–1550
  55. Sambridge M (1998) Exploring multi-dimensional landscapes without a map. *Inverse Probl* 14:427–440
  56. Sambridge M (1999) Geophysical inversion with a Neighbourhood algorithm, vol I. Searching a parameter space. *Geophys J Int* 138:479–494
  57. Sambridge M (1999) Geophysical inversion with a neighbourhood algorithm, vol II. Appraising the ensemble. *Geophys J Int* 138:727–746
  58. Sambridge M (2003) Nonlinear inversion by direct search using the neighbourhood algorithm. In: *International Handbook of Earthquake and Engineering Seismology*, vol 81B. Academic Press, Amsterdam, pp 1635–1637
  59. Sambridge M, Drijkoningen G (1992) Genetic algorithms in seismic waveform inversion. *Geophys J Int* 109:323–342
  60. Sambridge M, Gallagher K (1993) Earthquake hypocenter location using genetic algorithms. *Bull Seism Soc Am* 83:1467–1491
  61. Sambridge M, Kennett BLN (1986) A novel method of hypocentre location. *Geophys J R Astron Soc* 87:679–697
  62. Sambridge M, Mosegaard K (2002) Monte Carlo Methods In Geophysical Inverse Problems. *Rev Geophys* 40:1009–1038
  63. Satriano C, Lomax A, Zollo A (2007) Optimal, Real-time Earthquake Location for Early Warning. In: Gasparini P, Gaetano M, Jochen Z (eds) *Earthquake Early Warning Systems*. Springer, Berlin
  64. Satriano C, Lomax A, Zollo A (2007) Real-time evolutionary earthquake location for seismic early warning. *Bull Seism Soc Am* 98:1482–1494
  65. Sen M, Stoffa PL (1995) *Global optimization methods in geophysical inversion*. Elsevier, Amsterdam, p 281
  66. Sethian JA (1999) *Level set methods and fast marching methods*. Cambridge University Press, Cambridge
  67. Shannon CE (1948) A mathematical theory of communication. *Bell Syst Tech J* 27:379–423
  68. Shearer PM (1994) Global seismic event detection using a matched filter on long-period seismograms. *J Geophys Res* 99:13,713–13,735
  69. Shearer PM (1997) Improving local earthquake locations using the L1 norm and waveform cross correlation: Application to the Whittier Narrows, California, aftershock sequence. *J Geophys Res* 102:8269–8283

70. Steinberg DM, Rabinowitz N, Shimshoni Y, Mizrahi D (1995) Configuring a seismographic network for optimal monitoring of fault lines and multiple sources. *Bull Seism Soc Am* 85:1847–1857
71. Stummer P, Maurer HR, Green AG (2004) Experimental Design: Electrical resistivity data sets that provide optimum subsurface information. *Geophysics* 69:120–139
72. Tarantola A (1987) *Inverse problem theory: Methods for data fitting and model parameter estimation*. Elsevier, Amsterdam
73. Tarantola A (2005) *Inverse Problem Theory and Methods for Model Parameter Estimation*. SIAM, Philadelphia
74. Tarantola A, Valette B (1982) Inverse problems = quest for information. *J Geophys Res* 50:159–170
75. Thurber CH, Kissling E (2000) Advances in travel-time calculations for three-dimensional structures. In: Thurber CH, Rabinowitz N (eds) *Advances in Seismic Event Location*. Kluwer, Amsterdam
76. Uhrhammer RA (1980) Analysis of small seismographic station networks. *Bull Seism Soc Am* 70:1369–1379
77. Um J, Thurber C (1987) A fast algorithm for two-point seismic ray tracing. *Bull Seism Soc Am* 77:972–986
78. van den Berg J, Curtis A, Trampert J (2003) Bayesian, nonlinear experimental design applied to simple, geophysical examples. *Geophys J Int* 55(2):411–421. Erratum: 2005. *Geophys J Int* 161(2):265
79. Vidale JE (1988) Finite-difference calculation of travel times. *Bull Seism Soc Am* 78:2062–2078
80. Winterfors E, Curtis A (2007) Survey and experimental design for nonlinear problems. *Inverse Problems* (submitted)
81. Wither M, Aster R, Young C (1999) An automated local and regional seismic event detection and location system using waveform correlation. *Bull Seism Soc Am* 8:657–669
82. Withers M, Aster R, Young C, Beiriger J, Harris M, Moore S, Trujillo J (1998) A comparison of select trigger algorithms for automated global seismic phase and event detection. *Bull Seism Soc Am* 88:95–106
83. Wittlinger G, Herquel G, Nakache T (1993) Earthquake location in strongly heterogeneous media. *Geophys J Int* 115:759–777
84. Zhou H (1994) Rapid 3-D hypocentral determination using a master station method. *J Geophys Res* 99:15439–15455

### Books and Reviews

- Gasparini P, Gaetano M, Jochen Z (eds) (2007) *Earthquake Early Warning Systems*. Springer, Berlin
- Lee WHK, Stewart SW (1981) *Principles and applications of microearthquake networks*. Academic Press, New York
- Thurber CH, Rabinowitz N (eds) (2000) *Advances in Seismic Event Location*. Kluwer, Amsterdam

## Earthquake Magnitude

PETER BORMANN, JOACHIM SAUL  
GeoForschungsZentrum Potsdam, Potsdam, Germany

### Article Outline

Glossary

Definition of the Subject

Introduction to Common Magnitude Scales:

Potential and Limitations

Common Magnitude Estimates

for the Sumatra 2004  $M_w$  9.3 Earthquake

Magnitude Saturation and Biases

Due to Earthquake Complexity

Proposals for Faster Magnitude Estimates

of Strong Earthquakes

Future Requirements and Developments

Bibliography

### Glossary

Technical terms that are written in the text in italics are explained in the Glossary.

**Corner frequency** The frequency  $f_c$  at which the curve that represents the Fourier amplitude spectrum of a recorded seismic signal abruptly changes its slope (see Fig. 5). For earthquakes, this frequency is related to the fault size, rupture velocity, rupture duration and stress drop at the source. Also the frequency at which the magnification curve of a recording system (e.g., Fig. 3) changes its slope.

**Dispersion** Frequency-dependence of the wave propagation velocity. Whereas seismic body-waves show virtually no dispersion, it is pronounced for seismic surface waves. It causes a significant stretching of the length of the surface-wave record and the rather late arrival of its largest amplitudes (Airy phases) from which the surface-wave magnitude  $M_S$  and the mantle magnitude  $M_m$ , respectively, are determined.

**Earthquake size** A frequently used, but not uniquely defined term. It may be related – more or less directly – to either the geometric-kinematic size of an earthquake in terms of area and slip of the fault or to the *seismic energy* radiated from a seismic source and its potential to cause damage and casualty (moment or energy *magnitude*).

**Earthquake source** In general terms, the whole area or volume of an *earthquake* rupture where seismic body waves are generated and radiated outwards. More specifically, one speaks either of the *source mechanism*

or the source location. The latter is commonly given as earthquake hypocenter (i. e. the location at the source depth  $h$  from where the seismic rupture, collapse or explosion begins) or as the point on the Earth's surface vertically above the hypocenter, called the epicenter. Earthquakes at  $h < 70$  km are shallow, those at larger depth either intermediate (up to  $h = 300$  km) or deep earthquakes ( $h = 300$ – $700$  km). The determination of the geographical coordinates latitude  $\varphi$ , longitude  $\lambda$ , and focal depth  $h$ , is the prime task of seismic source location. However, for extended seismic sources, fault ruptures of great earthquakes in particular, the hypocenter is generally not the location of largest fault slip and/or seismic moment/energy release and the epicenter is then also not the location where the strongest ground shaking is felt. The locations of largest effects may be dozens of kilometers in space and many seconds to minutes in time away from the hypocenter or epicenter, respectively.

**Fundamental modes** The longest period oscillations of the whole Earth with periods of about 20 min (spheroidal mode), 44 min. (toroidal mode) and some 54 min (“rugby” mode), excited by great earthquakes.

**Magnitude** A number that characterizes the relative *earthquake size*. It is usually based on measurement of the maximum motion recorded by a seismograph (sometimes for waves of a particular type and frequency) and corrected for the decay of amplitudes with epicenter distance and source depth due to geometric spreading and attenuation during wave propagation. Several magnitude scales have been defined. Some of them show *saturation*. In contrast, the moment magnitude ( $M_w$ ), based on the concept of *seismic moment*, is uniformly applicable to all earthquake sizes but is more difficult to compute than the other types, similarly the energy magnitude,  $M_e$ , which is based on direct calculation of the *seismic energy*  $E_s$  from broadband seismic records.

**Saturation** (of magnitudes) Underestimation of *magnitude* when the duration of the earthquake rupture significantly exceeds the seismic wave period at which the magnitude is measured. The shorter this period, the earlier respective magnitudes will saturate (see relation (13) and Figs. 4 and 5).

**Seismic energy** Elastic energy  $E_s$  (in joule) generated by, and radiated from, a seismic source in the form of seismic waves. The amount of  $E_s$  is generally much smaller than the energy associated with the non-elastic deformation in the seismic source (see *seismic moment*  $M_o$ ). The ratio  $E_s/M_o = (\Delta\sigma/2\mu) = \tau_a/\mu$ , i. e., the seismic energy released per unit of  $M_o$ , varies for earthquakes

in a very wide range between some  $10^{-6}$  and  $10^{-3}$ , depending on the geologic-tectonic environment, type of *source mechanism* and related stress drop  $\Delta\sigma$  or apparent stress  $\tau_a$ .

**Seismic moment  $M_o$**  A special measure of earthquake size. The moment tensor of a shear rupture (see *earthquake source*) has two non-zero eigenvalues of the amount  $M_o = \mu \bar{D} F_a$  with  $\mu$ -shear modulus of the ruptured medium,  $\bar{D}$ -average source dislocation and  $F_a$ -area of the ruptured fault plane.  $M_o$  is called the scalar seismic moment. It has the dimension of Newton meter (Nm) and describes the total non-elastic (i. e., ruptural and plastic) deformation in the seismic source volume. Knowing  $M_o$ , the moment *magnitude*  $M_w$  can be determined via Eq. (11).

**Source mechanism** Depending on the orientation of the earthquake fault plane and slip direction in space, one discerns different source mechanisms. Strike-slip faults are vertical (or nearly vertical) fractures along which rock masses have mostly shifted horizontally. Dip-slip faults are inclined fractures. If the rock mass above an inclined fault moves down (due to lateral extension) the fault is termed normal, whereas, if the rock above the fault moves up (due to lateral compression), the fault is termed reverse (or thrust). Oblique-slip faults have significant components of both slip styles (i. e., strike-slip and dip-slip). The greatest earthquakes with the largest release of seismic moment and the greatest potential for generating tsunamis are thrust faults in subduction zones where two of Earth's lithosphere plates (e. g., ocean-continent or continent-continent) collide and one of the two plates is subducted underneath the overriding plate down into the Earth's mantle. Different source mechanisms are characterized by different radiation patterns of seismic wave energy.

**Transfer function** The transfer function of a seismic sensor-recorder system (or of the Earth medium through which seismic waves propagate) describes the frequency-dependent amplification, damping and phase distortion of seismic signals by a specific sensor-recorder (or medium). The modulus (absolute value) of the transfer function is termed the amplitude-frequency response or, in the case of seismographs, also magnification curve (see Fig. 3).

### Definition of the Subject

Besides earthquake location (i. e., the determination of the geographical coordinates of the epicenter, the hypocenter depth and the origin time; for definition of these terms see

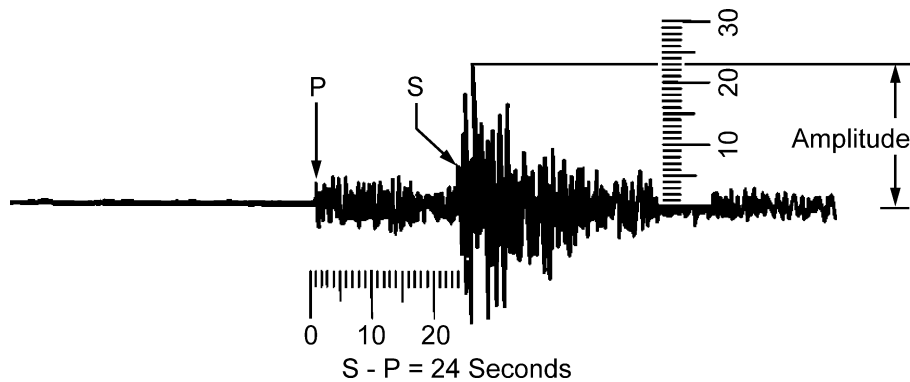
*earthquake source* in the Glossary), the *magnitude* is the most frequently determined and commonly used parameter to characterize an earthquake. Despite its various imperfections, it provides important information concerning the earthquake source spectrum at the period where the magnitude is measured and current source theories (cf. [3]) allow one to understand differences in the source spectra of different earthquakes in terms of source dimension and stress drop, i. e., the difference between the stress level before and after the earthquake. Via various empirical relations, magnitudes enable estimates of the *seismic moment* and the *seismic energy* released by the earthquake. These parameters are important in the discussion of various global problems such as the seismic slip rates between lithosphere plates and the excitation of Chandler Wobble [25]. Besides these more academic issues, magnitude values have an immense practical value in providing:

- Rapid simple parameter estimates of the strength of an earthquake that can help to realistically assess the related ground shaking or tsunami potential and thus assist efficient disaster management response;
- Mass data in earthquake catalogs and data banks, covering long time periods over many decades – and hopefully centuries in future, which allows one to assess the seismic activity and related hazards of Earth's regions and their possible variability in space and time. This is not only of high scientific interest, but also the very basis for realistic long-term disaster preparedness and risk mitigation efforts.

The term magnitude and the basic method of its determination were introduced by Charles F. Richter in 1935 [71]. He intended to compare the relative *earthquake size* in southern California in terms of differences in the maximum amplitudes  $A$  recorded at a network of seismic stations that were equipped with standard short-period Wood-Anderson (WA) torsion seismometers.

The WA seismometer response is depicted in Fig. 3 and Fig. 1 shows a WA record and magnitude measurement example. In order to make amplitudes recorded by stations at different epicentral distances  $D$  from the earthquake comparable, Richter had to compensate for the amplitude decay with  $D$  using an appropriate correction term  $-A_0(D)$ . Since the strength and thus the radiated amplitudes of earthquakes vary in a wide range Richter defined his local magnitude scale  $M_L$ , determined from records at source distances up to 600 km, as follows:

*“The magnitude of any shock is taken as the logarithm of the maximum trace amplitude, expressed in microns, with which the standard short-period tor-*



Earthquake Magnitude, Figure 1

Record of a short-period Wood-Anderson seismograph (frequency-magnification curve see Fig. 3) of a local earthquake. *P* marks the onset of the first arriving longitudinal *P* wave, and *S* the onset of the much stronger secondary, transverse polarized shear wave. Note the long tail of coda-waves following *S*. From the time difference  $S - P = 24$  s follows a hypocentral distance  $R = 190$  km. The maximum record amplitude is  $A_{\max} = 23$  mm. Applying the amplitude-distance correction  $-\log A_o(190 \text{ km}) = 3.45$  according to Richter [72] results in a magnitude  $M_L = 4.8$

*sion seismometer ... would register that shock at an epicentral distance of 100 km."*

Thus:

$$M_L = \log A_{\max} - \log A_o(D). \quad (1)$$

According to the above definition, an amplitude of  $1 \mu\text{m}$  in a WA record at a distance  $D = 100$  km from the epicenter would correspond to  $M_L = 0$ . Amplitude means in (1) and the following text either the center-to-peak or half of the peak-to-trough amplitude.

Wood-Anderson (WA) seismographs record horizontal short-period ground motions with an amplification of only about 2080 times [82]. Modern electronic seismographs may achieve magnifications larger than  $10^6$  and thus are able to record local earthquakes with even negative magnitudes, down to about  $-2$ . The largest values determined with the  $M_L$  scale are around seven. Later it was found that all magnitudes derived from short-period waves (typically with periods  $T < 3$  s) show *saturation* (see Glossary, Fig. 4 and Sect. "Magnitude Saturation and Biases Due to Earthquake Complexity"). Therefore, it was necessary to develop complementary magnitude scales that use medium to long-period ( $T \approx 5$  s – 30 s) as well as very long-period waves ( $T \approx 50$  s – 3000 s) in order to enable less or non-saturating magnitude estimates (see Sect. "Introduction to Common Magnitude Scales: Potential and Limitations"). For the so far strongest instrumentally recorded earthquake (Chile 1960) a value of  $M = 9.5$  was determined that way. Accordingly, instrumental seismic monitoring currently covers the magnitude range of

about  $-2 \leq M < 10$ . This roughly corresponds to ruptures of some millimeters to more than 1000 km long. They radiate approximately the same amount of seismic wave energy  $E_s$  as well-contained underground explosions with yields ranging from a few milligrams ( $10^{-9}$  t) to several 10 to 100 Gt ( $1 \text{ Gt} = 10^9$  t) Trinitrotoluol (TNT) equivalent, thus covering about 20 orders in energy. Earthquakes with magnitudes around four may cause only minor local damage, those with magnitudes  $> 6$  heavy damage, and those with magnitudes  $> 7$  already widespread devastating damage. Shallow submarine earthquakes with magnitudes  $> 7$  may generate significant local tsunamis with damage potential to nearby shores whereas those with magnitudes  $> 8.5$  may stimulate ocean-wide tsunamis causing destruction and casualties even at shores thousands of kilometers away from such earthquakes.

In order to measure and classify *earthquake size* in the wide range of magnitudes from about  $-2$  to  $< 10$  and satisfy specific requirements in research and application which are based on magnitude data, it was indispensable to develop different magnitude scales that are complementary, but properly scaled to the original Richter  $M_L$ . Thus, there exists today a host of magnitude scales applicable in a wide range of source distances from less than 1 km up to more than 10,000 km. These scales, their specifics, potential and limitations are discussed in detail (with many reference given) in Chapter 3 of the IASPEI New Manual of Seismological Observatory Practice [6]. The early pioneers of magnitude scales, Beno Gutenberg and Charles Richter, had hoped that different magnitude scales could be cross-calibrated to yield a unique value for any given



earthquake (cf. [25,30]). In their joint book [29] “Seismicity of the Earth” (1954; first edition 1949) and later in Richter’s [72] famous text book “Elementary Seismology” as well as in Duda [22] only one magnitude value  $M$  was given per earthquake. However, this approach proved only partially realistic under certain conditions and within limited magnitude ranges because of the often significant differences in measurement procedures as well as period and bandwidth ranges used in later magnitudes scales. Decades later it took significant efforts (cf. [1,2,25]) to reconvert these  $M$  values, which turned out to be not even compatible (cf. [25]) into their original body or surface wave magnitudes in order to get values that agree with the original definition of these specific magnitude scales and can be compared with current data of the same type.

In general, such magnitude conversion relations strongly depend on initial data errors and the type of least-square regression procedure applied [11,14]. Moreover, the latter have often not been interpreted and used in a correct way. This may result in the case of noisy magnitude data for events at the upper and lower end of the investigated magnitude range, in conversion errors of more than 0.5 magnitude units (m.u.) with serious consequences on seismic hazard estimates based on such converted magnitudes (cf. [7,11,14,15]). Moreover, magnitude values determined within the *saturation* range of a given scale cannot reliably be converted via empirical regression relations into the equivalent magnitude values of another less or non-saturating magnitude scale (see Fig. 4 and [44]). Furthermore, some magnitudes relate best to the released *seismic energy* while others are scaled to the static *seismic moment*, i. e., they measure equally important but fundamentally different physical aspects of the source and the radiated seismic waves and may differ by sometimes more than 1 m.u. Thus there is no way to characterize *earthquake size* in all its different aspects by just a single magnitude value. Proper interpretation and use of different types of magnitude data, however, requires one to understand the physics behind such values and how these may be affected by the complexity and duration of the earthquake rupture process. Further, this necessitates one to discriminate unambiguously the different types of magnitude values by using a unique nomenclature and to assure that magnitude values published with a given nomenclature have been determined with an internationally agreed standard procedure. With this in mind, the most important magnitude scales and related problems are summarized in Sects. “**Introduction to Common Magnitude Scales: Potential and Limitations**” and “**Common Magnitude Estimates for the Sumatra 2004  $M_w$  9.3 Earthquake**”.

## Introduction to Common Magnitude Scales: Potential and Limitations

### Magnitude Scales Used in the Local and Regional Distance Range ( $D < 2000$ km)

The original Richter local magnitude scale for Southern California [71] has been further developed since its invention [38]. In its expanded form (with the nomenclature  $M_L$  common in the United States), the following relation now holds:

$$M_L = \log_{10}(A_{\max}) + 1.11 \log_{10} R + 0.00189 R - 2.09 \quad (2)$$

with  $R$  = distance from the station to the hypocenter in kilometers and  $A_{\max}$  = maximum trace amplitude in nanometers (instead of  $\mu\text{m}$  in a WA record). This amplitude is measured on the output from a horizontal-component seismograph that is filtered so that the response of the seismograph/filter system replicates that of a WA standard seismograph but with a static magnification of one. The underlying procedure of  $M_L$  determination according to relation (2) was adopted by the International Association of Seismology and Physics of the Earth’s Interior (IASPEI) in 2004 as the standard procedure for determining local magnitudes in the distance range up to typically less than 1000 km [42]. For earthquakes in the Earth’s crust of regions with attenuation properties that differ from those of coastal California, and for measuring  $M_L$  with vertical component seismographs, the standard equation takes the form:

$$M_L = \log_{10}(A_{\max}) + F(R) + G \quad (3)$$

where  $F(R)$  is an  $R$ -dependent calibration function and  $G$  a constant which have to compensate for different regional attenuation and/or for any systematic biases of amplitudes measured on vertical instead on horizontal seismographs. Examples of regional  $M_L$  calibration functions developed for different parts of the world have been compiled by Borrmann (Chap. 3, p. 26, and DS 3.1 in [6]).

A few decades ago, analog seismic records prevailed. They had a rather limited dynamic range of only some 40 dB. This caused record traces often to go off-scale when stronger seismic events were recorded at local or regional distances. Then  $A_{\max}$  could not be measured. Yet, it was found that the duration  $\mathbf{d}$  of the coda that follows  $A_{\max}$  with exponentially decaying amplitudes (see Fig. 1) increases with magnitude and distance  $D$ . On this basis, local duration magnitude formulas of the following general form

$$M_d = a + b \log \mathbf{d} + cD \quad (4)$$

have been developed with  $a$ ,  $b$  and  $c$  being coefficients to be determined locally. When using only recordings at distances  $D < 100$  km the distance term  $cD$  is not even needed. However, crustal structure, scattering and attenuation conditions vary from region to region. Moreover, the resulting specific equations will also depend on the chosen definition for  $\mathbf{d}$ , the local signal-to-noise (SNR) conditions and the sensor sensitivity at the considered seismic station(s) of a network. Therefore,  $M_d$  scales have to be determined locally for a given source-network configuration and scaled to the best available amplitude-based  $M_L$  scale.

Nowadays digital recorders with large usable dynamic range of about 140 dB are common. Thus even sensitive modern broadband seismographs remain on scale when recording local or regional earthquakes up to  $M \approx 7$ . This reduces the need for  $M_d$  scales. Moreover, the increasing availability of modern strong-motion (SM) recorders with comparably large dynamic range, which will not clip even in the case of very strong nearby earthquakes, have led to the development of (partially) frequency-dependent  $M_L^{SM}$  scales. They are usually based on the calculation of synthetic WA seismograph outputs from strong-motion accelerograms [35,54].

Also, amplitudes of short-period  $L_g$  waves with periods around 1 s are sometimes used to determine magnitudes, termed  $m_b(L_g)$ .  $L_g$  waves travel with group velocities of 3.6 to 3.2 km/s and arrive after the (secondary, shear) S wave onset (Fig. 1). They propagate well in continental platform areas. Recently, the IASPEI [42] adopted a measurement procedure for  $m_b(L_g)$  as international standard, which had been developed for eastern North America [62] with the aim to improve yield estimates of Nevada Test Site explosions. However, as for all other local or regional magnitude scales, the calibration term is strongly influenced by the local/regional geologic-tectonic conditions in the Earth's crust and requires a proper scaling to this standard, when applied to other areas than eastern North America.

Tsuboi developed for the Japan Meteorological Agency (JMA) in 1954 [79] a magnitude formula for shallow earthquakes (depth  $h < 60$  km) that have been recorded at epicentral distances  $D$  up to 2000 km:

$$M_{JMA} = \log_{10} A_{\max} + 1.73 \log_{10} D - 0.83. \quad (5)$$

$A_{\max}$  is the largest ground motion amplitude (in  $\mu\text{m}$ ) in the total event record of a seismograph with an eigenperiod of 5 s. If horizontal seismographs are used then  $A_{\max} = (A_{NS}^2 + A_{EW}^2)^{1/2}$  with  $A_{NS}$  and  $A_{EW}$  being half the maximum peak-to-trough amplitudes measured in the two horizontal components. This formula was devised to be equivalent to the medium to long-period Gutenberg–

Richter [29] magnitude  $M$ . Therefore,  $M_{JMA}$  agrees rather well with the seismic moment magnitude  $M_w$ . The average difference is less than 0.1 in the magnitude range between 4.5 and 7.5 but becomes  $> 0.5$  for  $M_w > 8.5$  (see Fig. 4). Katsumata [49,50] has later modified the  $M_{JMA}$  formula for earthquakes deeper than 60 km.

Another, more long-period regional moment magnitude scale, termed  $M_{wp}$ , has been developed in Japan as well [80]. It provides quick and less saturating magnitude estimates for tsunami early warning. Velocity-proportional records are twice integrated and approximately corrected for geometrical spreading and an average  $P$ -wave radiation pattern (see *source mechanism*) to obtain estimates of the scalar seismic moment  $M_o$  at each station. Usually the first maximum in the integrated displacement trace, called “moment history”  $M_o(t)$ , is assumed to represent  $M_o$ . From these  $M_o$  values moment magnitudes  $M_w$  are then calculated for each station according to Eq. (11) and averaged.  $M_{wp}$  results from adding an empirically derived correction of 0.2 m.u. to the averaged station  $M_w$  [80]. Finally, a magnitude-dependent correction is applied to  $M_{wp}$  [86] in order to get an even better estimate of the recognized “authoritative” Global Centroid Moment Tensor magnitude  $M_w$  (GCMT) which is calculated according to the Harvard procedure [23] and now published under [41].

The  $M_{wp}$  concept was originally developed for earthquakes at  $5^\circ \leq D^\circ \leq 15^\circ$ , but can be applied for  $M_w < 7.5$  (down to about  $M_w \approx 5$ ) even to shorter local distances as long as this distance is significantly larger than the rupture length. Later the  $M_{wp}$  procedure has been adopted for application to records of deep and teleseismic earthquakes as well [81].  $M_{wp}$  estimates are standard routine in Japan, at the Alaska and the Pacific Tsunami Warning Centers (ATWC and PTWC), and the National Earthquake Information Center (NEIC) of the United States Geological Survey (USGS). However, each of these centers use slightly different procedures. Values for most strong earthquakes are usually available some 10 to 15 min after the origin time (OT). On average  $M_{wp}$  data scale well with  $M_w$ . Exceptions, however, are extremely slow or very large complex earthquakes. Then  $M_{wp}$  is usually too small, up to about 1 m.u.

In recent years great attention is paid to the development of even more rapid earthquake early warning systems (EWS). They aim at event location and magnitude estimates from the very first few seconds of broadband acceleration, velocity or displacement records and within about 10 to 30 s after origin time (OT) of strong damaging earthquakes on land. These data are to be used for instantaneous public alarms and/or automatically triggered risk

mitigation actions after strong earthquakes with damage potential. The goal is to minimize the area of “blind zones” which are left without advanced warning before the arrival of the  $S$  waves which have usually the largest strong-motion amplitudes (see Fig. 1). This necessitates very dense and robust local seismic sensor networks within a few tens of kilometers from potentially strong earthquake sources. Such networks are at present available only in very few countries, e. g. in Japan, Taiwan, Turkey, and Italy.

Their principles of rapid magnitude estimates differ from those mentioned above and below and the data analysis from such systems is largely based on still much debated concepts such as the hypothesis of the deterministic nature of earthquake rupture [66,73]. Data presented in [66] seem to suggest that in the range  $3.0 < M$  (not specified)  $< 8.4$  the magnitude can be estimated with an average absolute deviation of 0.54 m.u. from the maximum period within the initial 4 s of the first arriving (primary, longitudinal)  $P$  wave when many low-pass filtered velocity records within 100 km from the epicenter are available. However, for  $M > 6$  the systematic increase of these greatly scattering periods becomes rather questionable. When analyzing waveforms of the Japanese Hinet seismic network [73], it could not be confirmed that such a dominant frequency scaling with magnitude exists. Also Kanamori [46], together with Nakamura [60,61], one of the fathers of this idea, expressed much more caution about the prospects of this method after he had run, together with Wu [89], an experiment with the Taiwan EWS. For each event they analyzed the first 3 s of at least eight  $P$ -wave records at epicentral distances  $< 30$  km. They knew that: “... the slip motion is in general complex and even a large event often begins with a small short-period motion, followed by a long-period motion. Consequently, it is important to define the average period during the first motion.” (termed  $\tau_c$  in [46,89]). However, after applying the  $\tau_c$  concept to the Taiwan EWS they concluded: “For EWS applications, if  $\tau_c < 1$  s, the event has already ended or is not likely to grow beyond  $M > 6$ . If  $\tau_c > 1$  s, it is likely to grow, but how large it will eventually become, cannot be determined. In this sense, the method provides a threshold warning”. Thus it seems that these new concepts work reasonably well only for earthquakes with  $M < 6.5$  and thus total rupture durations that are according to Eq. (13) on average not more than about 2–3 times the measurement time windows of 3 s or 4 s used in [46,66,89]. Nakamura and Saita [61] reported data from a much smaller set of events ( $N = 26$ ) recorded at local distances in the range  $4.6 < M < 6.9$ . We calculated the average absolute deviation of their rapid UrEDAS system magnitudes (0.47 m.u.) from the official magnitudes  $M_{JMA}$

published later by the Japan Meteorological Agency. This error decreases to 0.32 m.u. when only earthquakes with magnitudes up to  $M_{JMA} = 6.0$  are considered. This seems to support our assessment that the reliability of real-time EMS magnitudes decreases rapidly if the analyzed time window is much shorter than the rupture duration.

### Magnitude Scales Used in the Teleseismic Distance Range ( $D > 2000$ km)

Ten years after the introduction of the local magnitude  $M_L$ , Beno Gutenberg [26,27,28] extended the concept of magnitude determination to teleseismic distances larger than about 1000–2000 km. He used both records of seismic waves that propagate along the Earth’s surface (or near to it with a period-dependent penetration depth) and waves which travel through the Earth. Accordingly, the former are termed surface waves and the latter body waves. For the surface-wave magnitude Gutenberg [28] gave the following relation:

$$M_S = \log_{10} A_{Hmax} + 1.656 \log D^\circ + 1.818 \quad (6)$$

with  $A_{Hmax}$  = maximum “total” horizontal displacement amplitude of surface-waves in  $\mu\text{m}$  for periods around  $20 \pm 2$  s measured in the distance range  $15^\circ < D^\circ < 130^\circ$  ( $1^\circ = 111,195$  km).

While the original Richter  $M_L$  and Gutenberg  $M_S$  magnitudes were calculated from the maximum ground displacement amplitudes, Gutenberg [26,27,30] proposed to determine the body-wave magnitudes  $m_B$  from the relation:

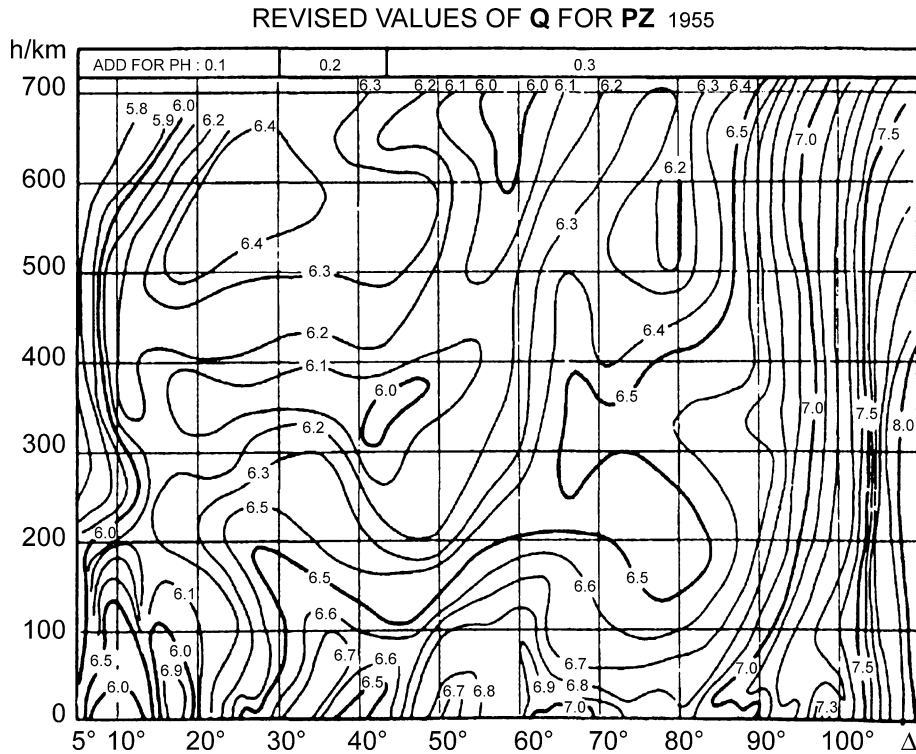
$$m_B = \log_{10} (A/T)_{max} + Q(D^\circ, h), \quad (7)$$

i. e., by measuring the maximum ratio of ground displacement amplitude  $A$  (in  $\mu\text{m}$ ) divided by the related period  $T$  (in s).  $A/T$  is equivalent to measuring the maximum ground motion velocity  $A_{vmax}/2\pi$  which is proportional to the square root of seismic energy, i. e.  $\sqrt{E_s}$ . Thus the magnitude becomes a measure of the elastic kinetic wave energy radiated by an earthquake. Only in this way comparable magnitude data could be obtained for different types of body waves and measurements at different sites. Another great advantage of  $m_B$  is that it permits magnitude estimates also from intermediate and deep earthquake sources, which produce only weak or no surface waves at all. Empirical relationships permit estimating  $E_s$  (in units of Joule) from body-wave magnitude  $m_B$  [30]

$$\log_{10} E_s = 2.4 m_B - 1.2 \quad (8)$$

or surface-wave magnitude  $M_S$  [72]

$$\log_{10} E_s = 1.5 M_S + 4.8. \quad (9)$$



Earthquake Magnitude, Figure 2  
 Calibration values  $Q(D^\circ, h)$  for vertical (Z) component P-wave amplitudes depending on epicentral distance  $D^\circ = \Delta$  and source depth  $h$  as used in the calculation of body-wave magnitudes  $m_b$  and  $m_B$  according to Gutenberg and Richter, 1956 [30]

Accordingly, an increase by 1 m.u. in  $m_B$  and  $M_S$  corresponds to an increase of radiated seismic energy by about 250 and 30 times, respectively.

Revised empirical distance-depth corrections for the calibration of body-wave magnitudes, so-called Q-functions, were published in 1956 by Gutenberg and Richter [30]. They are given as separate tables and charts for the body-wave phases P, PP (a P wave reflected at the surface of the Earth about the half way between source and station) and S. They are still in use, especially  $Q_{PV}$  for calibrating amplitude measurements made on vertical component P-wave records (Fig. 2). However, for epicenter distances between 5° and 20° these calibration values are not reliable enough for global application. In this range the wave propagation is strongly affected by regional variations of the structure and properties of the Earth’s crust and upper mantle. And for  $D > 100^\circ$  the P-wave amplitudes decay rapidly because of the propagation of P waves is influenced by the Earth’s core (so-called core shadow). Therefore, in agreement with current IASPEI recommendations [42],  $m_B$  and its short-period complement  $m_b$  (see below), should be determined by using  $Q_{PV}$  only between  $21^\circ \leq D \leq 100^\circ$ .

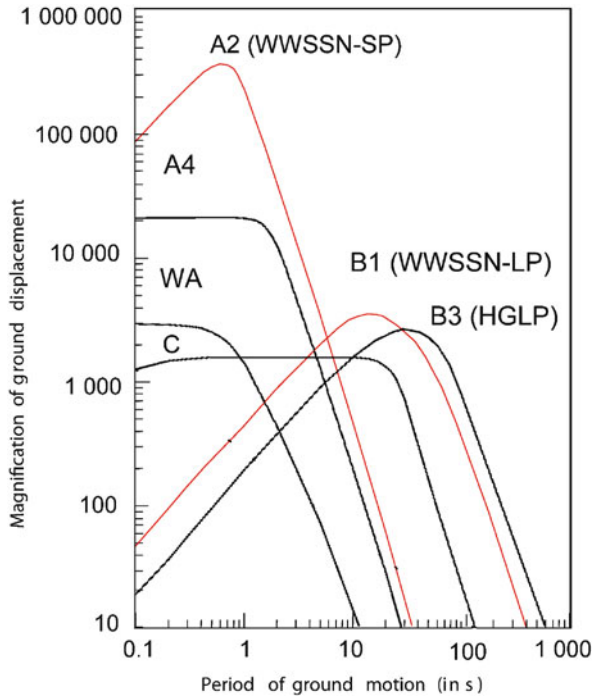
These body-wave magnitude calibration functions had been derived from amplitude measurements made mostly on medium-period broadband displacement records which dominated during the first half of the 20th Century at seismological stations. Their period-dependent magnification curve resembled more or less that of the classical standard seismograph type C shown in Fig. 3, although for some of these instruments the roll-off of the amplification occurred already at periods  $T > 10$  s.

Another, so-called Prague–Moscow formula for surface-wave magnitudes was proposed in 1962 by Vaněk et al. [84]. It is based on the measurement of  $(A/T)_{\max}$  in records of shallow earthquakes ( $h < 60$  km) in wide period and distance ranges ( $3 \text{ s} < T < 30 \text{ s}$ ;  $2^\circ \leq D^\circ \leq 160^\circ$ ):

$$M_S = \log_{10}(A/T)_{\max} + 1.66 \log_{10} D^\circ + 3.3 . \quad (10)$$

This relationship, which is – as Eq. (7) – more directly related to  $E_s$ , was adopted by the IASPEI in 1967 as international standard.

The NEIC adopted Eq. (10), but continues to limit the range of application to distances between  $20^\circ \leq D^\circ \leq$

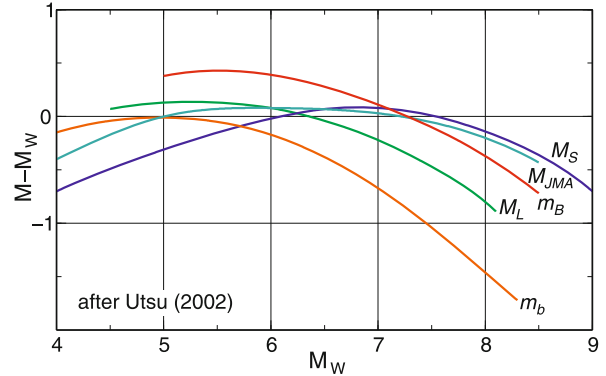


Earthquake Magnitude, Figure 3

**Magnification of ground displacement amplitudes by common standard types of seismographs.** WA = Wood–Anderson seismograph; WWSSN-SP and WWSSN-LP = short-period and long-period seismographs used in the former United States World-Wide Seismograph Standard Network; HGLP = US type of High Gain Long Period seismographs; A2, A3, B1, B3 and C = standard types of seismographs according to Willmore [87]. Reprint from [6] with © granted by IASPEI

$160^\circ$  and displacement amplitudes in the very limited period range as in formula (6) although Soloviev [74] had shown already in 1955 that  $(A/T)_{\max}$  is a stable quantitative feature of surface waves whatever the period of their maximum at all epicentral distances. Also theory has confirmed [63] that using the ratio  $(A/T)$  is a partial and ad hoc compensation for a large number of frequency-dependent terms ignored in (10). In fact, the periods at the surface-wave maximum used for  $M_S$  determination vary in a wide range between some 3 s and 25 s and show – despite large scatter – a clear distance dependence [84,87]. Therefore, several authors [36,70] showed that using Eq. (10) only for amplitude readings around 20 s results in systematic distance-dependent biases. However, their proposed revised calibration functions for 20 s waves are not yet used in routine practice at international seismological data centers.

The formulas (6) and (10) had originally been developed for horizontal component amplitude readings. Be-



Earthquake Magnitude, Figure 4

**Average relationships between different common types of magnitudes and the moment magnitude  $M_w$ .** Modified from Fig. 1 in [83]

ginning in the 1960s, however, more and more long-period and broadband vertical component instruments became available and are now commonly used for magnitude determination from surface waves. This procedure is easier and better defined than measuring and combining the amplitude measurements made in two horizontal components, yields on average values that are largely comparable with the Gutenberg  $M_S$  [25] and has recently been adopted as IASPEI [42] standard. Herak et al. [37] published theoretical and observed depth corrections for  $M_S(20)$  when determined according to (10). These corrections allow determination of more reliable surface-wave magnitudes for earthquakes in all depth ranges and improve significantly the relationship between  $M_S$  and the seismic moment  $M_0$ .

In the 1960s, the United States deployed a World-Wide Standard Seismograph Network (WWSSN) equipped with short-period (SP) and long-period (LP) seismographs of limited bandwidth (cf. Fig. 3). This network had two priority tasks. Firstly, to significantly increase the signal-to-noise ratio of the seismic records by narrow-band short-period filtering, thus improving the global detection threshold for teleseismic events down to magnitudes around 4–4.5 and the location accuracy for seismic events. Secondly, to realize an effective discriminator between underground nuclear explosions (UNE) and natural earthquakes based on the ratio of a short-period body-wave and a long-period surface-wave magnitude. Natural earthquakes have a much longer source duration (seconds to minutes) than explosions of comparable size (typically milliseconds). Also, at comparable seismic moment magnitude, UNEs radiate significantly more high-frequency energy (see dotted curve in Fig. 5) than earthquakes. Therefore, a better discrimination of the two types of events was

achieved by measuring the  $P$ -wave amplitude only at periods  $< 3$  s (typically around 1 s) and calculating a short-period  $P$ -wave magnitude termed  $m_b$ . In contrast, the Gutenberg  $m_B$  is based on measuring  $A_{\max}$  at periods  $T$  usually between 2 s and 30 s. Further, during the first two decades of the WWSSN, the  $P$ -wave amplitude was not – as required by Gutenberg’s procedure for  $m_B$  determination – always measured at the maximum of the whole  $P$ -wave train (whose length depends on the source duration and thus on the magnitude itself) but initially within the first five half-cycles only and later by USGS requirement in the first 5 s of the record.

Because of the short source duration of explosions, their  $P$ -waves will always reach maximum amplitudes within such a short time-interval. However,  $P$  waves radiated by large earthquakes of much longer source duration will reach their maximum amplitude usually much later in the rupture process. For magnitude 6 the average rupture duration is on average already 6 s, and may increase to about 600 s for the strongest earthquakes (cf. relation (13)). Both effects together, plus the fact, that  $m_b$  was still computed using the  $Q_{PV}$  function derived for mainly medium-period  $P$  waves, which are much less affected by frequency-dependent attenuation than 1 Hz  $P$  waves, resulted in a systematic underestimation of the earthquake size for magnitudes larger than 5 and a *saturation* of  $m_b$  at around 6.5.

In the late 1970s, the NEIC switched back to a longer time window of about 15 s and more recently, with an automatic procedure, to a window covering the first 10 cycles of short-period teleseismic  $P$  waves. In the case of strong earthquakes this window may later be extended interactively up to 60 s. This mitigates to some extent the saturation of  $m_b$ . However, no  $m_b$ -values larger than 7.2 have ever been measured with this procedure. On the other hand  $m_b$  yields rather reliable values for magnitudes  $< 5$  when the corner frequency of the average source spectra falls within the passband of the short-period seismograph or is even more high frequency (cf. Figs. 3, 4, 5). For magnitudes  $< 5$   $m_B$  can usually no longer be determined because of too small signal-to-noise ratio (SNR) in broadband records. Then  $m_b$  is often the only available teleseismic estimator of earthquake size for small earthquakes.

Most seismic stations and networks worldwide adopted the US procedure for  $m_b$  measurement and – with the exception of Russia, China and their former allies – completely abandoned measuring  $m_B$  as originally defined. This change in attitude was stimulated by the fact that the NEIC, which serves in fact as one of the leading international data centers for seismology, did not accept reported  $P$ -wave amplitudes other than those obtained

from short-period measurements. Some stations, national and global data centers continue (at least up to 2008) to measure for  $m_b$  the maximum amplitude of  $P$  exclusively within the first 5 s after the  $P$ -wave *first arrival*, such as the China Earthquake Network Center and the International Data Center (IDC) of the Comprehensive Test-Ban Treaty Organization (CTBTO) in Vienna.

Because of these inconsistencies in  $m_b$  and  $M_S$  determination and the proven merits of both broadband  $m_B$  and  $M_S$  (see also [11]) the IASPEI Working Group on Magnitude Measurements recommended that in future:

- $m_b$  is always determined from  $A_{\max}$  at periods  $T < 3$  s within the whole  $P$ -wave train;
- The band-limited magnitudes  $m_b$  and  $M_S(20)$  be complemented by true broadband magnitudes  $m_B$  and  $M_S(BB)$ . The latter two will be obtained by measuring  $A_{v\max}$  on unfiltered velocity broadband records and thus always include the maximum velocity amplitudes of the source spectrum in the magnitude range of interest (cf. Fig. 5). This will link these two broadband magnitudes to the seismic energy released by an earthquake, more closely than the common band-limited magnitudes.

These recommendations have been adopted by the IASPEI Commission on Seismic Observation and Interpretation (CoSOI) in 2005 as new magnitude measurement standards. More details about the new measurement procedures for  $m_b$ ,  $m_B$ ,  $M_S(20)$  and  $M_S(BB)$  are given on the CoSOI web site [42]. Beginning in 2007 they are gradually implemented at the main seismological data centers and networks.

Since all magnitudes discussed so far show more or less pronounced *saturation* for large earthquakes (cf. Fig. 4 and [44]) a non-saturating magnitude, termed  $M_w$ , has been proposed [31,43,69]. The moment magnitude  $M_w$  is derived from the scalar *seismic moment*  $M_o$  via the relation

$$M_w = (2/3)(\log_{10} M_o - 9.1). \quad (11)$$

$M_o$  has the dimension of Newton meter (Nm) and expresses the total inelastic “work” required for rupturing and displacing the considered earthquake fault. It can be determined either by waveform analysis and inversion in the time domain or by measuring the spectral amplitude  $u_{0p,s}$  of the low-frequency level (plateau) of the displacement spectrum of  $P$  or  $S$  waves (cf. Fig. 5) via the relationship

$$M_o = 4\pi r \rho v_{p,s}^3 u_{0p,s} / R_{\theta,\phi}^{p,s} \quad (12)$$

with  $r$  = hypocenter distance,  $\rho$  = average density of rocks in the source and receiver area,  $v_{p,s}$  = average velocity of the  $P$  or  $S$  waves from the source to the receiver area and  $R_{\theta,\phi}^{p,s}$  = a factor correcting the observed seismic amplitudes for the influence of the radiation pattern of the given *source mechanism*, which is different for  $P$  and  $S$  waves.

$M_0$  is expected to show no *saturation*, provided that the amplitude level is measured only at periods significantly larger than the magnitude-dependent *corner period* of the seismic source spectrum (cf. Fig. 5). In Sects. “**Common Magnitude Estimates for the Sumatra 2004  $M_w$  9.3 Earthquake**” and “**Magnitude Saturation and Biases Due to Earthquake Complexity**” we will show, however, that incorrect determination of  $M_0$  may still result in an underestimation of the earthquake size. Since  $M_w$  is derived from  $M_0$  it is related to the tectonic effect of earthquakes, i. e., to the product of rupture area and average fault slip and thus also relevant to assess the tsunami potential of strong shallow marine earthquakes. An example is the off-shore Nicaragua earthquake of 2 September 1992. Its  $m_b = 5.3$  was too weak to alert the people ashore, some 70–120 km away from the source area. However, its  $M_w = 7.6$  was much larger and caused a damaging local tsunami with almost 200 casualties.

Yet,  $M_0$  and thus  $M_w$  do not carry any direct information about the dominant frequency content and thus of the seismic energy released by the earthquake (cf. Sect. “**Magnitude Saturation and Biases Due to Earthquake Complexity**”). In fact, relation (11) was derived by assuming constant stress drop and an average ratio of  $E_s/M_0 = 5 \times 10^{-5}$  on the basis of elastostatic considerations and empirical data [43] and then replacing in Eq. (9)  $M_S$  by  $M_w$ .

As source theory has advanced and broadband digital data have become readily available, the radiated *seismic energy*  $E_s$  could be computed explicitly rather than from an empirical formula. Boatwright and Choy (cf. [5,16]) developed such an algorithm for computing  $E_s$  as well as a related energy magnitude  $M_e$  which agrees with  $M_w$  for  $E_s/M_0 = 2 \times 10^{-5}$ .  $E_s$  is computed by integrating squared velocity-proportional broadband records over the duration of the  $P$ -wave train, corrected for effects of geometrical spreading, frequency-dependent attenuation during wave propagation and source radiation pattern. According to [16], the radiated seismic energy may vary for a given seismic moment by two to three orders of magnitude. Further, it was found that a list of the largest events is dominated by earthquakes with thrust mechanisms when size is ranked by moment, but dominated by strike-slip earthquakes when ranked by radiated seismic energy. Choy and Kirby [18] gave a striking example for differences between

$M_e$  and  $M_w$  for two Chile earthquakes in 1997 which occurred in the same area but with different *source mechanisms*. One was interplate-thrust with  $M_w = 6.9$  and relatively low  $M_e = 6.1$ , whereas the other was intraslab-normal with  $M_w = 7.1$  and rather large  $M_e = 7.6$ . The first earthquake had a low potential to cause shaking damage and was felt only weakly in a few towns. In contrast, the second one caused widespread damage, land- and rock-slides, killed 300 people and injured 5000. Thus,  $M_w$ , although it theoretically does not saturate, may strongly underestimate or overestimate the size of an earthquake in terms of its potential to cause damage and casualties. Shaking damage is mainly controlled by the relative amount of released high-frequency energy at  $f > 0.1$  Hz which is better measured by  $M_e$ .

The quantity  $\tau_a = \mu E_s/M_0$  is termed apparent stress [90]. It represents the dynamic component of stress acting on the fault during slip, which is responsible for the generation of radiated kinetic seismic wave energy  $E_s$ . On average it holds that  $\tau_a \approx 2\Delta\sigma$  (with  $\Delta\sigma$  = stress drop = difference between the stress in the source area before and after the earthquake rupture). Both  $\tau_a$  and  $\Delta\sigma$  depend strongly on the seismotectonic environment, i. e., the geologic-tectonic conditions, fault maturity and type of earthquake *source mechanisms* prevailing in seismically active regions [16,17,18,19]. However,  $M_e \approx M_w$  holds only for  $\tau_a \approx 0.6$  MPa.

Another important teleseismic magnitude is called mantle magnitude  $M_m$ . It uses surface waves with periods between about 60 s and 410 s that penetrate into the Earth’s mantle. The concept has been introduced by Brune and Engen [13] and further developed by Okal and Talandier [64,65].  $M_m$  is firmly related to the seismic moment  $M_0$ . Best results are achieved for  $M_w > 6$  at distances  $> 15\text{--}20^\circ$  although the  $M_m$  procedure has been tested down to distances of  $1.5^\circ$  [77]. However, at  $D < 3^\circ$  the seismic sensors may be saturated in the case of big events. Also, at short distances one may not record the very long periods required for unsaturated magnitude estimates of very strong earthquakes, and for  $M_w < 6$ , the records may become too noisy at very long-periods. A signal-to-noise ratio larger than 3 is recommended for reliable magnitude estimates.  $M_m$  determinations have been automated at the PTWC and the CPPT [39,85] so that estimates are available in near real-time within about 10 min after OT from near stations, however typically within about half an hour, plus another few minutes for great earthquakes measured at the longest periods. Since  $M_m$  is determined at variable very long periods this magnitude does not – or only marginally – saturate even for very great, slow or complex earthquakes.

### Common Magnitude Estimates for the Sumatra 2004 $M_w$ 9.3 Earthquake

On 26 December 2004, the great Sumatra–Andaman Island earthquake with a rupture length of more than 1000 km occurred. It caused extensive damage in Northern Sumatra due to strong earthquake shaking. Moreover, it generated an Indian Ocean-wide tsunami with maximum run-up heights of more than 10 m. In total, this event claimed more than 200,000 victims and caused widespread damage on the shores of Sumatra, Thailand, India and Sri Lanka that were reached by the tsunami wave within some 15 min to about two hour’s time. This earthquake put the current procedures for magnitude determination to a hard test both in terms of the reliability and compatibility of calculated values and the timeliness of their availability to guide early warning and disaster management activities. Here we address only seismological aspects, not the additional major problems of inadequate global monitoring and insufficient regional communication and disaster response infrastructure. The earliest magnitudes reported by or made available to the Pacific Tsunami Warning Center (PTWC) were:

- $m_b > 7$ , about 8 min after origin time (OT);
- $M_{wp} = 8.0$ , available after some 12 minutes at the PTWC (including a magnitude-dependent correction [86]);
- somewhat later in Japan  $M_{wp} = 8.2$  after magnitude-dependent correction [48];
- $M_m \geq 8.5$  at the PTWC about 45 min after OT, hours later upgraded to  $M_m = 8.9$  by using mantle surface waves with longer periods ( $T \approx 410$  s);
- a first surface-wave magnitude estimate  $M_S = 8.5$ , some 65 min after OT;
- $M_w = 8.9$  (later revised to 9.0) released by Harvard Seismology more than 6 h after OT.

Other available measurements were:  $m_b = 5.7$  and  $M_S = 8.3$  by the IDC of the CTBTO,  $m_b = 7.0$ ,  $M_S = 8.8$ ,  $M_e = 8.5$  and another long-period  $P$ -wave based  $M_w = 8.2$  by the NEIC. All these values were too small and mostly available only after several hours or days (e. g., IDC data). Weeks later, after the analysis of Earth’s *fundamental modes* with periods up to 54 min and wavelength of several 1000 km, the now generally accepted value  $M_w = 9.3$  was published [76]. Why were the other magnitude values all too low and/or too late?:

- $m_b$  NEIC suffers from the combined effect of both spectral and time-window dependent saturation that we will discuss in more detail in Sect. “Magnitude Saturation and Biases Due to Earthquake Complexity”;

- $m_b$  IDC is even more affected by these saturation effects, because of the very short measurement time window of only 5 s after the first  $P$ -wave onset. In the case of the Sumatra 2004 earthquake, the first  $P$ -wave maximum occurred after some 80 s and another, with comparable amplitude, after about 330 s (cf. Fig. 8). Further, prior to  $m_b$  measurement, the IDC broadband data are filtered with a more narrow-band response peaked at even higher frequencies (3–4 Hz) than at NEIC ( $\approx 2.5$  Hz) [11];
- The reported surface-wave magnitudes ranged between  $M_S = 8.3$  (IDC), 8.8 (NEIC and Japan Meteorological Agency) and 8.9 (Beijing), i. e., some of them are close to the moment magnitudes. However, because of the late arrival of long-period teleseismic surface waves, good estimates are usually not available within 1–2 h after OT. This leaves a sufficient tsunami warning lead time only for shores more than 1000–2000 km away from the source.
- The NEIC  $P$ -wave moment magnitude  $M_w = 8.2$  was too small because its procedure is, similar as for  $M_{wp}$  determinations, based on relatively short-period (typically  $T < 25$  s)  $P$ -wave recordings and a single-source model (cf. Sect. “Magnitude Saturation and Biases Due to Earthquake Complexity”).
- The preliminary  $M_e = 8.5$ , computed a few hours after the December 2004 earthquake agreed with the final  $M_e$  computed later using a more robust method [20]. Another algorithm simulating a near-real-time computation would have yielded  $M_e = 8.3$ . Yet  $M_e$ , by its very nature as an energy magnitude and because of the relation  $E_s = \Delta\sigma/2\mu M_e$ , will generally be smaller than  $M_w$  for slow, long duration earthquakes with low stress drop. This is often the case for shallow thrust earthquakes in subduction zones. Extreme examples are four well-known slow tsunami earthquakes of 1992 (Nicaragua;  $M_w = 7.6$ ,  $\Delta M_e = -0.9$ ), 1994 (Java;  $M_w = 7.8$ ,  $\Delta M_e = -1.3$ ), 2000 (New Britain Region;  $\Delta M_w = 7.8$ ,  $\Delta M_e = -1.0$ ) and 2006 (Java;  $M_w = 7.7$ ,  $\Delta M_e = -0.9$ ) [40].

### Magnitude Saturation and Biases Due to Earthquake Complexity

Currently, the most common magnitude scales, especially those based on band-limited short-period data, still suffer *saturation*, e. g., the magnitudes  $m_b$ ,  $M_L$ ,  $m_B$  and  $M_S$ , which are typically measured at periods around 1 s, 2 s, 5–15 s and 20 s, respectively begin to saturate for moment magnitudes  $M_w$  larger than about 5.5, 6.5, 7.5 and 8.0. Earthquakes with  $m_b > 6.5$ ,  $M_L > 7.0$ ,  $m_B > 8.0$  and



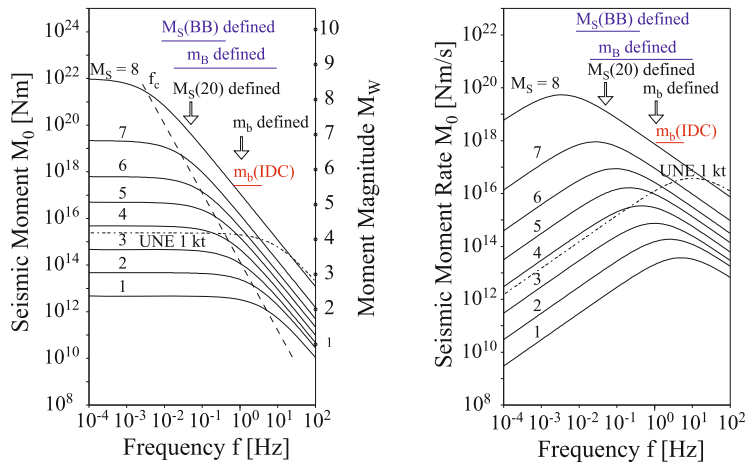
$M_S > 8.5$  are rare been found due to saturation (cf. Fig. 4 and [44]). Magnitude saturation has two causes: spectral saturation and saturation due to insufficient time-window length for the amplitude measurements. Source complexity may cause additional biases between different magnitude scales.

**Spectral Saturation of Magnitudes**

Spectral saturation occurs when the magnitude-dependent *corner frequency*  $f_c$  (for energy-related magnitudes) or the low-frequency plateau of displacement amplitudes (for moment magnitude) fall outside of the passband range of the seismographs *transfer function* or magnification curve (Fig. 3) of the recording seismograph or of the filter applied to broadband data before magnitude measurements are made. The reason for spectral saturation can best be explained by way of idealized average “source spectra” of ground displacement  $u(f)$  and ground velocity  $v(f)$  that have been corrected for the instrument response, for the decay of the wave amplitudes due to attenuation that is caused by internal friction and scattering of the seismic

waves at heterogeneities of the Earth and for amplification effects at the surface of the receiver site. For better understanding such source spectra have been multiplied in Fig. 5 by the factor  $4\pi r \rho v_{p,s}^3 / R_{\theta,\phi}^{p,s}$  given in Eq. (12) in order to get for the displacement amplitudes  $u_0 = \text{constant}$  at  $f < f_c$  the related scalar seismic moment  $M_0$  (Fig. 5, left) and its time-derivative, the so-called moment rate (Fig. 5, right).

The shape of the source spectrum can be interpreted as follows: The critical wavelength, which corresponds to  $f_c$ , is  $\lambda_c = v_{p,s} / f_c = c_{m1} \pi R = c_{m2} (L \times W)^{1/2}$  with  $v_{p,s}$  - velocity of the P or S waves in the source region, depending on whether  $f_c$  relates to a P-wave or an S-wave spectrum, R-radius of a circular fault rupture model, L- length and W- width of a rectangular fault rupture model;  $c_{m1}$  and  $c_{m2}$  are model dependent constants. For very long fault ruptures, i. e.,  $L \gg W$ , one can even write  $\lambda_c = c_{m3} L$ . Thus,  $\lambda_c$  is proportional to the linear dimension of the fault. For  $f < f_c$ ,  $\lambda_c$  becomes larger than the fault. Rupture details along the fault can then no longer be resolved and the fault is “seen” by these long wavelengths just as a point source. Therefore, all frequencies  $f < f_c$  have the



Earthquake Magnitude, Figure 5

“Source spectra” of ground displacement (left) and velocity (right) for an average single rupture seismic shear source, scaled on the left ordinates to seismic moment  $M_0$  (left diagram) and moment rate (right diagram), respectively. The black spectral lines have been scaled according to Aki [3] to integer surface-wave magnitudes  $M_S$  between 1 and 8. For reference the respective integer moment magnitude values  $M_w$  between 1 and 10, calculated according to Eq. (11), have been marked with equidistant dots on the right-side ordinate of the left diagram. The broken line shows the increase of the corner frequency  $f_c$  with decreasing seismic moment of the event, the dotted curve gives the approximate “source spectrum” for a well contained underground nuclear explosion (UNE) of an equivalent yield of 1 kt TNT. Note the plateau in the displacement spectrum towards low frequencies (corresponding to  $u_0 = \text{constant}$  for  $f < f_c$ ), from which  $M_0$  is determined according to Eq. (11) when using the frequency-domain approach. For  $f > f_c$  the amplitudes decay  $\sim f^{-2}$ . The open arrows point to the center frequencies on the abscissa at which the 1 Hz body-wave magnitude  $m_b$  and the 20 s surface-wave magnitude  $M_S(20)$ , respectively, are determined and the blue horizontal interval bars mark the range of frequencies within which the maximum P-wave and Rayleigh-wave amplitudes for  $m_b$  and  $M_S(BB)$  should be measured according to the new IASPEI standards [37]. In contrast, the red bar marks the frequency range of maximum velocity-proportional magnification of the bandpass filter between 1 Hz and 4 Hz which is used for  $m_b$  determination at the IDC.

same displacement amplitudes. Accordingly,  $M_0$ , which is proportional to the fault area and the average slip over the fault, has to be determined either in the spectral domain from the low-frequency asymptote  $u_0$  to the displacement spectrum or in the time domain by fitting synthetic long-period waves with  $f < f_c$  to observed ones that have been low-pass filtered in the same frequency range.

For radiated frequencies  $f > f_c$  with  $\lambda < \lambda_c$ , the shape of the spectrum changes drastically. Related displacement amplitudes are then excited by successively smaller patches of the rupture plane. The area of the rupture elements decreases with the second order of their linear dimension. Accordingly, the generated displacement amplitudes are  $A_d \sim f^{-2}$ , while the related velocity amplitudes  $A_v = A_d 2\pi f$  decay only  $\sim f^{-1}$ . In the seismological literature this is usually called the  $\omega^{-2}$  rupture model [3], based on the concept of similarity, which implies a constant stress drop independent of source size. More complicated rupture models yield a high-frequency amplitude decay  $\sim \omega^{-3}$  [33,34] and even more rapid decays have sometimes been found in empirical data (up to 5th order). Steeper than  $\omega^{-2}$  amplitude decay would further amplify the spectral saturation of magnitude data discussed below.

The Harvard standard procedure for  $M_0$  determination assumes a single point source model with a prescribed, triangular moment-rate function in the time domain (as an approximation to moment-rate curves such the ones shown in Fig. 7) as well a minimum period of 200 s for strong earthquakes with magnitudes  $> 8$ . Assuming an average rupture velocity of 2.5 km/s, this period would correspond to a wavelength of 500 km. This is much shorter than the total rupture length of more than 1100 km for the great Sumatra 2004 earthquake and explains why  $M_w(\text{HRV}) = 9.0$  was smaller than the moment magnitude  $M_w = 9.3$  determined by using fundamental Earth's modes with periods of 1000 s and more [76].

The relationship between the two currently most common magnitudes,  $m_b$  and  $M_S(20)$ , can be understood with reference to Fig. 5.  $m_b$  is measured in the period range  $0.5 < T < 3$  s, typically around 1 s. This corresponds approximately to the *corner frequencies* of earthquakes with  $M_S \approx 3$  to 4.5. According to Utsu [83] this is equivalent to an  $m_b$  between about 3.5 and 5.0. For  $M_S < 4.5$  or  $m_b < 5$ ,  $m_b$  is thus likely to be determined from amplitude measurements near or below the *corner frequency* of the source spectrum. In that case  $m_b$  is a good measure of seismic moment. However, for larger magnitudes  $m_b$  samples spectral amplitudes well above  $f_c$ , resulting in systematically too small  $m_b$  values as compared to  $M_S$  and  $M_w$ . For great earthquakes this difference may reach 2 m.u. (Fig. 4).

In contrast,  $M_S(20)$  is measured at periods around 20 s and thus saturates much later at values between about 8.5 to 9.

However, these arguments only hold on average. The stress drop  $\Delta\sigma$  of individual events may vary by about 2 to 3 orders, as apparent stress  $\tau_a$ , especially for earthquakes with  $M_w < 7.5$  [16,17]. According to the relation  $M_0 = (16/7)\Delta\sigma R^3$  given by Keilis-Borok [52] this may change source radii  $R$  and associated  $f_c$  by about one order. As an example, the dotted curve in Fig. 5 shows the approximate seismic source spectrum for a well contained underground nuclear explosion (UNE) of an equivalent yield of 1 kt TNT which corresponds to a magnitude  $m_b \approx 4$ . Its source volume is much smaller than that of an earthquake with same seismic moment. Hence the *corner frequency* of its source spectrum is not around 1 Hz but around 10 Hz. This is the reason why  $m_b$  determined from UNE records does not saturate, even for the strongest UNE ever tested with  $m_b \approx 7$ . Moreover, Fig. 5 also illustrates that an earthquake and an UNE with seismic moment around  $4 \times 10^{15}$  Nm and  $M_w \approx 4$  have different maximum seismic moment-rate release at about  $4 \times 10^{15}$  and  $4 \times 10^{16}$  Nm/s, respectively. The latter corresponds to 100 times higher seismic energy release or to an energy magnitude  $M_e$  that is 1.3 m.u. larger. Large differences have also been observed amongst earthquakes, e. g., the Balleny Island earthquake of 25.03.1998 had  $M_w(\text{HRV}) = 8.1$  and  $M_e(\text{NEIC}) = 8.8$ . The opposite will happen in the case of low stress drop earthquakes propagating with very low rupture velocity [38]. The Java tsunami earthquake of 17 July 2006 was a striking example with  $M_e = 6.8$ ,  $m_B = 7.0$  and  $M_w = 7.7$ .

Similar observations had already been made in the 1970s when comparing  $m_b$  and  $M_S$  values of identical events. This prompted the Russian scientist Prozorov to propose a “creepex” parameter  $c = M_S - a \times m_b$  (with  $a = \text{constant}$  to be determined empirically for different source types and stress drop conditions). It aims at discriminating between normal, very slow (creeping) and explosion-like (fast rupture, high stress drop) earthquakes. World-wide determination of this parameter for earthquakes in different regions revealed interesting relations of  $c$  to source-geometry and tectonic origin [51]. Similar systematic regional differences were also reported for  $M_S - M_w$  [24,67] and  $M_e - M_w$  [16,19], suggesting systematic regional differences in stress drop.

### Magnitude Saturation Due to Insufficient Time-Window Length for Amplitude Measurement

The second reason for magnitude saturation is insufficient time-window length for measuring  $(A/T)_{\text{max}}$  in seismic

records. It is most relevant when determining body-wave magnitudes, but it has been a subject of controversy, misconceptions and disregard of earlier recommendations for decades. The reason is that in teleseismic seismograms the  $P$ -wave group does not always appear sufficiently well separated in time from later phase arrivals such as the depth phases  $pP$  and  $sP$ . These do not directly travel from the seismic source at depth  $h$  to the recording station but travel first to the Earth's surface above the source and from there, after reflection or conversion from  $S$  to  $P$ , propagate back into the Earth. Depending on  $h$ , which may vary from a few kilometers up to 700 km, and the type of depth phase recorded, they may arrive from a few seconds up to about 4.5 min after the onset of direct  $P$ . Depending on the radiation pattern of the *source mechanism*, some stations may even record the depth phases with larger amplitudes than the direct  $P$  wave. This is one of the concerns that led many researchers to propose measuring the  $P$ -wave amplitudes for magnitude measurements within a short time window after the  $P$  onset. On average, however, the depth phases have smaller amplitudes than  $P$  and will not bias  $m_b$  estimates at all. If, however, a seismic station is situated near to the nodal line of the so-called focal sphere, corresponding to strongly reduced  $P$ -wave radiation in these directions, the amplitude of the depth phase is a better estimator for the body-wave energy radiated by this seismic source and thus of its corresponding magnitude.

Two or three more phases of longitudinal waves may arrive close to the direct  $P$  at teleseismic distances between  $20^\circ$  and  $100^\circ$ . These include  $PcP$ , which results from  $P$ -wave energy reflected back from the surface of the Earth's core at 2900 km depth, and the phases  $PP$  and  $PPP$ , which are  $P$  waves that have been reflected back from the Earth's surface once at half-way or twice at 1/3- and 2/3-way between the seismic source and the recording station, respectively. However, in short-period records the amplitudes of  $PP$  and  $PPP$  are generally smaller and those of  $PcP$  even much smaller than the amplitudes of direct  $P$  waves. These later arrivals will therefore, never bias  $m_b$  estimates. Yet on broadband records  $PP$  may sometimes have equal or even slightly larger amplitudes than primary  $P$ . However,  $P$  and  $PP$  phases are usually well separated by more than 1 min (up to 4 min) and not likely misinterpreted. Only for rare large earthquakes with  $M > 7.5$  the rupture duration and related  $P$ -wave radiation may extend into the time window where  $PP$  should arrive. But even then, wrongly taking  $PP_{\max}$  for  $P_{\max}$ , the bias in  $m_b$  estimate will not exceed 0.2 m.u. and usually be much smaller.

This experience from extensive seismogram analysis practice led Bormann and Khalturin [7] to state in 1974:

... "that the extension of the time interval for the measurement of  $(A/T)_{\max}$  up to 15 or 25 sec., as proposed ... in the Report of the first meeting of the IASPEI Commission on Practice (1972) ... is not sufficient in all practical cases, especially not for the strongest earthquakes with  $M > 7.5$  ...".

This was taken into account in the Manual of Seismological Observatory Practice edited by Willmore [87]. It includes the recommendation to extend the measurement time window for  $P$ -wave magnitudes up to 60 s for very large earthquakes. But still, this has not yet become common practice (see Sect. "Introduction to Common Magnitude Scales: Potential and Limitations") although even a limit of 60 s may not be sufficient for extreme events such as the Sumatra  $M_w$  9.3 earthquake when the first  $P1_{\max}$  appeared around 80 s and a second  $P2_{\max}$  of comparable amplitude at about 330 s after the first  $P$ -wave onset (cf. Fig. 8).

To allow a quick rough estimate of earthquake rupture duration  $\tau_d$  as a function of magnitude we derived from extrapolation of data published in [66] the average relation

$$\log \tau_d \approx 0.6M - 2.8. \quad (13)$$

It yields for  $M = 6, 7, 8$  and  $9$   $\tau_d \approx 6$  s, 25 s, 100 s and 400 s, respectively. Measurement time windows of 5 s, 25 s or 60 s may therefore underestimate the magnitude of earthquakes with  $M_w > 6, > 7$  or  $> 8$ , respectively. We call this effect the time-window component of magnitude saturation. It aggravates the pure spectral saturation component. To avoid this in future, the new IASPEI standards of amplitude measurements for  $m_b$  and  $m_B$  (cf. [42]) recommend to measure  $(A/T)_{\max} = A_{v\max}/2\pi$  in the entire  $P$ -phase train (time span including  $P$ ,  $pP$ ,  $sP$ , and possibly  $PcP$  and their codas but ending preferably before  $PP$ ).

In fact the pioneers of the magnitude scales, Richter and Gutenberg, knew this, because they were still very familiar with the daily analysis of real seismic records and their complexity. Regrettably, they never wrote this down, with respect to magnitude measurements, in detail for easy reference. In the current era of automation and scientific depreciation of alleged "routine processes" the younger generation of seismologists usually had no chance to gather this experience themselves and occasionally introduced technologically comfortable but seismologically questionable practices. In an interview given in 1980 [75] Prof. Richter remembered that Gutenberg favored the body-wave scale in preference to the surface-wave scale because it is theoretically better founded. However, he said:

“... it gives results comparable with Gutenberg’s only if his procedure is closely followed. Experience has shown that misunderstanding and oversimplified misapplications can occur. For instance, magnitude is sometimes assigned on the first few waves of the  $P$  group rather than the largest  $P$  waves as Gutenberg did.”

In order to avoid too-short measurement time windows when searching for the largest  $P$  amplitude one can estimate the rupture duration independently from the duration of large  $P$ -wave amplitudes in high-frequency filtered BB records because the generation of high-frequency waves is directly related to the propagation of the rupture front. Thus one may find  $P_{\max}$  for great earthquakes even beyond the theoretically expected  $PP$  arrival (cf. Fig. 8).

### Magnitude Biases Due to Neglecting Multiple Source Complexity

Realizing that strong earthquakes usually consist of multiple ruptures Bormann and Khalturin [7] also wrote:

*“In such cases we should determine the onset times and magnitudes of all clear successive  $P$ -wave onsets separately, as they give a first rough impression of the temporal and energetic development of the complex rupture process. ... The magnitude  $MP = \log \sum_n (A_i/T_i) + Q(D, h)$  ( $n$  is the number of successive  $P$ -wave onsets) could be considered as a more realistic measure of the  $P$ -wave energy released by such a multiple seismic event than the  $m_b$ -values from ... (single amplitude)  $(A/T)_{\max}$  within the first five half cycles or within the whole  $P$ -wave group.”*

This magnitude, which is based on summed amplitudes in broadband records, is now called  $m_{bc}$  [10], which stands for cumulative body-wave magnitude.

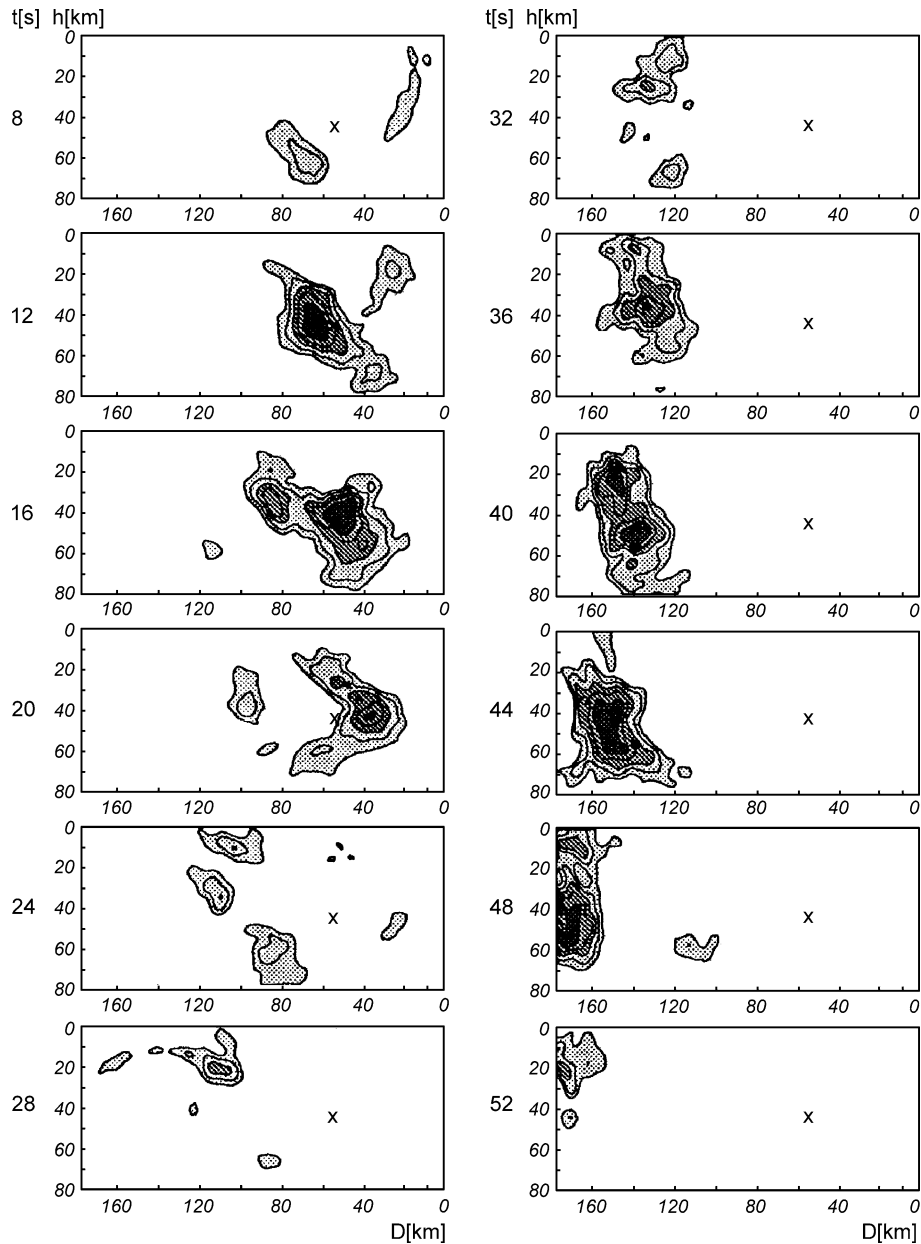
The 1985  $M_w = 8.1$  Mexico earthquake was a striking example for the development of such a multiple rupture process in space and time ([58], Fig. 6). A dense network of high-frequency strong-motion recordings near to the source revealed that the earthquake had a total rupture duration of about 60 s and consisted of two main sub-ruptures with significantly increased slip-velocities (12–32 cm/s). These two fault segments were separated in space by roughly 100 km in strike direction and ruptured between 10–22 s and 34–50 s after rupture start. Such a complicated rupture is not well represented by calculating the average slip and rupture velocity for a single point-source model. Also the *corner frequencies* related to these smaller sub-ruptures will be higher and not correspond to  $(L \times W)^{-1/2}$  of the total rupture area.

Such multiple ruptures are not an exception but rather the rule for earthquakes with magnitudes above 7.5 (and often also for smaller ones, even down to events with magnitudes around 5.0). The detailed patterns of the respective moment-rate curves differ from event to event (Fig. 7). Often they can not be approximated by a single-source triangular moment-rate function, as commonly assumed in the standard procedure for moment tensor solutions practiced at Harvard [78] and other centers.

Therefore, the Harvard group [78] re-analyzed the data of the great Sumatra 2004 earthquake for which  $M_w = 9.0$  had been calculated with the standard procedure. Interactively fitting synthetic records for five successive point sources to the observed mantle surface-wave data in the 200–500 s period range yielded the same value of  $M_w = 9.3$  as derived by [76] for a single-source model but using much longer periods between 20 min and 54 min. In fact, the multiple Centroid Moment Tensor (CMT) source analysis applied in [78] resembles the concept proposed in [7] for  $P$ -wave magnitudes of strong earthquakes, but applied to long-period surface waves. Presently, a multiple CMT source analysis still requires human interaction and takes too much time for early warning applications. Possible alternative procedures such as the automatic line source inversion [21] have been developed but demonstrated so far only for earthquakes with magnitudes  $< 7$  for which classical  $m_b$ ,  $M_S$  or  $M_w$  do not saturate due to source complexity.

### Proposals for Faster Magnitude Estimates of Strong Earthquakes

Soon after the great Sumatra earthquake of 2004 several authors suggested improvements to obtain more reliable and faster magnitude estimates of strong earthquakes. Menke and Levin [59] proposed to use a representative selection of 25 globally distributed high quality stations of the IRIS (Incorporated Research Institutions for Seismology) Global Seismic Network as a reference data base of available strong long-period master-event records with known  $M_w$ . In case of a new strong earthquake, a search for the nearest (within a few hundred kilometers) reference event in the data base is performed and waveforms are compared for a time window of about 30 min. By adding the  $\log_{10}$  of the average amplitude ratio of the two events to the  $M_w$  of the master event, a moment magnitude estimate of the actual event is obtained. This procedure is based on the assumption of similarity of *source mechanisms* and radiation patterns, slip rates and stress drops, at least within the reference regions. The authors expect reasonably good magnitude estimates, with only small un-



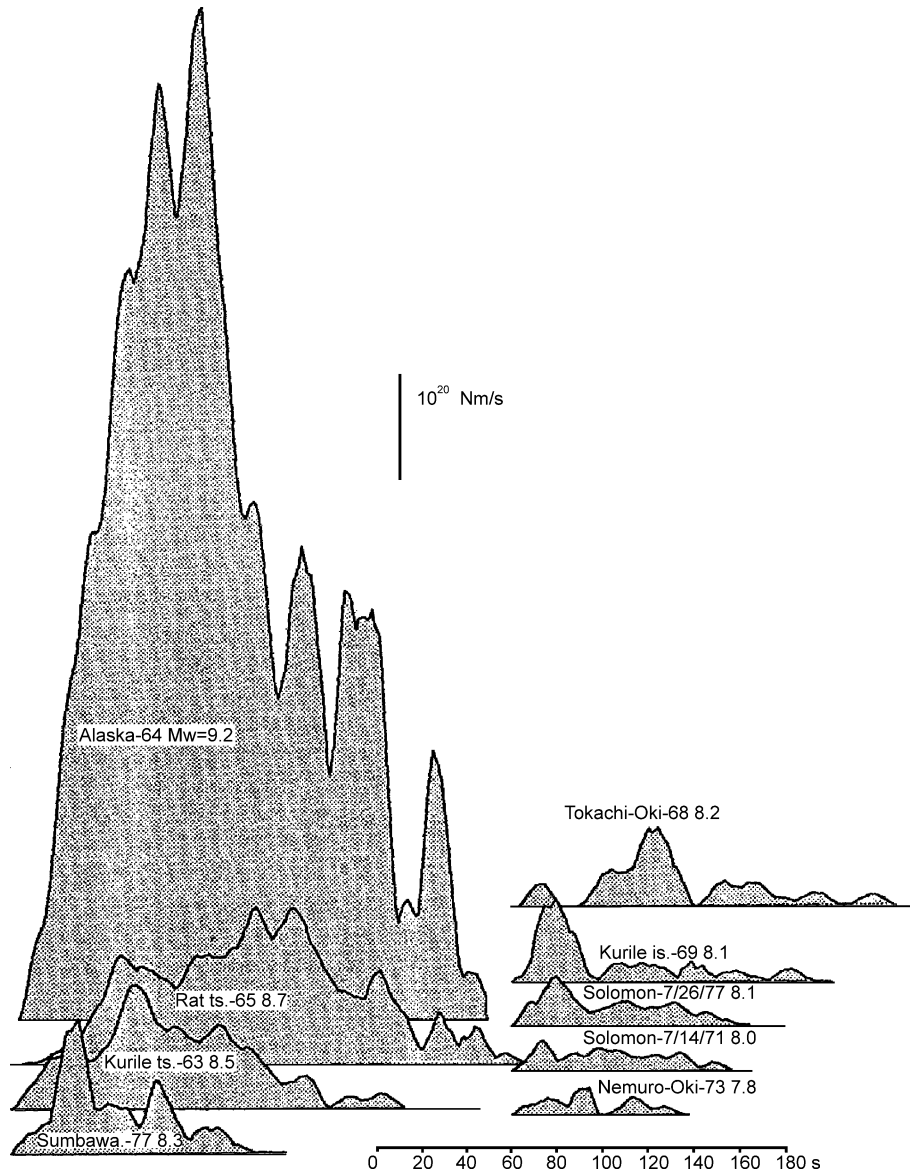
Earthquake Magnitude, Figure 6

Snapshots of the development in space and time of the inferred rupture process of the 1985 Michoacán, Mexico earthquake. The cross denotes the NEIC hypocenter position, the shading of the patches (from the outer part inwards *dotted*, *hatched* and *black*) relate to areas with velocities of dip slip (see *source mechanism*) in the ranges between 12 and 22 cm/s, 22 and 32 cm/s and greater than 32 cm/s. Redrawn and modified from Fig. 6 in [58]; taken from Fig. 3.8 in Vol. 1 of [6], © Seismological Society of America and IASPEI; with permission of the authors

derestimation for events with  $M_w > 8.6$ . Thus warnings could be issued within about 40 min after OT (measurement time window plus travel-time to stations of a global network). This would still be relevant for distant coasts that might be affected by a tsunami. However, no data have

been published until now that demonstrate the near-real-time operational capability of this procedure for a representative set of strong events.

Another approach by Lomax et al. [55,56,57] uses high-frequency seismograms ( $f \geq 1$  Hz) that contain pre-



Earthquake Magnitude, Figure 7

Moment-rate functions for the largest earthquakes in the 1960 and 1970s (modified from Fig. 9, p. 1868 in [53]), taken from Fig. 3.7 in Vol. 1 of [6], © Seismological Society of America and IASPEI; with permission of the authors

dominantly  $P$  signals radiated directly from the propagating rupture front and show little interference with later secondary waves such as  $PP$  or  $S$ , thus providing a direct estimate of the rupture duration. Such recordings are available at teleseismic distances ( $30^\circ$ – $90^\circ$ ) within about 20 min after OT, even after strong events with long durations and provide an early picture of the total rupture process. When assuming constant rupture velocity and mean slip for stronger and weaker earthquakes, the seismic moment  $M_0$  and thus moment magnitude  $M_w$  could be esti-

mated by comparing the actual rupture duration (averaged from observations at several seismic stations) with that of a reference event with known  $M_0$  and rupture duration. This is conceptually similar to the approach in [59] but with high-frequency observations and the ratio of rupture duration instead of amplitudes.

However, Hara [32] demonstrated with a large data set of strong earthquakes that it is difficult to estimate earthquake size reliably only from durations  $t$  of high-frequency radiation. Therefore, he measured duration  $t$  in combi-

nation with the maximum displacement amplitude  $A_{dmax}$  within this time interval and derived the following empirical relation:

$$M = 0.79 \log A_{dmax} + 0.83 \log D + 0.69 \log t + 6.47 \quad (14)$$

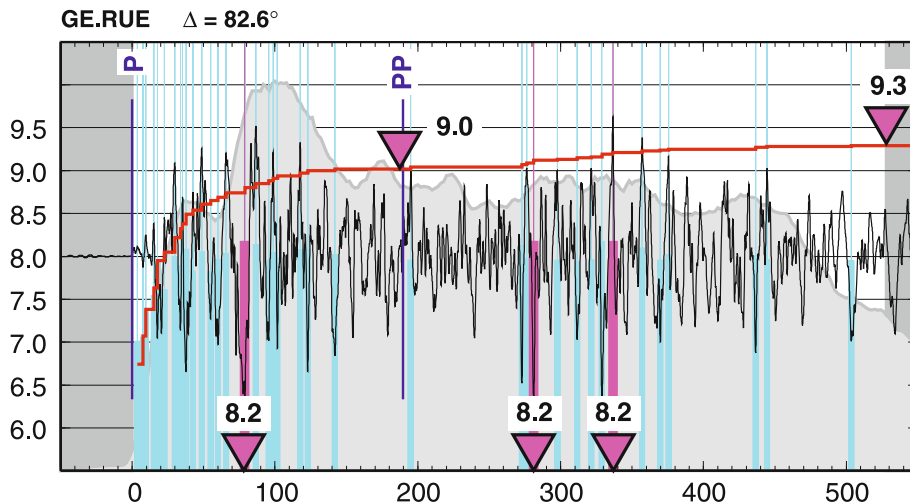
with  $A_{dmax}$ ,  $D$  and  $t$  in units of m, km and s, respectively. He applied Eq. (14) to 69 shallow earthquakes in the magnitude range  $7.2 \leq M_w(\text{HRV}) \leq 9.0$  at distances between  $30^\circ$  and  $85^\circ$  and on average got a 1:1 relation between  $M_w(\text{HRV})$  and his magnitude with a standard deviation of 0.18 m.u. All event estimates were within  $\pm 0.5$  m.u. of  $M_w(\text{HRV})$ , with the exception of the heavily underestimated Denali/Alaska earthquake of 3 November 2002 (7.1 instead of 7.8). This is a promising and simple procedure.

Bormann and Wylegalla [9] applied the earlier proposal in [7] to recordings with a velocity passband between 40 Hz and 125 s. They interactively summed up the maximum amplitudes of all visually discernible sub-ruptures in the recordings of several recent great earthquakes with  $M_w \geq 8.3$ , amongst them the tsunamigenic  $M_w = 9.3$  Sumatra earthquake of 2004. For the latter they obtained a cumulative broadband body-wave magnitude  $m_{Bc} = 9.3$  in records of just a single German station (RUE;  $D = 82.5^\circ$ ) at the time of the second major amplitude maximum, some 330 s after the first  $P$  onset and 18 min after OT. For three more events with magnitudes  $M_w$  8.3, 8.4 and 8.6 they calculated  $m_{Bc}$  values of 8.4,

8.4 and 8.6, respectively, i. e., excellent agreement. Subsequently, 50 more earthquakes in the magnitude range 6 to 9 were analyzed interactively [10] with the following results:

- Average difference  $m_B - M_w(\text{HRV}) = 0.00 \pm 0.27$  in the range  $6.0 \leq M_w(\text{HRV}) < 8$ . For magnitudes  $> 7.8$ –8, however,  $m_B$  tends to underestimate  $M_w$ , e. g.,  $m_B = 8.3$  for the Sumatra earthquake of 26 December 2004 based on the BB record of station RUE. Remarkably this  $m_B$  value is still very close to  $M_w$ ,  $M_{wp}$  and  $M_e$  of the NEIC, which ranged between 8.2 and 8.5.
- The average difference  $m_{Bc} - M_w(\text{HRV}) = +0.18 \pm 0.26$  in the range  $6.0 \leq M_w(\text{HRV}) \leq 9.0$ , i. e.,  $m_{Bc}$  has a tendency to slightly overestimate  $M_w(\text{HRV})$  on average, but not for  $M_w > 8$  (see the four values above).

In [10] also first results of a fully automatic determination of  $m_B$  and  $m_{Bc}$  have been presented. The algorithm has been improved by incorporating automatic estimates of the rupture duration calculated from the envelope of the high-frequency  $P$ -wave radiation from filtered broadband records of globally distributed stations in a wide range of azimuths and source distances. In the case of strong earthquakes with long rupture duration this justifies the search for broadband  $P_{max}$  even beyond the onset of  $PP$  and to sum-up the amplitudes of major sub-ruptures over the whole rupture duration as defined above. Figure 8 gives an



Earthquake Magnitude, Figure 8

Velocity broadband record at the Berlin station RUE in  $D^\circ = 82.6^\circ$  epicentral distance of the great  $M_w 9.3$  tsunamigenic Sumatra earthquake of Dec. 26, 2004. The record is projected into a time-magnitude diagram as plotted by the automatic  $m_B - m_{Bc}$  algorithm. The red inverted triangles mark the times and give the values of  $m_B$  for the three largest sub-ruptures. The red step curve shows the development of the cumulative magnitude  $m_{Bc}$  as a function of time. The inverted red triangles on this curve give the  $m_{Bc}$  before the onset of  $PP$  and at the end of the rupture process, about 530 s after the first  $P$ -wave onset, as estimated from the decay of the amplitude envelope of short-period filtered  $P$ -waves (see text)

example for a BB record of the Sumatra earthquake of 26 December 2004. The largest  $P$ -wave amplitudes at about 80 s, 280 s and 330 s after the  $P$  onset each yield a single amplitude  $m_B = 8.2$ , whereas the cumulative magnitude  $m_{BC} = 9.3$  is in perfect agreement with the best moment magnitude estimates for this event.

The automatic algorithm for  $m_B$  and  $m_{BC}$  determination has been in use since spring 2007 in the operational Indonesian prototype tsunami early warning system and yields online estimates of  $m_B$ . Before the implementation it had been tested whether the automatic procedure produces results that are comparable with those determined earlier interactively by two experienced seismogram analysts. Identical broadband records of 54 earthquakes in the magnitude range  $6 \leq M_w(\text{HRV}) \leq 9$  were used for this comparison based on 138  $m_B$  and 134  $m_{BC}$  values. The average difference between the interactively and automatically determined magnitudes was 0.03 and 0.02 m.u. with standard deviations of  $\pm 0.13$  and  $\pm 0.19$  m.u., respectively. This is in the range of other high-quality magnitude measurements. Even single station  $m_B$  and  $m_{BC}$  estimates differed on average  $< 0.08$  m.u. from average global network estimates based on up to hundreds of stations. Their standard deviations were  $< \pm 0.25$  m.u. and decreased to  $\pm 0.10$  m.u. for  $m_B$  and  $\pm 0.14$  m.u. for  $m_{BC}$  when just a few stations (between two and seven) were used to estimate the  $m_B$  and  $m_{BC}$  event magnitudes. This documents both the reliability of the automatic procedure as well as the reliability of  $m_B$  and  $m_{BC}$  estimates, even if derived from a few records of globally distributed high-fidelity stations. Thus, the automatic procedure is suitable for reproducibly determining the IASPEI recommended standard magnitude  $m_B$  and its proposed non-saturating extension  $m_{BC}$  in near real-time. When using only observations in the distance range  $21^\circ \leq D^\circ \leq 100^\circ$  saturation-free teleseismic magnitude estimates of earthquakes with potential for strong shaking damage and tsunami generation could be made available in near real-time within about 4 to 18 min after OT, depending on epicentral distance and rupture duration.

Compared to other more theoretically based methods such as  $M_{wp}$  and  $M_w$ , the empirical  $m_B - m_{BC}$  method is free of any hypothesis or model assumptions about the rupture process (single or multiple source), type of rupture mechanism, rupture velocity, average slip and/or stress drop, complexity or simplicity of the moment-release function, etc. It just measures velocity amplitudes on the unfiltered broadband record, complex or not, sums them up over the duration of the rupture process and calibrates them with the classical empirical broadband  $Q_{pV}$  function (Fig. 2 and [30]). However, one has to consider

that – in contrast to all types of moment magnitudes –  $m_B$  and  $m_{BC}$  are not determined from the maximum long-period displacement amplitudes, but from the maximum velocity amplitudes. Therefore,  $m_B$  (for earthquakes with  $M_w < 8.0$ ) and  $m_{BC}$  (for earthquakes with  $M_w > 7.8$ ) are better estimators than  $M_w$  for the seismic energy released by the earthquake and thus of its shaking-damage potential. Figure 9 compares  $m_B$  and  $m_{BC}$  with  $M_w(\text{HRV})$  for 76 earthquakes in the range  $6 \leq M_w \leq 9.3$ . The respective standard regression relations are:

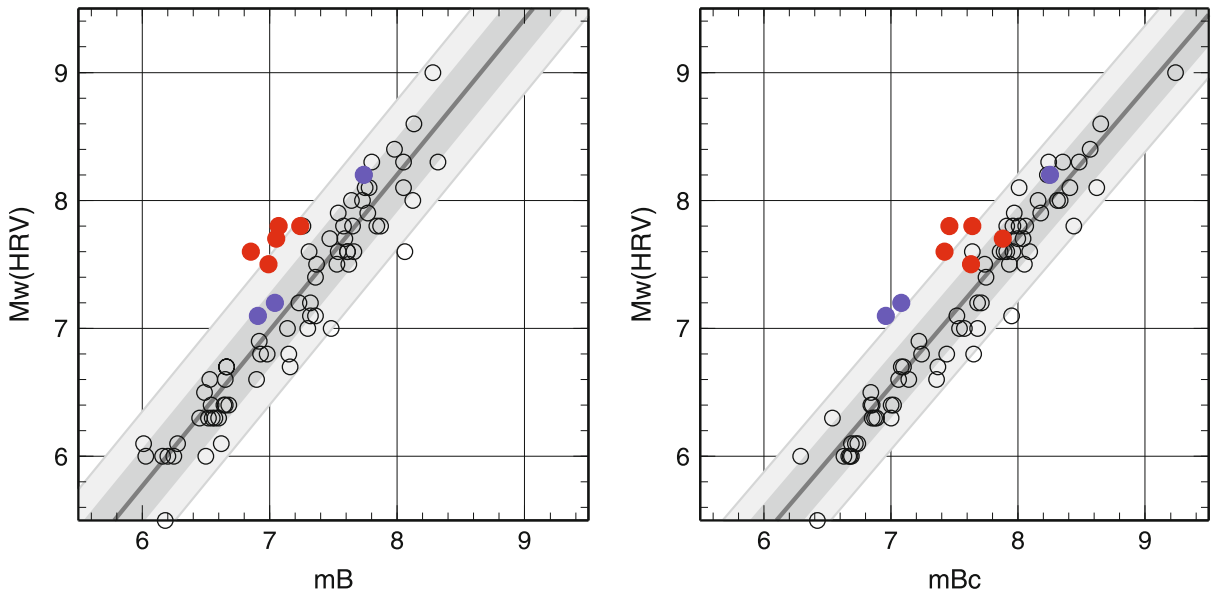
$$M_w(\text{HRV}) = 1.22 m_B - 1.54 \pm 0.29 \quad (15)$$

and

$$M_w(\text{HRV}) = 1.16 m_{BC} - 1.59 \pm 0.25 \quad (16)$$

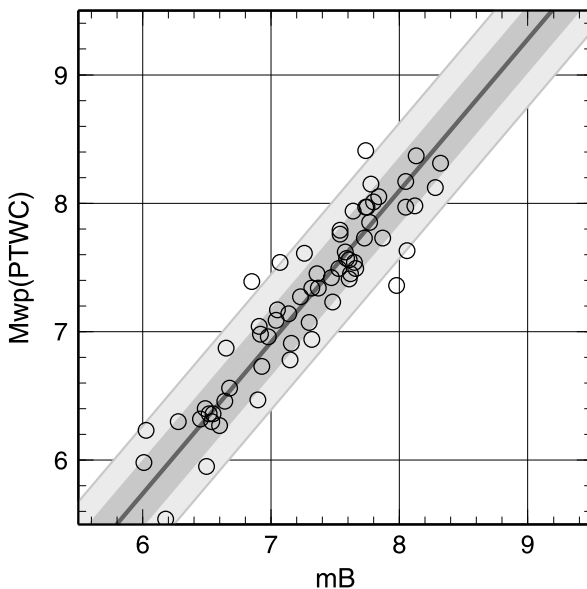
These scaling relations allow much faster estimates of  $M_w$  than current routine standard  $M_w$  procedures. The rough moment estimates derived from  $m_B$  and  $m_{BC}$  data,  $M_w(m_B)$  or  $M_w(m_{BC})$ , are sufficiently reliable for initial earthquake and tsunami alarms with standard deviations of the  $M_w$  estimates of about  $\pm 0.29$  and  $\pm 0.25$  m.u., respectively. However, looking into details of the somewhat irregular data scatter one realizes how *seismic source* complexity may “spoil” such regression relations. The five data points marked red in Fig. 9(left and right) are distinct outliers in the left diagram, i. e., the respective  $m_B$  values are 0.5 to 0.75 m.u. smaller than  $M_w(\text{HRV})$  although usually  $m_B$  scales rather well with  $M_w(\text{HRV})$  between  $6.5 < m_B < 8.0$ . These points correspond to slow earthquakes, one in Peru (1996) and four are tsunami earthquakes as mentioned at the end of Sect. “Common Magnitude Estimates for the Sumatra 2004  $M_w$  9.3 Earthquake”. Their rupture durations ranged from about 100 s to 200 s, i.e, according to relationship (13) about 2–3 times longer than expected on average for their  $M_w$  value. Both  $m_B$  and  $M_e$  are usually much smaller than  $M_w$  for such events. In contrast, when calculating  $m_{BC}$ , then these five data points all move close to the (not marked) 1:1 line in the  $m_{BC}-M_w(\text{HRV})$  diagram Fig. 9(right). Thus  $m_{BC}$  becomes a good direct estimator of  $M_w$  for typical slow earthquakes, much better than via relation (16), which compensates for the usually too large  $m_{BC}$  values of shallow depth and “normal” rupture earthquakes with  $M_w < 8$ . Thus, by determining rupture duration independently and treating very slow events separately, the standard deviation in relations (15) and (16) can be reduced. Moreover, the blue dots in Fig. 9 belong to very deep earthquakes with  $h = 525$  km, 583 km and 631 km, respectively. Such deep earthquakes are “explosion-like” with comparably short rupture durations. Both  $m_B$  (for  $M_w < 8$ ) and  $m_{BC}$  yield values very close to





Earthquake Magnitude, Figure 9

Standard regression relationships of  $M_w(\text{HRV})$  over  $m_B$  (left) and  $m_{Bc}$  (right). Red dots correspond to very slow earthquakes (Nicaragua 1992, Java 1994, New Britain Region 2000, Peru 2001 and Java 2006) and the blue dots belong to very deep earthquakes (Bolivia 1994, Philippines 2005 and Fiji Island 2006) with source depths  $h = 631$  km, 525 km and 583 km, respectively. The gray band and the two white bands around the average straight line correspond to the width of one and two standard deviations in  $y$ -direction



Earthquake Magnitude, Figure 10

Standard regression of  $M_{wp}(\text{PTWC})$  over  $m_B$ . The standard deviations in  $y$ -direction are marked as in Fig. 9. The  $M_{wp}$  data have been kindly provided by the PTWC (courtesy of B. Hirshorn)

$M_w$ . In the  $m_{Bc}-M_w$  diagram, which is dominated by shallow and normal rupture earthquakes, such deep events ap-

pear as outliers. However, rapid event locations with good depth estimates allow one to identify such events and  $m_{Bc}$  (or  $m_B$ ) should then be taken directly as estimator of  $M_w$  and not via relation (16).

$M_{wp}$  has so far been the fastest operationally determined estimator of  $M_w$ . Comparably fast automatic  $m_B$  determination is now implemented, complementary to  $M_{wp}$ , in the German Indonesian Tsunami Early Warning System (GITEWS). Figure 10 compares the relation between  $m_B$  and  $M_{wp}$  for our test data set. These two magnitudes scale almost 1:1, following the standard regression relation:

$$M_{wp} = 1.08 m_B - 0.638 \pm 0.24. \quad (17)$$

### Future Requirements and Developments

Few national seismological data centers and stations report amplitude, period and/or magnitude data to the international data centers. The main reason is usually the lack of manpower to make competent measurements of these parameters interactively for the large amount of data recorded nowadays. Instrument responses of the seismographs used are sometimes not known accurately enough. There is, however, a growing practical and research need for such parameter data that have been determined ac-

ording to international standards. Therefore, the most urgent requirements in the field of magnitudes are:

- Training of station and network operators to understand and practice proper magnitude measurements, instrument calibration and control;
- Implementation of the IASPEI magnitude standards [42];
- Making the tested and calibrated automatic algorithms available worldwide to data producers so that lack of manpower is no longer a hindrance to mass-produce such data;
- Use of such standardized mass data with significantly reduced procedure-dependent errors for improved research into the attenuation properties of the Earth and deriving better magnitude calibration functions for all distance ranges;
- Comparison of magnitude data derived from identical record sets by applying both traditional and new standard measurement procedures and to derive standardized conversion relationships. This is a precondition for assuring long-term compatibility of magnitude data in national and international data catalogs and their usefulness for seismic hazard assessment and research;
- Improvement of current procedures for direct determination of seismic moment and energy in a wider magnitude range than currently possible, down to small magnitudes that are at present well covered only by  $M_L$  and  $m_b$ ;
- Development of regional calibration functions for  $m_b$  and  $m_B$ , which will permit more reliable and much faster body-wave magnitude estimates from records at distances down to about  $5^\circ$
- Development and consequent use of standard procedures for  $M_o$  and  $E_s$  measurements that assure non-saturating and globally compatible estimates of seismic moment and energy and of their related magnitude scales  $M_w$  and  $M_c$ ;
- Use of these data for in-depth studies in the regional variability of apparent stress conditions and their relevance for improving (time-variable) regional earthquake and tsunami hazard and risk assessment;
- Comprehensive testing of speed and reliability of the various methods recently proposed for more rapid (near) real-time magnitude estimates (e. g. [9,10,32,46,55,56,57,59,61,66]) under operational EWS conditions;
- Development of faster automated procedures for direct non-saturating  $M_w$  and  $M_c$  determination for improving quick and realistic disaster response;
- Development of alternative automatic (near) real-time procedures of magnitude determination such as the

rapid finite-source analysis [21], their scaling to both seismic energy and moment and operational testing also for very large earthquakes.

## Bibliography

### Primary Literature

1. Abe K (1981) Magnitudes of large shallow earthquakes from 1904 to 1980. *Phys Earth Planet Int* 27:72–92
2. Abe K (1984) Complements to Magnitudes of large shallow earthquakes from 1904 to 1980. *Phys Earth Planet Int* 34:17–23
3. Aki K (1967) Scaling law of seismic spectrum. *J Geophys Res* 72(4):1217–1231
4. Båth M (1981) Earthquake magnitude – recent research and current trends. *Earth Sci Rev* 17:315–398
5. Boatwright J, Choy GL (1986) Teleseismic estimates of the energy radiated by shallow earthquakes. *J Geophys Res* 91(B2):2095–2112
6. Bormann P (ed) (2002) IASPEI New manual of seismological observatory practice, vol 1 and 2. GeoForschungsZentrum, Potsdam
7. Bormann P, Khalturin V (1975) Relations between different kinds of magnitude determinations and their regional variations. In: Proceed XIVth General Ass European Seism. Comm Trieste Sept pp 16–22, 1974. Academy of Sciences of DDR, Berlin, pp 27–39
8. Bormann P, Baumbach M, Bock G, Gresser H, Choy GL, Boatwright J (2002) Seismic sources and source parameters. In: Bormann P (ed) IASPEI New manual seismological observatory practice. GeoForschungsZentrum Potsdam, chap 3, pp 1–94
9. Bormann P, Wylegalla K (2005) Quick estimator of the size of great earthquakes. *EOS* 86(46):464
10. Bormann P, Wylegalla K, Saul J (2006) Broadband body-wave magnitudes  $m_B$  and  $m_{BC}$  for quick reliable estimation of the size of great earthquakes. <http://spring.msi.umn.edu/USGS/Posters/>
11. Bormann P, Liu R, Ren X, Gutdeutsch R, Kaiser D, Castellaro S (2007) Chinese national network magnitudes, their relation to NEIC magnitudes, and recommendations for new IASPEI magnitude standards. *Bull Seism Soc Am* 97(1B):114–127
12. Brune JN (1970) Tectonic stress and the spectra of shear waves from earthquakes. *J Geophys Res* 75:4997–5009
13. Brune JN, Engen GR (1969) Excitation of mantle Love waves and definition of mantle wave magnitude. *Bull Seism Soc Am* 49:349–353
14. Castellaro S, Mulargia F, Kagan YY (2006) Regression problems for magnitudes. *Geophys J Int* 165:913–930
15. Castellaro S, Bormann P (2007) Performance of different regression procedures on the magnitude conversion problem. *Bull Seism Soc Am* 97:1167–1175
16. Choy GL, Boatwright J (1995) Global patterns of radiated seismic energy and apparent stress. *J Geophys Res* 100, B9:18,205–18,228
17. Choy GL, Boatwright J, Kirby SH (2001) The radiated seismic energy and apparent stress of interplate and intraslab earthquakes at subduction zone environments: Implications for seismic hazard estimation. US Geological Survey Open-File Report 01–0005:18

18. Choy GL, Kirby S (2004) Apparent stress, fault maturity and seismic hazard for normal-fault earthquakes at subduction zones. *Geophys J Int* 159:991–1012
19. Choy GL, McGarr A, Kirby SH, Boatwright J (2006) An overview of the global variability in radiated energy and apparent stress. In: Abercrombie R, McGarr A, Kanamori H (eds) Radiated energy and the physics of earthquake faulting, AGU. *Geophys Monogr Ser* 170:43–57
20. Choy GL, Boatwright J (2007) The energy radiated by the 26 December 2004 Sumatra-Andaman earthquake estimated from 10-minute *P*-wave windows. *Bull Seism Soc Am* 97:18–24
21. Dregers DS, Gee L, Lombard P, Murray MH, Romanowicz B (2005) Rapid finite-source analysis and near-fault strong ground motions: Application to the 2003  $M_w$ 6.5 San Simeon and 2004  $M_w$ 6.0 Parkfield earthquakes. *Seism Res Lett* 76:40–48
22. Duda SJ (1965) Secular seismic energy release in the circum-Pacific belt. *Tectonophysics* 2:409–452
23. Dziewonski AM, Chou TA, Woodhouse JH (1981) Determination of earthquake source parameters from waveform data for studies of global and regional seismicity. *J Geophys Res* 86:2825–2852
24. Ekström G, Dziewonski AM (1988) Evidence of bias in estimations of earthquake size. *Nature* 332:319–323
25. Geller RJ, Kanamori H (1977) Magnitudes of great shallow earthquakes from 1904 to 1952. *Bull Seism Soc Am* 67:587–598
26. Gutenberg B (1945) Amplitudes of *P*, *PP*, and *S* and magnitude of shallow earthquakes. *Bull Seism Soc Am* 35:57–69
27. Gutenberg B (1945) Magnitude determination of deep-focus earthquakes. *Bull Seism Soc Am* 35:117–130
28. Gutenberg B (1945) Amplitude of surface waves and magnitude of shallow earthquakes. *Bull Seism Soc Am* 35(3):3–12
29. Gutenberg B, Richter CF (1954) Seismicity of the earth and associated phenomena, 2nd Edn. Princeton University Press, Princeton
30. Gutenberg B, Richter CF (1956) Magnitude and energy of earthquakes. *Ann Geofis* 9:1–15
31. Hanks TC, Kanamori H (1979) A moment magnitude scale. *J Geophys Res* 84(B5):2348–2350
32. Hara T (2007) Measurement of duration of high-frequency energy radiation and its application to determination of magnitudes of large shallow earthquakes. *Earth Planet Space* 59:227–231
33. Haskell N (1964) Total energy and energy spectral density of elastic wave radiation from propagating faults, vol 1. *Bull Seismol Soc Am* 54:1811–1842
34. Haskell N (1964) Total energy and energy spectral density of elastic wave radiation from propagating faults, vol 2. *Bull Seismol Soc Am* 56:125–140
35. Hatzidimitriou P, Papazachos C, Kiratzi A, Theodoulidis N (1993) Estimation of attenuation structure and local earthquake magnitude based on acceleration records in Greece. *Tectonophysics* 217:243–253
36. Herak M, Herak D (1993) Distance dependence of  $M_S$  and calibrating function for 20 second Rayleigh waves. *Bull Seism Soc Am* 83:1881–1892
37. Herak M, Panza GF, Costa G (2001) Theoretical and observed depth corrections for  $M_S$ . *Pure Appl Geophys* 158:1517–1530
38. Hutton LK, Boore DM (1987) The  $M_L$  scale in Southern California. *Bull Seism Soc Am* 77:2074–2094
39. Hyvernaud O, Reymond D, Talandier J, Okal EA (1993) Four years of automated measurements of seismic moments at Paapeete using the mantle magnitude  $M_m$ : 1987–1991. In: Duda SJ, Yanovskaya TB (eds) Special section: Estimation of earthquake size. *Tectonophysics* 217:175–193. Elsevier Science
40. <http://neic.usgs.gov/neis/sopar>
41. <http://www.globalcmt.org/CMTsearch.html>
42. [http://www.iaspei.org/commissions/CSOI/Summary\\_of\\_WG\\_recommendations.pdf](http://www.iaspei.org/commissions/CSOI/Summary_of_WG_recommendations.pdf)
43. Kanamori H (1977) The energy release in great earthquakes. *J Geophys Res* 82:2981–2987
44. Kanamori H (1983) Magnitude scale and quantification of earthquakes. *Tectonophysics* 93:185–199
45. Kanamori H (1988) The importance of historical seismograms for geophysical research. In: Lee WHK (ed) Historical seismograms and earthquakes of the world. Academic Press, New York, pp 16–33
46. Kanamori H (2005) Real-time seismology and earthquake damage prediction. *Ann Rev Earth Planet Sci* 33:195–214
47. Kanamori H, Hauksson E, Heaton T (1997) Real-time seismology and earthquake hazard mitigation. *Nature* 390:461–464
48. Kanjo K, Furudate T, Tsuboi S (2006) Application of  $M_{wp}$  to the great December 26, 2004 Sumatra earthquake. *Earth Planet Space* 58:121–126
49. Katsumata M (1964) A method of determination of magnitude for near and deep-focus earthquakes (in Japanese with English abstract). *A J Seism* 22:173–177
50. Katsumata M (1996) Comparison of magnitudes estimated by the Japan Meteorological Agency with moment magnitudes for intermediate and deep earthquakes. *Bull Seism Soc Am* 86:832–842
51. Kaverina AN, Lander AV, Prozorov AG (1996) Global creep distribution and its relation to earthquake – source geometry and tectonic origin. *Geophys J Int* 135:249–265
52. Keilis-Borok VI (1959) On the estimation of displacement in an earthquake source and of source dimension. *Ann Geofis* 12:205–214
53. Kikuchi M, Ishida M (1993) Source retrieval for deep local earthquakes with broadband records. *Bull Seism Soc Am* 83:1855–1870
54. Lee V, Trifunac M, Herak M, Živčić M, Herak D (1990)  $M_L^{SM}$  computed from strong motion accelerograms recorded in Yugoslavia. *Earthq Eng Struct Dyn* 19:1167–1179
55. Lomax A (2005) Rapid estimation of rupture extent for large earthquakes: Application to the 2004, M9 Sumatra-Andaman mega-thrust. *Geophys Res Lett* 32:L10314
56. Lomax A, Michelini A (2005) Rapid determination of earthquake size for hazard warning. *EOS* 86(21):202
57. Lomax A, Michelini A, Piatanesi A (2007) An energy-duration procedure for rapid and accurate determination of earthquake magnitude and tsunamigenic potential. *Geophys J Int* 170:1195–1209
58. Mendez AJ, Anderson JG (1991) The temporal and spatial evolution of the 19 September 1985 Michoacan earthquake as inferred from near-source ground-motion records. *Bull Seism Soc Am* 81:1655–1673
59. Menke W, Levin R (2005) A strategy to rapidly determine the magnitude of great earthquakes. *EOS* 86(19):185–189
60. Nakamura Y (1989) Earthquake alarm system for Japan railways. *Jpn Railw Eng* 109:1–7
61. Nakamura Y, Saita J (2007) UrEDAS, the earthquake warning system: today and tomorrow. In: Gasperini P, Manfredi G,

- Zschau J (eds) Earthquake early warning systems. Springer, Berlin, pp 249–281
62. Nuttli OW (1986) Yield estimates of Nevada test site explosions obtained from seismic Lg waves. *J Geophys Res* 91:2137–2151
  63. Okal EA (1989) A theoretical discussion of time domain magnitudes: The Prague formula for  $M_S$  and the mantle magnitude  $M_m$ . *J Geophys Res* 94:4194–4204
  64. Okal EA, Talandier J (1989)  $M_m$ : A variable-period mantle magnitude. *J Geophys Res* 94:4169–4193
  65. Okal EA, Talandier J (1990)  $M_m$ : Extension to Love waves of the concept of a variable-period mantle magnitude. *Pure Appl Geophys* 134:355–384
  66. Olson EL, Allen R (2005) The deterministic nature of earthquake rupture. *Nature* 438:212–215
  67. Patton HJ (1998) Bias in the centroid moment tensor for central Asian earthquakes: Evidence from regional surface wave data. *J Geophys Res* 103(26):885–898
  68. Polet J, Kanamori H (2000) Shallow subduction zone earthquakes and their tsunamigenic potential. *Geophys J Int* 142:684–702
  69. Purcaru G, Berckhemer H (1978) A magnitude scale for very large earthquakes. *Tectonophysics* 49:189–198
  70. Rezapour M, Pearce RG (1998) Bias in surface-wave magnitude  $M_S$  due to inadequate distance corrections. *Bull Seism Soc Am* 88:43–61
  71. Richter CF (1935) An instrumental earthquake magnitude scale. *Bull Seism Soc Am* 25:1–32
  72. Richter CF (1958) Elementary seismology. W.H. Freeman, San Francisco
  73. Rydelek P, Horiuchi S (2006) Is earthquake rupture deterministic? *Nature* 444:E5–E6
  74. Soloviev SL (1955) Classification of earthquakes by energy value (in Russian). *Trudy Geophys Inst Acad Sci USSR* 39(157):3–31
  75. Spall H (1980) Charles F. Richter – an interview. *Earthq Inf Bull* 12(1):5–8
  76. Stein S, Okal E (2005) Speed and size of the Sumatra earthquake. *Nature* 434:581–582
  77. Talandier J, Okal EA (1992) One-station estimates of seismic moments from the mantle magnitude  $M_m$ : The case of the regional field ( $1.5^\circ \leq \Delta \leq 15^\circ$ ). *Pure Appl Geophys* 138:43–60
  78. Tsai VC, Nettles M, Ekström G, Dziewonski AM (2005) Multiple CMT source analysis of the 2004 Sumatra earthquake. *Geophys Res Lett* 32(L17304):1–4
  79. Tsuboi C (1954) Determination of the Gutenberg-Richter's magnitude of earthquakes occurring in and near Japan (in Japanese with English abstract). *Zisin Second Ser* 7:185–193
  80. Tsuboi S, Abe K, Takano K, Yamanaka Y (1995) Rapid determination of  $M_w$  from broadband P waveforms. *Bull Seism Soc Am* 85:606–613
  81. Tsuboi S, Whitmore PM, Sokolovski TJ (1999) Application of  $M_{wp}$  to deep and teleseismic earthquakes. *Bull Seism Soc Am* 89:1345–1351
  82. Uhrhammer RA, Collins ER (1990) Synthesis of Wood-Anderson seismograms from broadband digital records. *Bull Seism Soc Am* 80:702–716
  83. Utsu T (2002) Relationships between magnitude scales. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of earthquake and engineering seismology, Part A*. Academic Press, Amsterdam, pp 733–746
  84. Vanek J, Zátopek A, Kárník V, Kondorskaya N, Riznichenko Y, Savarenski S, Solovév S, Shebalin N (1962) Standardization of magnitude scales. *Izv Acad Sci USSR, Geophys Ser*, pp 108–111 (English translation)
  85. Weinstein S, Okal E (2005) The mantle magnitude  $M_m$  and the slowness parameter  $\Theta$ : Five years of real-time use in the context of tsunami warning. *Bull Seism Soc Am* 95:779–799
  86. Whitmore PM, Tsuboi S, Hirshorn B, Sokolowski TJ (2002) Magnitude-dependent correction for  $M_{wp}$ . *Sci Tsunami Hazard* 20(4):187–192
  87. Willmore PL (ed) (1979) *Manual of seismological observatory practice, World data center A for solid earth geophysics*. Report SE–20. Boulder, Colorado
  88. Wu Z (2001) Scaling of apparent stress from broadband radiated energy catalogue and seismic moment catalogue and its focal mechanism dependence. *Earth Planets Space* 53:943–948
  89. Wu KM, Kanamori H (2005) Experiment on an onsite early warning method for the Taiwan early warning system. *Bull Seism Soc Am* 95:347–353
  90. Wyss M, Brune JN (1968) Seismic moment, stress, and source dimensions for earthquakes in the California-Nevada regions. *J Geophys Res* 73:4681–4694

## Books and Reviews

- Båth M (1979) *Introduction to seismology*. Birkhäuser, Basel
- Bolt BA (1999) *Earthquakes*, 4th edn. W.H. Freeman, San Francisco
- Duda S, Aki K (Eds) (1983) *Quantification of earthquakes*. *Tectonophysics* 93(3/4):183–356
- Gasperini P, Manfredi G, Zschau J (eds) (2007) *Earthquake early warning systems*. Springer, Berlin
- Kulhánek O (1990) *Anatomy of seismograms*. *Developments in solid earth geophysics*, vol 18. Elsevier, Amsterdam
- Lay T, Wallace TC (1995) *Modern global seismology*. Academic Press, San Diego
- Lee WHK (ed) (1988) *Historical seismograms and earthquakes of the world*. Academic Press, New York
- Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) (2002) *International handbook of earthquake and engineering seismology, part A and B*. Academic Press, London (an imprint of Elsevier Science)
- Scholz CH (2002) *The mechanics of earthquake faulting*, 2nd edn. Cambridge University Press, Cambridge
- Shearer PM (1999) *Introduction to seismology*. Cambridge University Press, Cambridge
- Stein S, Wyssession M (2002) *Introduction to seismology, Earthquakes and earth structure*. Blackwell Publishing, Malden

## Earthquake Monitoring and Early Warning Systems

WILLIAM H. K. LEE<sup>1</sup>, YIH-MIN WU<sup>2</sup>

<sup>1</sup> US Geological Survey, Menlo Park, USA

<sup>2</sup> Department of Geosciences, National Taiwan University, Taipei, Taiwan

### Article Outline

Glossary

Definition of the Subject

Introduction

Earthquake Monitoring: Instrumentation

Earthquake Monitoring:

Regional and Local Networks

Seismograms and Derived Products

Earthquake Early Warning (EEW) Systems

Future Directions

Acknowledgments

Appendix: A Progress Report on Rotational Seismology

Bibliography

### Glossary

**Active fault** A *fault* (q.v.) that has moved in historic (e. g., past 10,000 years) or recent geological time (e. g., past 500,000 years).

**Body waves** Waves which propagate through the interior of a body. For the Earth, there are two types of seismic body waves: (1) compressional or longitudinal (*P* wave), and (2) shear or transverse (*S* wave).

**Coda waves** Waves which are recorded on a *seismogram* (q.v.) after the passage of *body waves* (q.v.) and *surface waves* (q.v.). They are thought to be back-scattered waves due to the Earth's inhomogeneities.

**Earthquake early warning system (EEWS)** An earthquake monitoring system that is capable of issuing warning message after an earthquake occurred and before strong ground shaking begins.

**Earthquake precursor** Anomalous phenomenon preceding an earthquake.

**Earthquake prediction** A statement, in advance of the event, of the time, location, and *magnitude* (q.v.) of a future earthquake.

**Epicenter** The point on the Earth's surface vertically above the *hypocenter* (q.v.).

**Far-field** Observations made at large distances from the *hypocenter* (q.v.), compared to the wave-length and/or the source dimension.

**Fault** A fracture or fracture zone in the Earth along which the two sides have been displaced relative to one another parallel to the fracture.

**Fault slip** The relative displacement of points on opposite sides of a *fault* (q.v.), measured on the fault surface.

**Focal mechanism** A description of the orientation and sense of slip on the causative fault plane derived from analysis of *seismic waves* (q.v.).

**Hypocenter** Point in the Earth where the rupture of the rocks originates during an earthquake and *seismic waves* (q.v.) begin to radiate. Its position is usually determined from arrival times of *seismic waves* (q.v.) recorded by *seismographs* (q.v.).

**Intensity, earthquake** Rating of the effects of earthquake vibrations at a specific place. Intensity can be estimated from instrumental measurements, however, it is formally a rating assigned by an observer of these effects using a descriptive scale. Intensity grades are commonly given in Roman numerals (in the case of the Modified Mercalli Intensity Scale, from I for "not perceptible" to XII for "total destruction").

**Magnitude, earthquake** Quantity intended to measure the size of earthquake at its source, independent of the place of observation. *Richter magnitude* ( $M_L$ ) was originally defined in 1935 as the logarithm of the maximum amplitude of seismic waves in a seismogram written by a Wood–Anderson seismograph (corrected to) a distance of 100 km from the epicenter. Many types of magnitudes exist, such as *body-wave magnitude* ( $m_b$ ), *surface-wave magnitude* ( $M_S$ ), and *moment magnitude* ( $M_W$ ).

**Moment tensor** A symmetric second-order tensor that characterizes an internal seismic point source completely. For a finite source, it represents a point source approximation and can be determined from the analysis of *seismic waves* (q.v.) whose wavelengths are much greater than the source dimensions.

**Near-field** A term for the area near the causative rupture of an earthquake, often taken as extending a distance from the rupture equal to its length. It is also used to specify a distance to a seismic source comparable or shorter than the wavelength concerned. In engineering applications, near-field is often defined as the area within 25 km of the fault rupture.

**Plate tectonics** A theory of global *tectonics* (q.v.) in which the Earth's lithosphere is divided into a number of essentially rigid plates. These plates are in relative motion, causing earthquakes and deformation along the plate boundaries and adjacent regions.

**Probabilistic seismic hazard analysis** Available information on earthquake sources in a given region is

combined with theoretical and empirical relations among earthquake *magnitude* (q.v.), distance from the source, and local site conditions to evaluate the exceedance probability of a certain ground motion parameter, such as the peak ground acceleration, at a given site during a prescribed time period.

**Seismic hazard** Any physical phenomena associated with an earthquake (e.g., ground motion, ground failure, liquefaction, and tsunami) and their effects on land use, man-made structure, and socio-economic systems that have the potential to produce a loss.

**Seismic hazard analysis** The calculation of the *seismic hazard* (q.v.), expressed in probabilistic terms (See *probabilistic seismic hazard analysis*, q.v.). The result is usually displayed in a *seismic hazard map* (q.v.).

**Seismic hazard map** A map showing contours of a specified ground-motion parameter or response spectrum ordinate for a given *probabilistic seismic hazard analysis* (q.v.) or return period.

**Seismic moment** The magnitude of the component couple of the double couple that is the point force system equivalent to a *fault slip* (q.v.) in an isotropic elastic body. It is equal to rigidity times the fault slip integrated over the fault plane. It can be estimated from the far-field seismic spectrum at wave lengths much longer than the source size. It can also be estimated from the near-field seismic, geologic and geodetic data. Also called “scalar seismic moment” to distinguish it from *moment tensor* (q.v.).

**Seismic risk** The risk to life and property from earthquakes.

**Seismic wave** A general term for waves generated by earthquakes or explosions. There are many types of seismic waves. The principle ones are *body waves* (q.v.), *surface waves* (q.v.), and *coda waves* (q.v.).

**Seismograph** Instrument which detects and records ground motion (and especially vibrations due to earthquakes) along with timing information. It consists of a *seismometer* (q.v.) a precise timing device, and a recording unit (often including telemetry).

**Seismogram** Record of ground motions made by a *seismograph* (q.v.).

**Seismometer** Inertial sensor which responds to ground motions and produces a signal that can be recorded.

**Source parameters of an earthquake** The parameters specified for an earthquake source depends on the assumed earthquake model. They are origin time, *hypo-center* (q.v.), *magnitude* (q.v.), *focal mechanism* (q.v.), and *moment tensor* (q.v.) for a point source model. They include fault geometry, rupture velocity, stress drop, slip distribution, etc. for a finite fault model.

**Surface waves** Waves which propagate along the surface of a body or along a subsurface interface. For the Earth, there are two common types of seismic surface waves: Rayleigh waves and Love waves (both named after their discoverers).

**Tectonics** Branch of Earth science which deals with the structure, evolution, and relative motion of the outer part of the Earth, the lithosphere. The lithosphere includes the Earth’s crust and part of the Earth’s upper mantle and averages about 100 km thick. See *plate tectonics* (q.v.).

**Teleseism** An earthquake at an epicentral distance greater than about 20° or 2000 km from the place of observation.

### Definition of the Subject

When a sudden rupture occurs in the Earth, elastic (seismic) waves are generated. When these waves reach the Earth’s surface, we may feel them as a series of vibrations, which we call an earthquake. *Seismology* is derived from the Greek word *σεισμός* (seismos or earthquake) and *λόγος* (logos or discourse); thus, it is the science of earthquakes and related phenomena. Seismic waves can be generated naturally by earthquakes or artificially by explosions or other means. We define earthquake monitoring as a branch of seismology, which systematically observes earthquakes with instruments over a long period of time.

Instrumental recordings of earthquakes have been made since the later part of the 19th century by seismographic stations and networks of various sizes from local to global scales. The observed data have been used, for example, (1) to compute the source parameters of earthquakes, (2) to determine the physical properties of the Earth’s interior, (3) to test the theory of plate tectonics, (4) to map active faults, (5) to infer the nature of damaging ground shaking, and (6) to carry out seismic hazard analyzes. Constructing a satisfactory theory of the complex earthquake process has not yet been achieved within the context of physical laws, e.g., realistic equations for modeling earthquakes do not exist at present. Good progress, however, has been made in building a physical foundation for the earthquake source process [62], partly as a result of research directed toward earthquake prediction.

Earthquakes release large amounts of energy that potentially can cause significant damage and human deaths. During an earthquake, potential energy (mainly elastic strain energy and some gravitational energy) that has accumulated in the hypocentral region over decades to centuries or longer is released suddenly [63]. This energy is partitioned into (1) radiated energy in the form of prop-

agating seismic waves, (2) energy consumed in overcoming fault friction, (3) the energy which expands the rupture surface area or changes its properties (e.g., by pulverizing rock), and (4) heat. The radiated seismic energy is a small fraction (about 7%) of the total energy budget, and it can be estimated using the recorded seismograms. Take, for example, the 1971 San Fernando earthquake ( $M_W = 6.6$ ) in southern California. Its radiated energy was about  $5 \times 10^{21}$  ergs, or about 120 kilotons of TNT explosives, or the energy released by six atomic bombs of the size used in World War II. The largest earthquake recorded instrumentally (so far) is the 1960 Chilean earthquake ( $M_W = 9.5$ ). Its radiated energy was about  $1.1 \times 10^{26}$  ergs, an equivalent of about 2,600 megatons of TNT explosives, the energy released by about 130,000 atomic bombs. It is, therefore, no surprise that an earthquake can cause up to hundreds of thousands of human deaths, and produce economic losses of up to hundreds of billions of dollars.

Monitoring earthquakes is essential for providing scientific data to investigate complex earthquake phenomena, and to mitigate seismic hazards. The present article is a brief overview of earthquake monitoring and early warning systems; it is intended for a general scientific audience, and technical details can be found in the cited references. Earthquakes are complex natural phenomena and their monitoring requires an interdisciplinary approach, including using tools from computer science, electrical and electronic engineering, mathematics, physics, and others. Earthquake early warning systems (which are based on earthquake monitoring) offer practical information for reducing seismic hazards in earthquake-prone regions.

After the “Introduction”, we will present a summary of earthquake monitoring, a description of the products derived from the analysis of seismograms, and a discussion of the limitations of these products. Earthquake early warning systems are then presented briefly, and we conclude with a section on future directions, including a progress report on rotational seismology (Appendix). We present overviews of most topics in earthquake monitoring, and an extensive bibliography is provided for additional reading and technical details.

## Introduction

Earthquakes, both directly and indirectly, have caused much suffering to mankind. During the 20th century alone about two million people were killed as a result of earthquakes. A list of deadly earthquakes (death tolls  $\geq 25$ ) of the world during the past five centuries was compiled by Utsu [115]. It shows that earthquakes of magnitude  $\geq 6$

( $\sim 150$  per year worldwide) can be damaging and deadly if they occur in populated areas, and if their focal depths are shallow ( $< 50$  km). Seismic risk can be illustrated by plotting the most deadly earthquakes of the past five centuries (1500–2000) over a map of current population density. This approach was used by Utsu [115], and his result is shown in Fig. 1. Most of these deadly earthquakes are concentrated (1) along the coasts of Central America, the Caribbean, western South America, and Indonesia, and (2) along a belt that extends from southern Europe, the Middle East, Iran, Pakistan and India, to China and Japan.

Table 1 lists the most deadly earthquakes (death toll  $> 20,000$ ) of the past 110 years based on official estimates (often under-estimated for political reasons, or lack of accurate census data in many areas of the world). In the first 5 years of the 21st century, four disastrous earthquakes occurred in India, Indonesia, Iran, and Pakistan. In the 20th century, the average death toll caused by earthquakes (and tsunamis they triggered) was about 16,000 per year. For the past seven years the yearly death toll was about 60,000 – four times higher than the average in the previous century. In Fig. 2 we extracted a portion of Fig. 1 to illustrate the relationship between past earthquakes and population in India, Pakistan, northern Indonesia, and adjoining regions. We numbered the four most recent disastrous earthquakes in Fig. 2. It is obvious that the large populations in India, Indonesia, Iran, Pakistan, and their adjoining regions (over 1.5 billion people) has been and will continue to be adversely affected by earthquakes. Fatalities depend largely on resistance of building construction to shaking, in addition to population density and earthquake occurrence.

In recent decades, population increases, accelerated urbanization, and population concentration along coastal areas prone to earthquakes suggest that many more earthquake-related fatalities will occur unless effective steps are taken to minimize earthquake and tsunami hazards.

## Earthquake Monitoring: Instrumentation

Besides geodetic data [28], the primary instrumental data for the quantitative study of earthquakes are *seismograms*, records of ground motion caused by the passage of seismic waves. Seismograms are written by *seismographs*, instruments which detect and record ground motion along with timing information. A seismograph consists of three basic components: (1) a seismometer, which responds to ground motion and produces a signal proportional to acceleration, velocity, or displacement over a range of amplitudes and frequencies; (2) a timing device; (3) either a local recording unit which writes seismograms on paper, film, or elec-

Earthquake Monitoring and Early Warning Systems, Table 1  
Deadly Earthquakes/Tsunamis from 1896–2005 ([115] and recent sources)

Origin Time Year MM/DD Hr:Min (UTC, except L=local)	Hypocenter			Magnitude	Location	Deaths (Approximate)
	Lat. (deg)	Lon. (deg)	Depth (km)			
2005 10/08 3:50	34.432	73.573	10	7.6	Pakistan, Kashmir	80,361+
2004 12/26 0:58	3.298	95.778	7	9.2	Indonesia, Sumatra	283,106+
2003 12/26 1:56	29.004	58.337	15	6.6	Iran, Bam	26,000
2001 01/26 3:16	23.420	70.230	16	7.7	India, Gujarat, Bhuj	20,000+
1990 06/20 21:00	37.008	49.213	18	7.4	Iran, western	~40,000
1988 12/07 7:41	40.919	44.119	7	6.8	Armenia, Spitak	~40,000
1976 07/27 19:42	39.605	117.889	17	7.6	China, Tangshan	~242,000
1976 02/04 9:01	15.298	-89.145	13	7.5	Guatemala	23,000
1970 05/31 20:23	-9.248	-78.842	73	7.5	Peru	67,000
1948 10/05 20:12	37.500	58.000	0	7.2	USSR, Ashgabat	~65,000
1939 12/26 23:57	39.770	39.533	35	7.7	Turkey, Erzincan	33,000
1939 01/25 3:32	-36.200	-72.200	0	7.7	Chile, Chillian	28,000
1935 05/30 21:32	28.894	66.176	35	8.1	Pakistan, Quetta	60,000
1932 12/25 2:04	39.771	96.690	25	7.6	China, Gansu	~70,000
1927 05/22 22:32	37.386	102.311	25	7.7	China, Tsinghai	~100,000
1923 09/01 2:58	35.405	139.084	35	7.9	Japan, Kanto	143,000
1920 12/16 12:05	36.601	105.317	25	8.6	China, Gansu	~240,000
1915 01/13 6:52	42.000	13.500	0	6.9	Italy, Avezzano	33,000
1908 12/28 4:20	38.000	15.500	0	7.0	Italy, Messina	~82,000
1906 08/17 0:40	-33.000	-72.000	0	8.2	Chile, Valparaiso	20,000
1905 04/04 0:50	33.000	76.000	0	8.1	India, Kangra	20,000
1896 06/15 19:32L	39.500	144.000	0	8.2	Japan, Sanriku-oki	22,000

"~" denotes large uncertainties because a range of deaths had been reported.

"+" denotes a minimum value.

tronic storage media, or more recently, a telemetry system for delivering the seismograms to a central laboratory for recording. Technical discussions of seismometry may be found, for example, in Wielandt [122], and of seismic instruments in Havskov and Alguacil [48]. An overview of challenges in observational earthquake seismology is given by Lee [71], and a useful manual of seismological observatory practice is provided by Bormann [12].

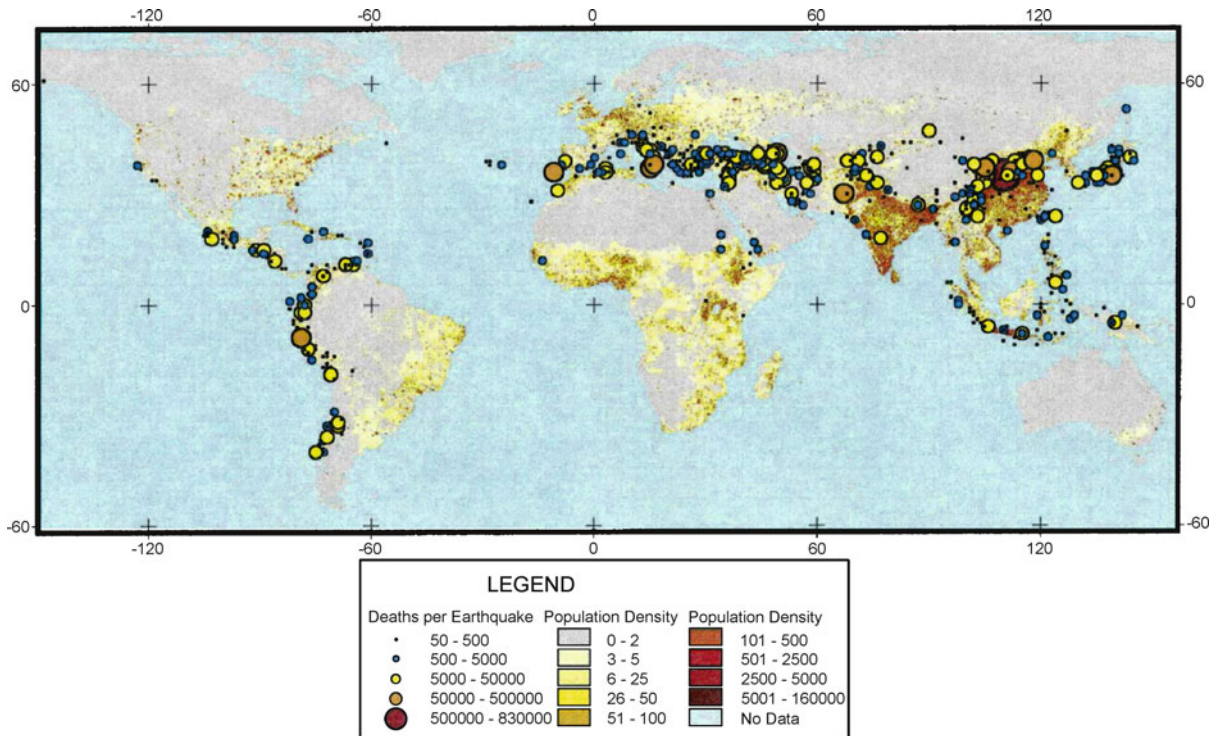
An *accelerograph* is a seismograph designed to record, on scale, the acceleration time history of strong ground motions. Measuring acceleration is important for studying response of buildings to strong ground motions close to earthquakes. Many modern sensitive seismographs are *velocigraphs* recording the time history of ground velocity. They are designed to measure seismic waves of small amplitudes (because seismic waves attenuate quickly from their sources) either from small earthquakes nearby, or from large earthquakes that are far away.

A seismic network (or an "array") is a group of seismographs "linked" to a central headquarters. Nowadays the

link is by various methods of telemetry, but in early days the links were by mail or telegrams, or simply by manual collection of the records. When we speak of a seismic *station*, we may mean an observatory with multiple instruments in special vaults or a small instrument package at a remote site.

Seismographs were first developed in the late 19th century, and individual seismographic observatories (often a part of astronomical or meteorological observatories) began earthquake monitoring by issuing earthquake information in their station bulletins and other publications. However, in order to accurately locate an earthquake, data from several seismographic stations are necessary. It was then natural for many governments to assume responsibility for monitoring earthquakes within their territories. However, because seismic waves from earthquakes do not recognize national boundaries, the need for international cooperation became clear. In the following subsections, we present an overview of the history and results of earthquake monitoring.





Earthquake Monitoring and Early Warning Systems, Figure 1

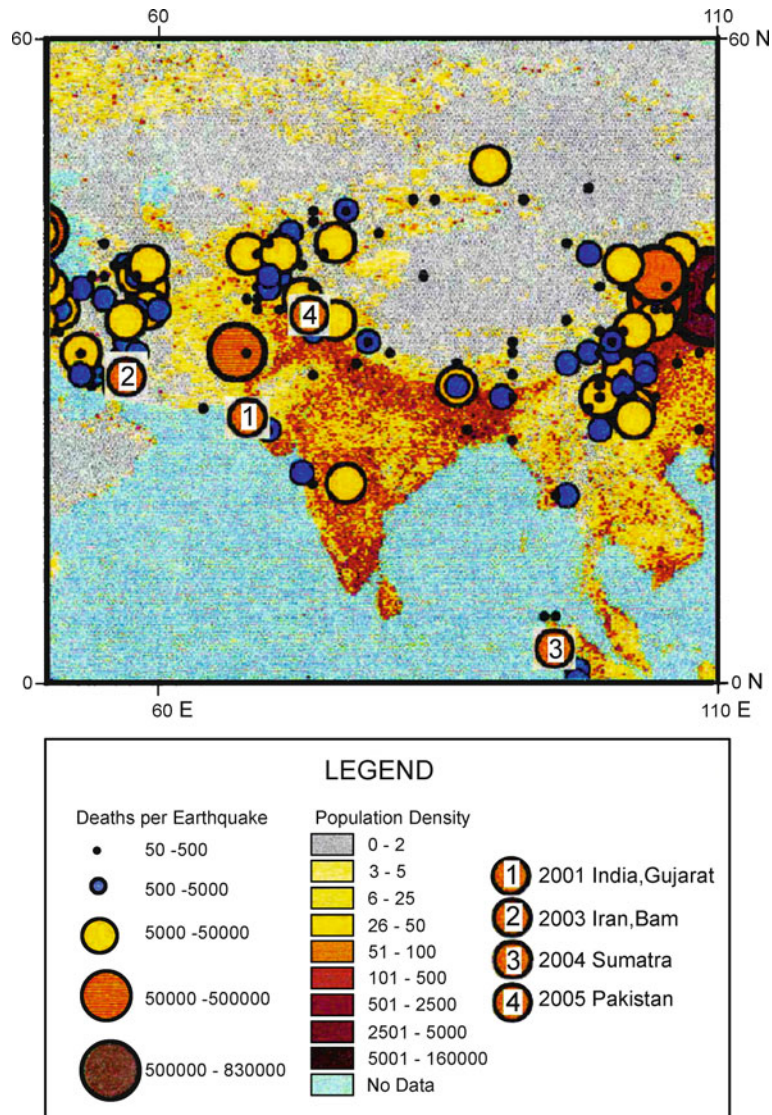
Location of deadly earthquakes around the world, 1500–2000. Population density is shown by the *background colors*. See [115] for details

## Historical Developments

In 1897, John Milne designed the first inexpensive seismograph, which was capable of recording very large earthquakes anywhere in the world. With a small grant from the British Association for the Advancement of Science (BAAS), a few other donations, and his own money, Milne managed to deploy about 30 of his instruments around the world, forming the first worldwide seismographic network. At the same time, seismogram readings were reported voluntarily to Milne's observatory at Shide on the Isle of Wight, England. A global earthquake summary with these seismogram readings was issued by Milne beginning in 1899. These summaries are now known as the "Shide Circulars". Milne also published progress and results in the "Reports of the BAAS Seismological Committee" from 1895 to 1913. A review of Milne's work and a reproduction of his publications as computer readable files were given by Schweitzer and Lee [101] and its attached CD-ROM. After Milne's death in 1913, Herbert H. Turner continued Milne's efforts, and in 1918 established publication of the International Seismological Summary (ISS).

The shortcomings of the Milne seismograph (low magnification, no damping, and poor time resolution) were soon recognized. Several improved seismographs (notably the Omori, Bosch–Omori, Wiechert, Galitzin, and Milne–Shaw) were developed and deployed in the first three decades of the 20th century. Figure 3 shows several of these classical seismographs (see Schweitzer and Lee [101] for further explanation). Although the ISS provided an authoritative compilation arrival-time data of seismic waves and determinations of earthquake hypocenters beginning in 1918, its shortcomings were also evident. These include difficulties in collecting the available arrival-time data around the world (which were submitted on a voluntary basis), and in the processing and analysis of data from many different types of seismographs. Revolutions and wars during the first half of the 20th century frequently disrupted progress, particularly impacting collection and distribution earthquake information.

In the late 1950s, attempts to negotiate a comprehensive nuclear test ban treaty failed, in part because of perceptions that seismic methods were inadequate for monitoring underground nuclear tests [95]. The influ-



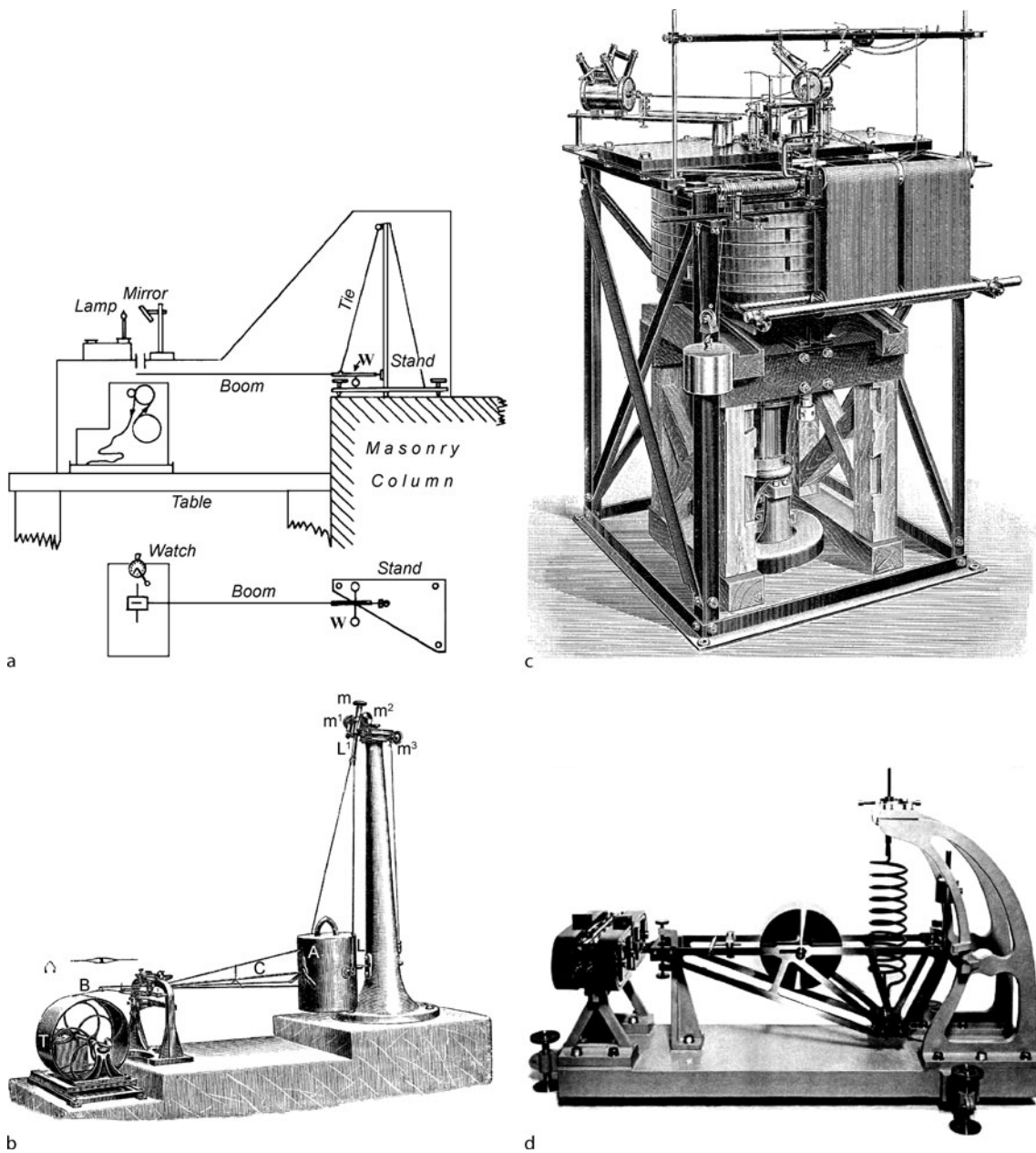
Earthquake Monitoring and Early Warning Systems, Figure 2

Location of the 4 most deadly earthquakes of the 21st century (up to the end of 2007) on a map showing the location of the deadly earthquakes from 16th to 20th centuries (after [115] and Table 1)

ential Berkner report of 1959 therefore advocated major support for seismology [66]. As a result, the Worldwide Standardized Seismograph Network (WWSSN) was created in the early 1960s with about 120 continuously recording stations located across much of the world, except China and the USSR [91]. Each WWSSN station was equipped with *identical* sets of short-period and long-period three-component seismographs and accurate chronometers. Figure 4 shows some of the equipment at a WWSSN station, including three-components of long-period seismometers, long-period recording and

test instruments, and the time and power console. A similar set of three-component short-period seismometers and recording and test instruments, nearly identical in appearance, was also deployed at each station. Seismograms from the WWSSN were sent to the United States to be photographed on 70 mm film chips for distribution (about US\$ 1 per chip as then sold to any interested person).

The WWSSN network is credited with making possible rapid progress in global seismology, and with helping to spark the plate tectonics revolution of the late



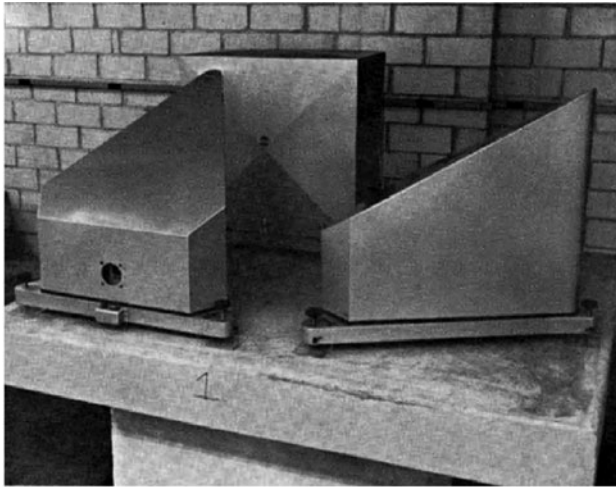
Earthquake Monitoring and Early Warning Systems, Figure 3  
Some classical seismographs: a Milne, b Bosch-Omori, c Wiechert, and d Galitzin (after [101])

1960s [117]. At about the same time, the Unified System of Seismic Observations (ESSN) of the former USSR and its allied countries was established, consisting of almost 100 stations equipped with Kirnos short-period, 1–20 s displacement sensors, and long-period seismographs.

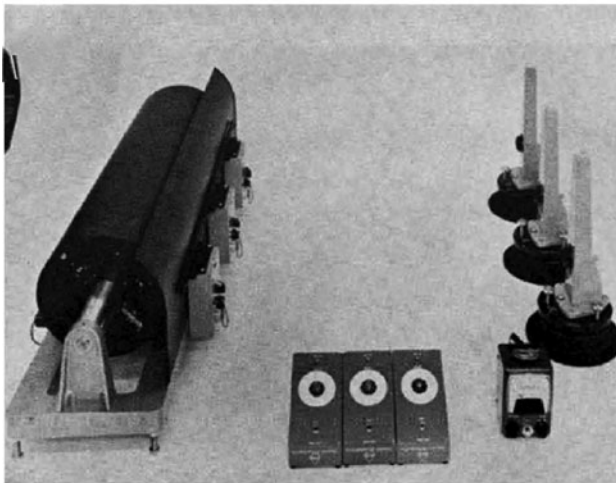
Samples of seismograms recorded on smoked paper and photographic paper or film by analog seismographs

are shown in Figs. 5 and 6. Two efforts to preserve and make such records available online are now underway: the *SeismoArchives* ([www.iris.edu/seismo/](http://www.iris.edu/seismo/) [72]), and *Sismos* ([sismos.rm.ingv.it](http://sismos.rm.ingv.it) [82]).

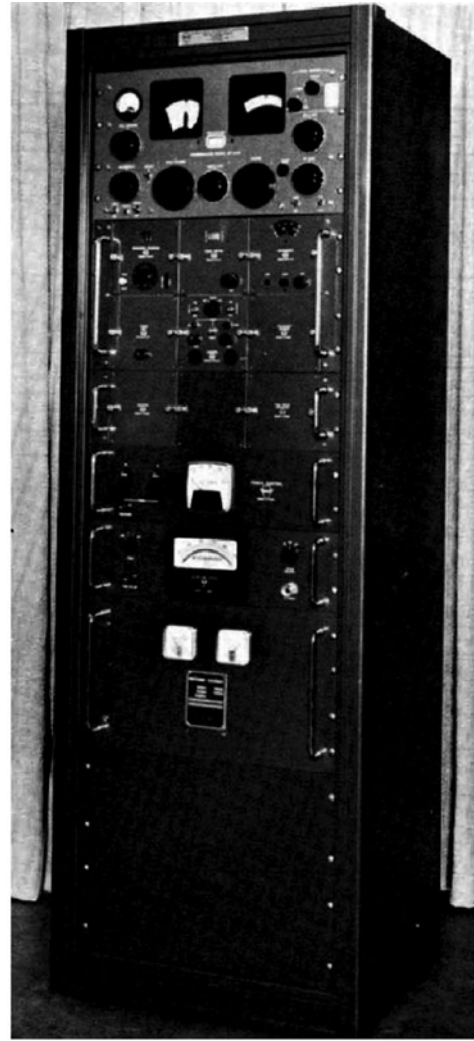
With the establishment of the WWSSN, the United States also assumed the task of monitoring earthquakes on a global scale beginning in the early 1960s. The mis-



a LONG-PERIOD SEISMOMETERS INSTALLED ON A PIER



b LONG-PERIOD AND TEST INSTRUMENTS



c TIME AND POWER CONSOLE

Earthquake Monitoring and Early Warning Systems, Figure 4

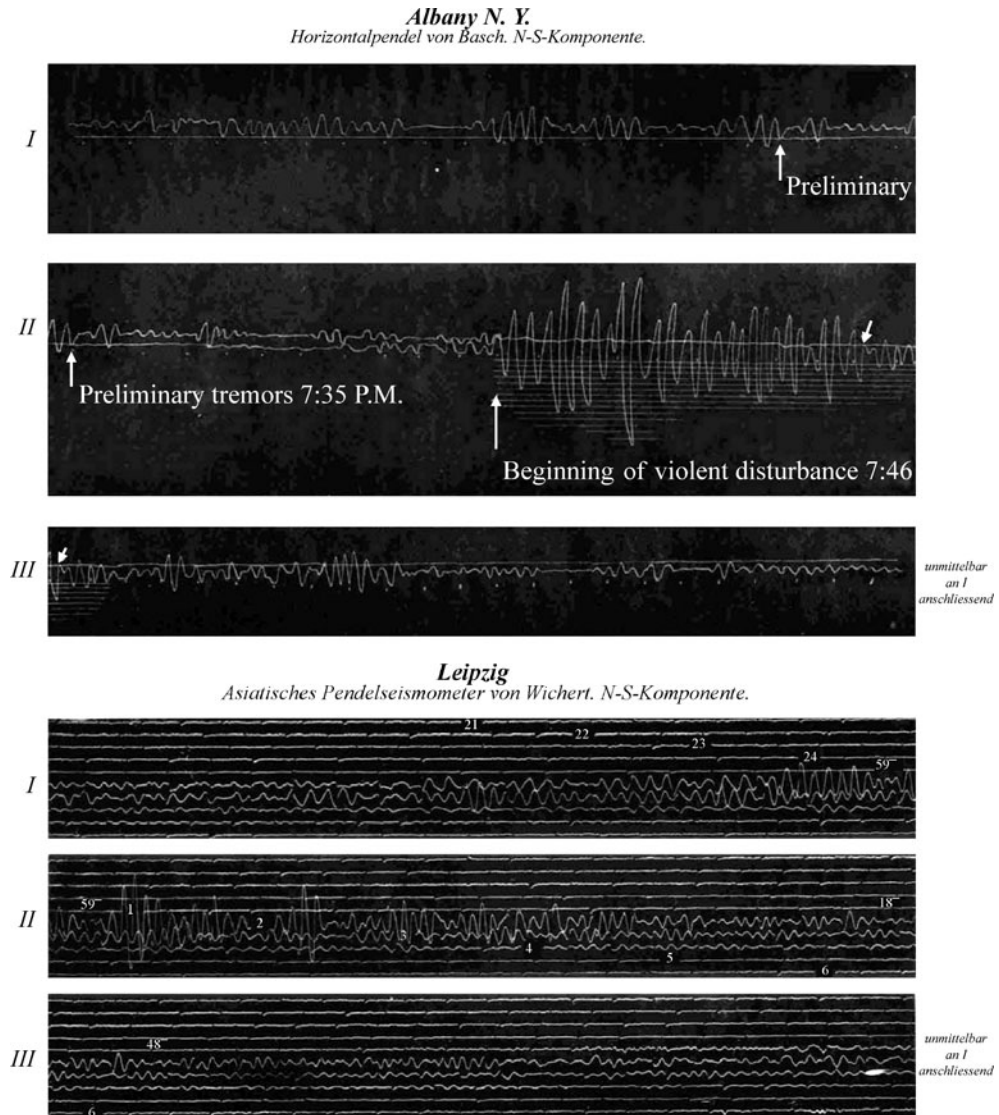
Some WWSSN station equipment: **a** Three-component, long-period seismometers installed on a seismic pier, **b** Long-period recording and test instruments, and **c** Time and power console. A similar set of three-component, short-period seismometers and recording/test instruments is not shown

sion of the US National Earthquake Information Center (NEIC, now part of the US, Geological Survey) is “to determine rapidly the location and size of all destructive earthquakes worldwide and to immediately disseminate this information to concerned national and international agencies, scientists, and the general public” (<http://earthquake.usgs.gov/regional/neic/>).

In 1964, the ISS was reorganized as the International Seismological Centre (ISC). Since then, the ISC (<http://www.isc.ac.uk/>) has issued annual global earthquake catalogs with a time lag of about two years [123].

### Technical Considerations

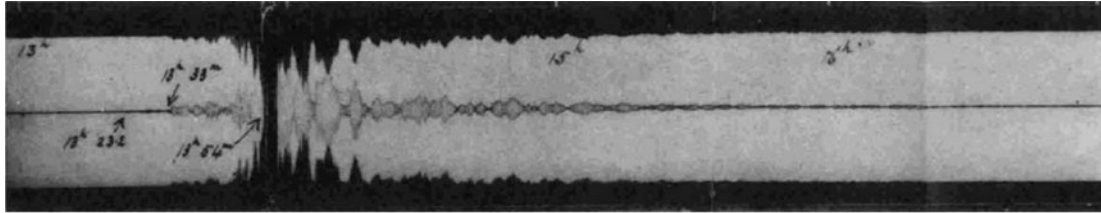
To record seismic waves, we must consider both the available technology for designing seismographs, and the nature of the Earth’s background noise [121]. The Earth is constantly in motion. This “background” noise is usually classified as either (1) *microseisms*, which typically have frequencies below about 1 Hz, are often the largest background signals, and are usually caused by natural disturbances (largely caused by ocean waves near shorelines); or (2) *microtremors*, which have frequencies higher



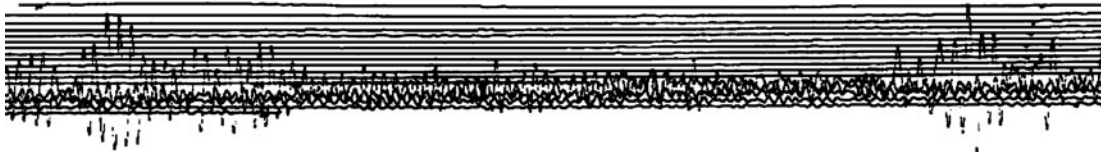
Earthquake Monitoring and Early Warning Systems, Figure 5  
Some sample analog seismograms recorded on smoked paper

than about 1 Hz, and are due to human activities (such as traffic and machinery) and local natural sources (such as wind and moving vegetation). Ground motions from earthquakes vary more than ten orders of magnitude in amplitude and six orders of magnitude in frequency, depending on the size of the earthquake and the distance at which it is recorded. Figure 7 illustrates the relative dynamic range of some common seismometers for global earthquake monitoring. A “low Earth noise” model [10,92] is the lower limit of Earth’s natural noise in its quietest locations – it is desirable to have instruments that are sensitive enough to detect this minimal background Earth

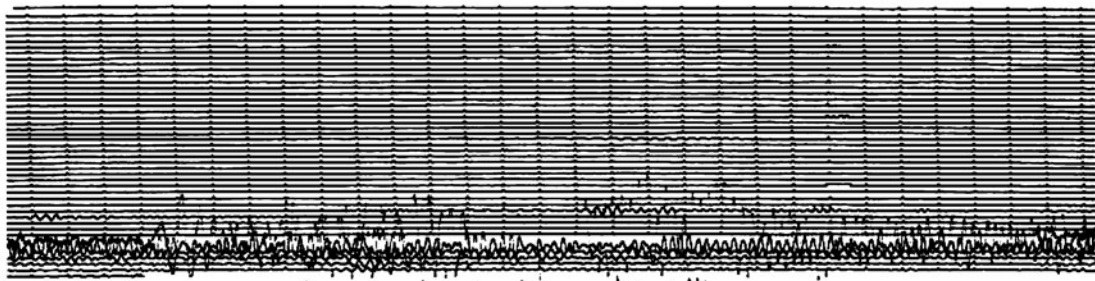
signal. In the analog instrument era (i. e., prior to about 1980), short-period and long-period seismometers were designed separately to avoid microseisms, which have predominant periods of about 6 s. Short-period seismometers were designed to detect tiny ground motions from smaller, nearby earthquakes, while long-period instruments were designed to recover the motions of distant, larger earthquakes (“teleseisms”). Additionally, strong-motion accelerometers, generally recording directly onto 70 mm-wide film strips, were used to measure large motions from nearby earthquakes. In today’s much more capable digital instrumentation, two major types of instru-



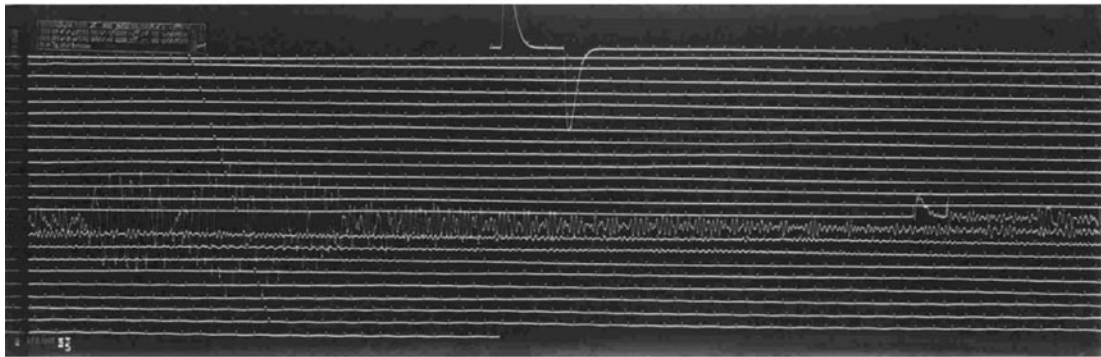
PAISLEY, SCOTLAND. Milne Seismograph. (From photographic copy.)



De Bilt, the Netherlands. Galitzin Seismograph.



Weston Observatory, USA. Benioff Seismograph



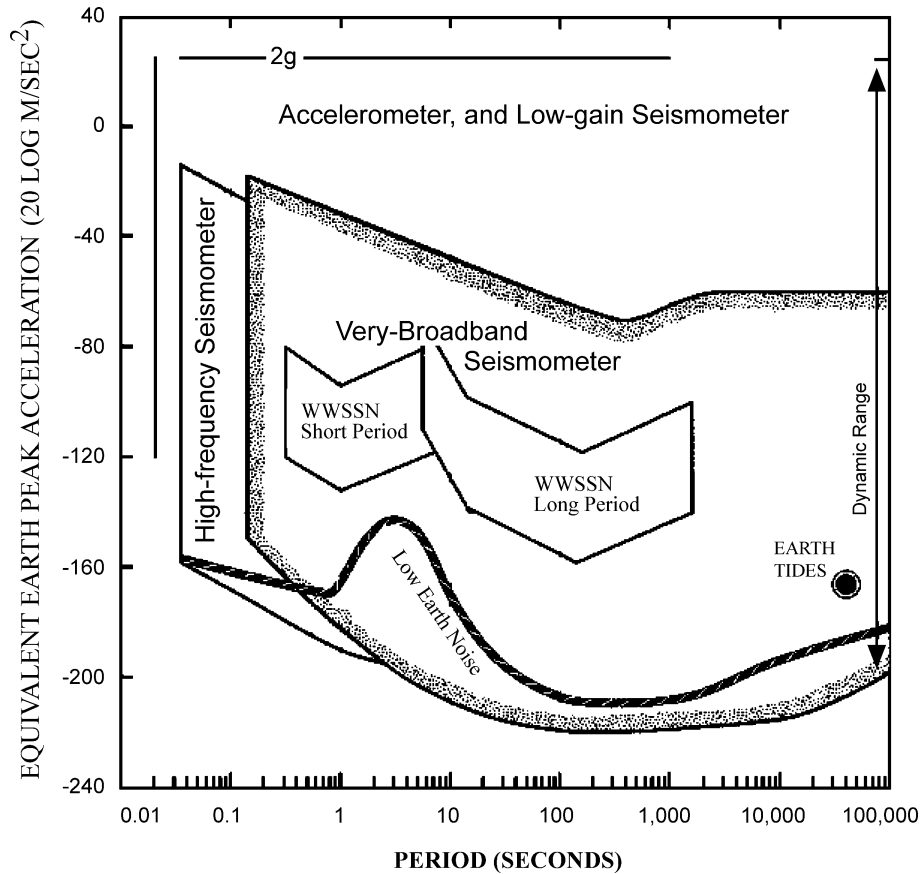
San Juan, Puerto Rico. WWSSN Long-Period Seismograph.

Earthquake Monitoring and Early Warning Systems, Figure 6  
Some sample analog seismograms recorded on photographic paper or film

ments are deployed: (1) “broadband” seismometers, which replace and improve upon both short-period and long-period seismometers, and (2) strong-motion accelerometers for high-amplitude, high-frequency, seismic waves from local earthquakes, which often drive broadband seismometers off scale. While rare examples of the old ana-

log instruments are still in use, the vast majority of instruments presently operating are digital.

In addition to having large variations in amplitudes and frequencies, seismic waves from earthquakes also attenuate rapidly with distance, that is, they lose energy as they travel, particularly at higher frequencies. We must



Earthquake Monitoring and Early Warning Systems, Figure 7

Relative dynamic range of some common seismometers for global earthquake monitoring (modified from Fig. 1 in [54]). The Y-axis is marked in decibel (dB) where  $\text{dB} = 20 \log(A/A_0)$ ;  $A$  is the signal amplitude, and  $A_0$  is the reference signal amplitude

consider these effects in order to monitor seismic waves effectively.

In 1935, C.F. Richter introduced the concept of *magnitude* to classify local earthquakes by their “size”, effectively the amount of energy radiated at the actual rupture surface within the Earth. See the entry by Bormann and Saul ► [Earthquake Magnitude](#) for a discussion of the various magnitude scales in use. While every effort is made to make these different scales overlap cleanly, each has strengths and weaknesses that make one or another preferable in a given situation. Probably the most general and robust of these methods is called a “moment magnitude”, symbolized as  $M_W$ . Existing instruments and environments are such that the smallest natural earthquakes we routinely observe close by are about magnitude = 1. The largest earthquake so far recorded by instrumentals is the  $M_W = 9.5$  Chilean earthquake in 1960. In 1941, B. Gutenberg and C.F. Richter discovered that over large

geographic regions the rate of earthquake occurrence is empirically related to their magnitudes by:

$$\log N = a - bM \quad (1)$$

where  $N$  is the number of earthquakes of magnitude  $M$  or greater, and  $a$  and  $b$  are numerical constants. It turns out that  $b$  is usually about 1, which implies that  $M = 6$  earthquakes are about ten times more frequent than  $M = 7$  earthquakes. Engdahl and Villasenor [24] show that there has been an *average* of about 15 major ( $M \geq 7$ ) earthquakes per year over the past 100 years, and about 150 large ( $M \geq 6$ ) earthquakes per year during this same time interval. Strong ground motions (above 0.1 g in acceleration) over sizeable areas are generated by  $M \geq 6$  earthquakes; these are potentially damaging levels of ground shaking.

Earthquakes are classified by magnitude ( $M$ ) as *major* if  $M \geq 7$ , as *moderate to large* if  $M$  ranges from 5 to 7, as

*small* if  $M$  ranges from 3 to 5, as *micro* if  $M < 3$ , and as *nano* if  $M < 0$ . An earthquake with  $M \geq 7\ 3/4$  is often called *great*, and if  $M \geq 9$ , *mega*.

### Earthquake Monitoring in the Digital Era

Figure 8 shows the expected amplitudes of seismic waves by earthquake magnitude. The top frame is a plot of the equivalent peak ground acceleration versus frequency. The two heavy curves denote the “minimum Earth noise”, and the “maximum Earth noise” (i. e., for seismographic station located in the continental interior versus near the coast).

The two domains of the WWSSN equipment, short-period long-period seismometers are shown as gray shading. The domains for two other instruments, SRO (Seismic Research Observatories Seismograph) and IDA (International Deployment of Accelerometers), are also shown; these were the early models of the current instruments now in operation in the *Global Seismographic Network* (GSN). The bottom two frames indicate expected amplitudes of seismic waves from earthquakes of a range of magnitudes (we use the moment magnitude,  $M_W$ ). For simplicity, we consider two cases: (bottom left) global earthquakes recorded at a large distance with a seismographic network spaced at intervals of about 1000 km, and (bottom right) local earthquakes recorded at short distances with a seismic array spaced at intervals of about 50 km. In the bottom left plot, the global-scale network, the expected amplitudes of  $P$ -wave and surface wave at 3000 km from the earthquake source are shown; for the bottom right plot, a local seismic array, the expected amplitudes of  $S$ -wave at 10 km and 100 km from the earthquake source are shown. Seismologists use this and similar figures in planning seismographic networks. Local noise surveys are usually conducted as well when designing specific seismographic networks.

With advances in digital technology, earthquake monitoring entered the digital era in the 1980s. Older analog equipment was gradually phased out as modern digital equipment replaced it [54]. The WWSSN was replaced by the *Global Seismographic Network* (GSN), a collaboration of several institutions under the IRIS consortium (<http://www.iris.edu/>). The goal of the GSN (<http://www.iris.edu/about/GSN/index.htm>) is “to deploy over 128 permanent seismic recording stations uniformly over the Earth’s surface”. The GSN project provides funding for two network operators: (1) the IRIS/ASL Network Operations Center, in Albuquerque, New Mexico (operated by the US Geological Survey), and (2) the IRIS/IDA Network Operations Center in La Jolla, California (operated by personnel from

the Scripps Institution of Oceanography). Components of a modern IRIS GSN seismograph system, which include broadband seismometers, accelerometers, and recording equipment, are shown in Fig. 9.

Figure 10 shows the station map of the Global Seismographic Network as of 2007. IRIS GSN stations continuously record seismic data from very broad band seismometers at 20 samples per second (sps), and also include high-frequency (40 sps) and strong-motion (1 and 100 sps) sensors where scientifically warranted. It is the goal of the GSN project to provide real-time access to its data via Internet or satellite. Since 1991, the IRIS Data Management Center has been providing easy access to comprehensive seismic data from the GSN and elsewhere [1].

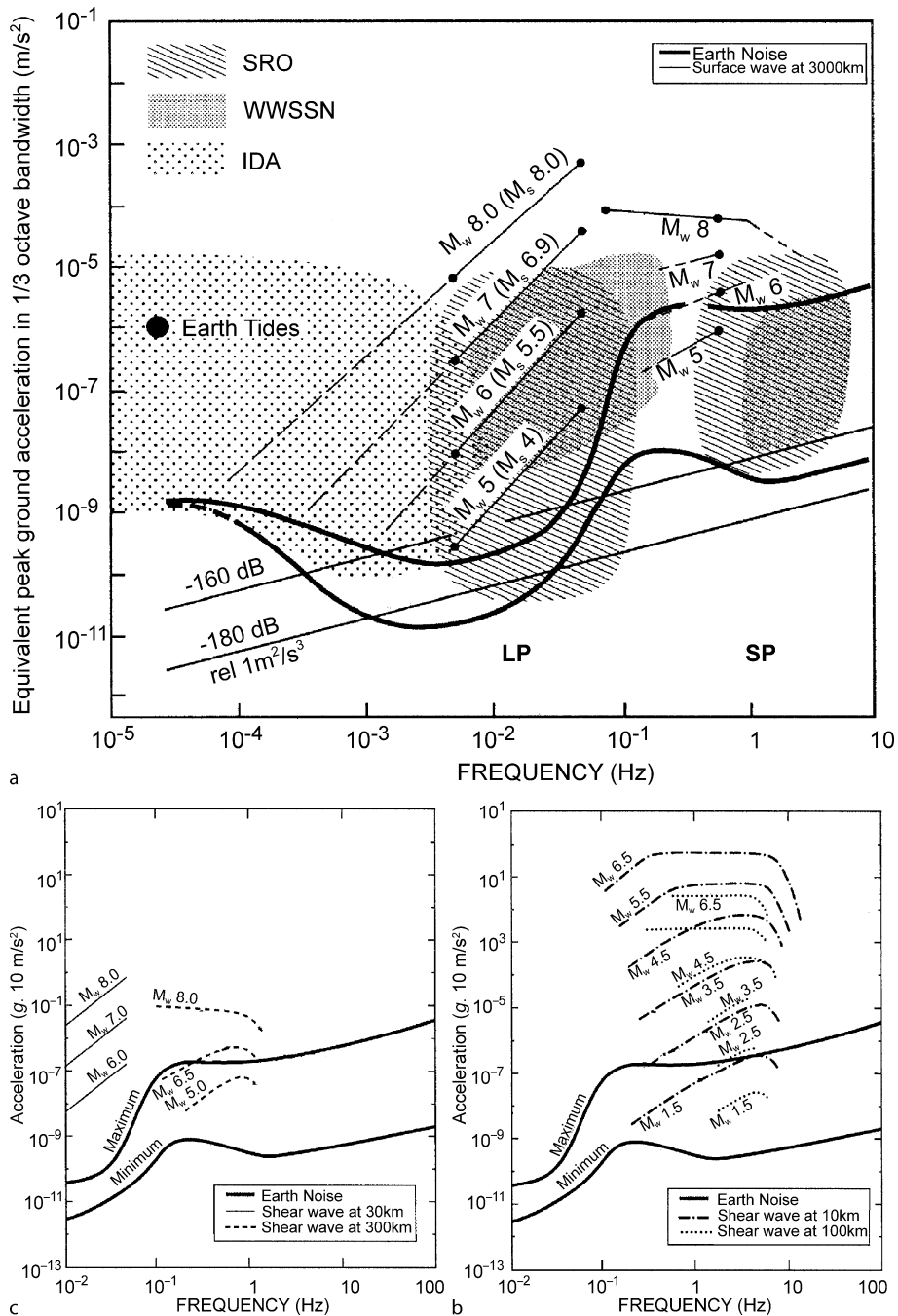
### Earthquake Monitoring: Regional and Local Networks

A major development in earthquake monitoring was the establishment of seismographic networks optimized to record the many *frequent* but *smaller* regional and local earthquakes occurring in many locations. To observe as many of these nearby earthquakes as possible, inexpensive seismographs with high magnifications and low dynamic-range telemetry are used to record the smallest earthquakes feasible with current technology and local background noise. As a result, the recorded amplitudes often overdrive the instruments for earthquakes with  $M \gtrsim 3$  within about 50 km of such seismographs. This is not a serious defect, since the emphasis for these networks is to obtain as many first arrival times as possible, and to detect and to locate the maximum number of earthquakes. Because seismic waves from small earthquakes are quickly attenuated with increasing distance, it is also necessary to deploy many instruments at small station spacing (generally from a few to a few tens of kilometers), and to cover as large a territory as possible in order to record at least a few earthquakes every week. Since funding often is limited, these local and regional seismic networks are commonly optimized for the largest number of stations rather than for the highest quality data.

### A Brief History

In the 1910s, the Carnegie Institution of Washington D.C. (CIW) was spending a great deal of money building the world’s then largest telescope (100 inch) at Mount Wilson Observatory, southern California [38]. Since astronomers were concerned about earthquakes that might disturb their telescopes, Harry O. Wood was able to persuade CIW to support earthquake investigations, and as a result, a regional network of about a dozen Wood–Anderson seis-



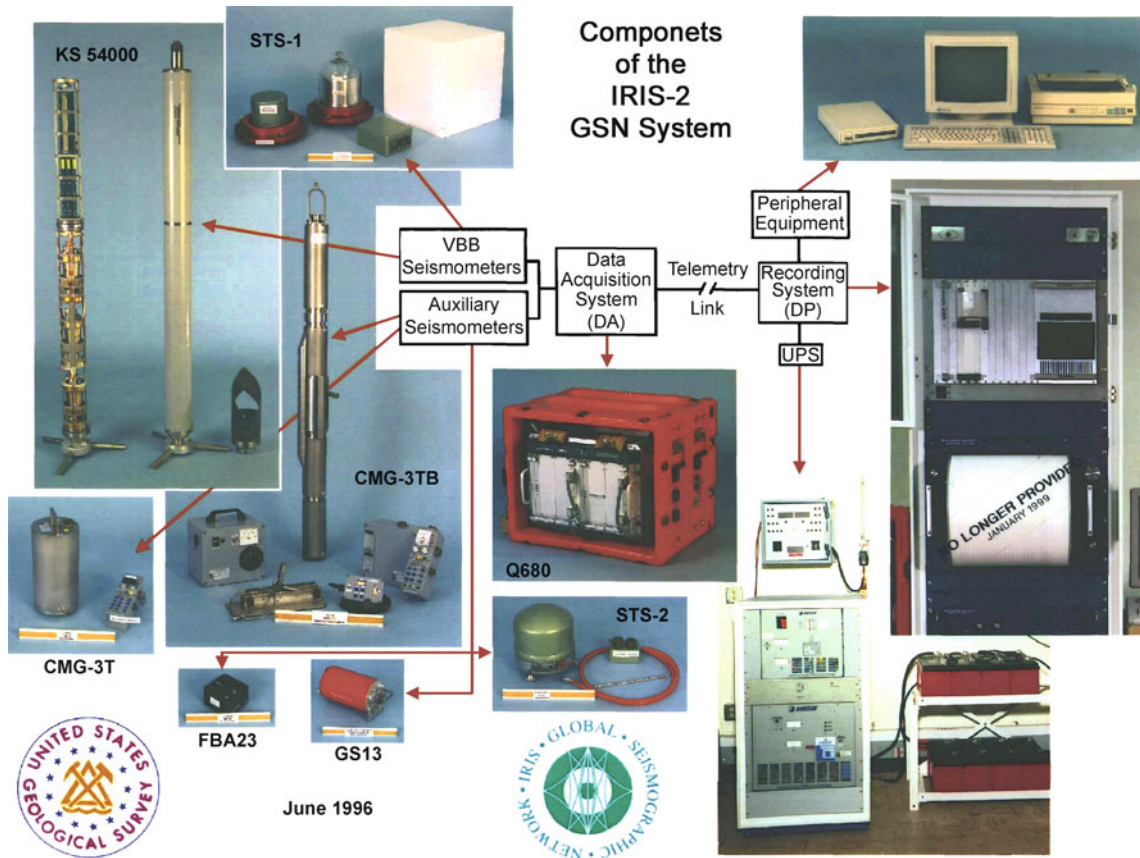


Earthquake Monitoring and Early Warning Systems, Figure 8  
 Expected amplitudes of seismic waves by earthquake magnitude. See text for explanations

mographs was established in southern California in the 1920s. See Goodstein [38] for the early history leading to the establishment of the California Institute of Technology (Caltech) and its Seismological Laboratory. Astronomers

played important roles in getting seismic monitoring established in various other regions of the world as well.

Regional networks using different types of seismographs were established in many countries about this time,



Earthquake Monitoring and Early Warning Systems, Figure 9  
 Components of the IRIS-2 GSN System: broadband seismometers, accelerometers and recording equipment

such as in Japan, New Zealand, and the USSR and its allies. In the 1960s, high-gain, short-period, telemetered networks were developed to study microearthquakes. To support detailed studies of local earthquakes and especially for the purpose of earthquake prediction, over 100 microearthquake networks were established in various parts of the world by the end of the 1970s [74]. These microearthquake networks comprised from tens to hundreds of short-period seismometers, generally with their signals telemetered into central recording sites for processing and analysis. High magnification was achieved through electronic amplification, permitting recording of very small earthquakes (down to  $M = 0$ ), though this came at the expense of saturated records for earthquakes of  $M \gtrsim 3$  within about 50 km. Unfortunately, the hope of discovering some sort of earthquake precursor from the data obtained by these microearthquake networks did not work out. For a review of the earthquake prediction efforts, please read Kanamori [60].

### Some Recent Advances

Because of recent advances in electronics, communications, and microcomputers, it is now possible to deploy sophisticated digital seismograph stations at global, national, regional, and local scales for *real-time* seismology [64]. Many such networks, including temporary portable networks, have been implemented in many countries. In particular, various real-time and near real-time seismic systems began operation in the 1990s: for example, in Mexico [25], California [32,47], and Taiwan [110]. The Real-Time Data (RTD) system operated by the Central Weather Bureau (CWB) of Taiwan is based on a network of telemetered digital accelerographs [102]; since 1995, this system has used pagers, e-mail, and other techniques to automatically and rapidly disseminate information about the hypocenter, magnitude, and shaking amplitude of felt earthquakes ( $M \gtrsim 4$ ) in the Taiwan region. The disastrous Chi-Chi earthquake ( $M_W = 7.6$ ) of 20

### Global Seismographic Network



Earthquake Monitoring and Early Warning Systems, Figure 10  
Station map of the Global Seismographic Network (GSN) as of 2007

September 1999 caused 2,471 deaths and total economic losses of US\$ 11.5 billion. For this earthquake sequence, the RTD system delivered accurate information to government officials 102 seconds after the origin time of the main shock (about 50 seconds for most aftershocks), and proved to be useful in the emergency response of the Taiwan government [37,131].

#### Recording Damaging Ground Shaking

Observing teleseisms on a global scale with station spacing of several hundreds of kilometers does not yield critical information about near-source strong ground shaking required for earthquake structural engineering purposes and seismic hazard reduction. Broadband seismometers, which are optimized to record earthquakes at great distances, do not perform well in the near-field of a major

earthquake. For example, during the 1999 Chi-Chi earthquake the nearest broadband station in Taiwan (epicentral distance of about 20 km) was badly overdriven, recorded no useful data beyond the arrival time of the initial *P*-wave, and finally failed about one minute into the shock.

A regional seismic network with station spacing of a few tens of kilometers cannot do the job either: the station spacing is still too large and the records are typically overdriven for earthquakes of  $M \gtrsim 3$  (any large earthquake would certainly overdrive these sensitive instruments in the entire network). In his account of early earthquake engineering, Housner [51] credited John R. Freeman, an eminent engineer, with persuading the then US Secretary of Commerce to authorize a strong-motion program, and, in 1930, the design of an accelerograph for engineering purposes. In a letter to R.R. Martel, Housner's professor at Caltech, Freeman wrote:

*I stated that the data which had been given to structural engineers on acceleration and limits of motion in earthquakes as a basis for their designs were all based on guesswork, that there had never yet been a precise measurement of acceleration made. That of the five seismographs around San Francisco Bay which tried to record the earthquake of 1906 not one was able to tell the truth.*

Strong-motion recordings useful to engineers must be on-scale for damaging earthquakes and, in particular, from instruments located on or near built structures in densely urbanized environments, within about 25 km of the earthquake-rupture zone for sites on rock, or within about 100 km for sites on soft soils. Recordings of motions sufficient to cause damage at sites at greater distances are also of interest for earthquake engineering in areas likely to be affected by major subduction-zone earthquakes and in areas with exceptionally low attenuation rates [11]. In addition, densely-spaced networks of strong-motion recorders are needed to study the large variations in these motions over short distances [26,29].

Although several interesting accelerograms were recorded in southern California in the 1930s and 1940s, most seismologists did not pursue strong-motion monitoring until much later. The 1971 San Fernando earthquake emphatically demonstrated the need for more strong-motion data [9]. Two important programs emerged in the United States – the National Strong-Motion Program (<http://nsmp.wr.usgs.gov/>), and the California Strong Motion Instrumentation Program (<http://docinet3.consrv.ca.gov/csmip/>). However, the budgets for these programs were and continue to be small in comparison to other earthquake programs. High levels of funding for strong-motion monitoring, comparable to that of the GSN and the regional seismic networks, became available in Taiwan in the early 1990s, and in Japan in the mid-1990s. The Consortium of Organizations for Strong-Motion Observation Systems (<http://www.cosmos-eq.org/>) was established recently to promote the acquisition and application of strong-motion data.

### Seismograms and Derived Products

Even before instruments were developed to record seismic waves from earthquakes, many scholars compiled catalogs of earthquake events noted in historical and other documents. Robert Mallet in 1852–1854 published the first extensive earthquake catalog of the world (1606 B.C.–A.D. 1842) totaling 6831 events [79]. Based on this compilation, Mallet prepared the first significant seismicity map of the Earth in 1858. Mallet's map is remarkable in that

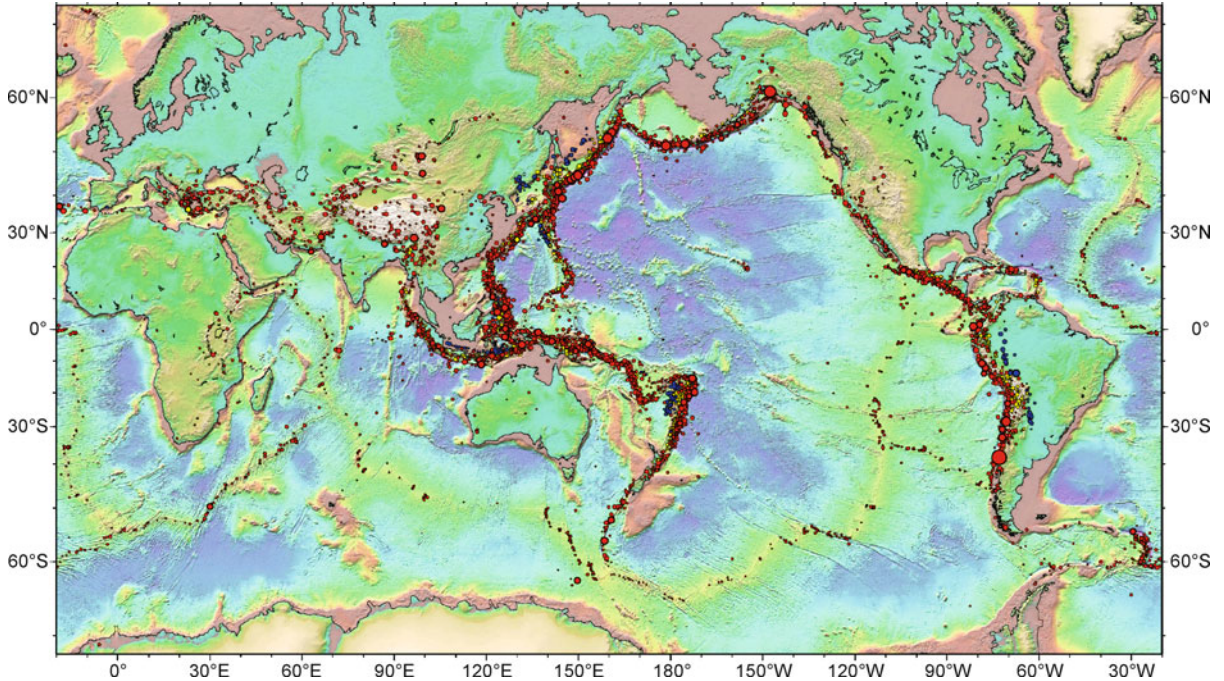
it correctly identifies the major earthquake zones of the Earth excepting for parts of the oceans. Although Mallet's earthquake catalog and similar compilations contain a wealth of information about earthquakes, they were made without the aid of instruments, and thus were subject to the biases of the observers as well as to population distributions. These non-instrumental earthquake catalogs also contain errors because the source materials were commonly incomplete and inconsistent regarding date, time, place names, and reported damage. Ambraseys et al. [8] discusses these difficulties for a regional case and Guidoboni [42] addresses the matter in general.

Today, seismograms are the fundamental data produced by earthquake monitoring. An analyst's first task is to find out when and where the earthquakes occurred, its size, and other characteristics. The accuracy of determining earthquake parameters, as well as the number of parameters used to characterize an earthquake, has progressed along with the availability of seismograms and computers, as well as advances in seismology. In the analog era, earthquake parameters were primarily the origin time, geographical location (epicenter), focal depth, and magnitude. A list of these parameters for earthquakes occurring over some time interval is called an *earthquake catalog*. A useful and common illustration of such results is a map showing the locations of earthquakes by magnitude (a seismicity map). Figure 11 is such a seismicity map for 1900–1999 as prepared by Engdahl and Villaseñor [24]. The map shows that moderate and large earthquakes are concentrated in tectonic active areas while most areas of the Earth are aseismic.

### Earthquake Location

Several methods have been developed to locate earthquakes (i.e., determine origin time, latitude and longitude of the epicenter, and focal depth). Common to most of these methods is the use of arrivals times of initial *P*- and *S*-waves. In particular, Geiger [33] applied the Gauss–Newton method to solve for earthquake location, which is a nonlinear problem, by formulating it as an inverse problem. However, since Geiger's method is computational intensive, it was not practical to apply it for the routine determinations of earthquake hypocenters until the advance of modern computers in the early 1960s.

Before computers became widely available starting in the 1960s, earthquakes were usually located by a manual, graphical method. In any location method, we assume that an earthquake is a point source and its sole parameters are origin time (time of occurrence,  $t_o$ ) and hypocenter position ( $x_o, y_o, z_o$ ). If both *P*- and *S*-arrival



Earthquake Monitoring and Early Warning Systems, Figure 11  
Seismicity of the Earth: 1900–1999 (see [24] for details)

times are available, one may use the time intervals between  $P$ - and  $S$ -waves at each station ( $S$ - $P$  times) and estimates of seismic wave velocities in the Earth to obtain a rough estimate of the epicentral distance,  $D$ , from that station:

$$D = [V_P V_S / (V_P - V_S)](T_S - T_P) \quad (2)$$

where  $V_P$  is the  $P$ -wave velocity,  $V_S$  the  $S$ -wave velocity,  $T_S$  the  $S$ -wave arrival time, and  $T_P$  the  $P$ -wave arrival time. For a typical crustal  $P$ -wave velocity of 6 km/s, and  $V_P/V_S \approx 1.8$ , the distance  $D$  in kilometers is about 7.5 times the  $S$ - $P$  interval measured in seconds. If three or more epicentral distances are available, the epicenter may be placed at the intersection of circles with the stations as centers and the appropriate  $D$  as radii. The intersection will seldom be a point, and its areal extent gives a rough estimate of the uncertainty of the epicenter and hypocentral (focal) depth. In the early days, the focal depth was usually assumed or occasionally determined using a “depth phase” (generally, a ray that travels upward from the hypocenter and reflects back from the Earth’s surface, then arcs through the Earth to reach a distant seismograph).

Although Geiger [33] presented a method for determining the origin time and epicenter, the method can be extended easily to include focal depth. To locate an earth-

quake using a set of arrival times,  $\tau_k$ , from stations at positions  $(x_k, y_k, z_k)$ ,  $k = 1, 2, \dots, m$ , we must assume a model of seismic velocities from which theoretical travel times,  $T_k$  for a trial hypocenter at  $(x^*, y^*, z^*)$  to the stations can be computed. Let us consider a given trial origin time and hypocenter as the trial vector  $\chi^*$  in a four-dimensional Euclidean space:

$$\chi^* = (t^*, x^*, y^*, z^*)^T \quad (3)$$

where the superscript  $T$  ( $^T$ ) denotes the vector transpose. Theoretical arrival time,  $t_k$ , from  $\chi^*$  to the  $k$ -th station is the theoretical travel time,  $T_k$ , plus the trial origin time,  $t^*$ . We now define the arrival time residual at the  $k$ -th station,  $r_k$ , as the difference between the observed and the theoretical arrival times. We may consider this set of station residuals as a vector in an  $m$ -dimensional Euclidean space and write:

$$\mathbf{r} = (r_1(\chi^*), r_2(\chi^*), \dots, r_m(\chi^*))^T. \quad (4)$$

We now apply the least squares method to obtain a set of linear equations solving for an adjustment vector,  $\delta\chi$ :

$$\mathbf{A}^T \mathbf{A} \delta\chi = -\mathbf{A}^T \mathbf{r}, \quad (5)$$

where  $A$  is the Jacobian matrix consisting of partial derivatives of travel time with respect to  $t$ ,  $x$ ,  $y$ , and  $z$ . A detailed derivation of the Geiger method is given by Lee and

Stewart (see, pp 132–134 in [74]). There are many practical difficulties in implementing Geiger’s method for locating earthquakes, as discussed by Lee and Stewart (see, pp 134–139 in [74]). Although standard errors for these earthquake locations can be computed, they are often not meaningful because errors in the measurement of arrival times usually do not obey a Gaussian probability distribution. In recent years, many authors applied various nonlinear methods to locate earthquakes; a review of these methods is given by Lomax et al. ► **Earthquake Location, Direct, Global-Search Methods.**

### Earthquake Magnitude

After an earthquake is located, the next question is: how big was it? The Richter magnitude scale was originally devised to measure the “size” of an earthquake in southern California. Richter [96] defined the local (earthquake) magnitude,  $M_L$ , of an earthquake observed at any particular station to be:

$$M_L = \log A - \log A_0(\Delta) \quad (6)$$

where  $A$  is the maximum amplitude in millimeters as recorded by a Wood–Anderson seismograph for an earthquake at epicentral distance of  $\Delta$  km. The correction factor,  $\log A_0(\Delta)$ , is the maximum amplitude at  $\Delta$  km for a “standard” earthquake. Thus, three arbitrary choices enter into the definition of local magnitude: (1) the use of the Wood–Anderson seismographs, (2) the use of the common logarithm (i. e., the logarithm to the base 10), and (3) the selection of the standard earthquake, whose amplitudes as a function of distance  $\Delta$  are represented by  $A_0(\Delta)$ .

In the 1940s, B. Gutenberg and C.F. Richter extended the local magnitude scale to include more distant earthquakes. Gutenberg [43] defined the surface-wave magnitude,  $M_S$ , as

$$M_S = \log(A/T) - \log A_0(\Delta^\circ) \quad (7)$$

where  $A$  is the maximum combined horizontal ground displacement in micrometers ( $\mu\text{m}$ ) for surface waves with a period of 20 s, and  $-\log A_0$  is tabulated as a function of epicentral distance  $\Delta$  in degrees, in a similar manner to that for the local magnitude’s  $A_0(\Delta)$ . Specifically, surface-wave magnitude is calculated from

$$M_S = \log A + 1.656 \log \Delta + 1.818 \quad (8)$$

using the prominent 20 s period surface waves commonly observed on the two horizontal-component seismograms from earthquakes of shallow focal depth.

Both magnitude scales were derived empirically and have scale-saturation problems, e. g., for very large earthquakes above a certain size the computed magnitudes of a particular type are all about the same. After the pioneering work of Charles F. Richter and Beno Gutenberg, numerous authors have developed alternative magnitude scales, as reviewed recently by Utsu [116] and by Bormann and Saul ► **Earthquake Magnitude.** A current magnitude scale widely accepted as “best” (as having the least saturation problem and being a close match to an earthquake’s total release of stress and strain) is the “moment magnitude”,  $M_W$ , computed from an earthquake’s “moment tensor”.

### Quantification of the Earthquake Source

As pointed out by Kanamori [59], it is not a simple matter to find a single measure of the “size” of an earthquake, simply because earthquakes result from complex physical processes. The elastic rebound theory of Harry F. Reid suggests that earthquakes originate from spontaneous slippage on active faults after a long period of elastic strain accumulation [94]. Faults may be considered the slip surfaces across which discontinuous displacement occurs in the Earth, while the faulting process may be modeled mathematically as a shear dislocation in an elastic medium (see [100], for a review). A shear dislocation (or slip) is equivalent to a double-couple body force [15,81]. The scaling parameter of each component couple of a double-couple body force is its *moment*. Using the equivalence between slip and body forces, Aki [2] introduced the *seismic moment*,  $M_0$ , as:

$$M_0 = \mu \int D(A) dA = \mu s A \quad (9)$$

where  $\mu$  is the shear modulus of the medium,  $A$  is the area of the slipped surface or source area, and  $s$  is the slip  $D(A)$  averaged over the area  $A$ . If an earthquake produces surface faulting, we may estimate its rupture length,  $L$ , and its average slip,  $s$ , from measurement of that faulting. The area  $A$  may be approximated by  $Lh$ , where  $h$  is the focal depth (it is often, but not always, found that the hypocenter is near the bottom of the rupture surface). A reasonable estimate for  $\mu$  is  $3 \times 10^{11}$  dynes/cm<sup>2</sup>. With these quantities, we can estimate the seismic moment from Eq. (9).

Seismic moment also can be estimated independently from seismograms. From dislocation theory, the seismic moment can be related to the far-field seismic displacement recorded by seismographs. For example, Hanks and Wyss [46] showed that

$$M_0 = (\Omega_0/\psi_{\theta\phi}) 4\pi\rho Rv^3 \quad (10)$$

where  $\Omega_0$  is the long-period limit of the displacement spectrum of either  $P$  or  $S$  waves,  $\psi_{\theta\phi}$  is a function accounting for the body-wave radiation pattern,  $\rho$  is the density of the medium,  $R$  is a function accounting for the geometric spreading of body waves, and  $v$  is the body-wave velocity. Similarly, seismic moment can be determined from surface waves or coda waves [2,3].

In 1977, Hiroo Kanamori recognized that a new magnitude scale can be developed using seismic moment ( $M_0$ ) by comparing the earthquake energy and seismic moment relation

$$E_S = (\Delta\sigma/2\mu)M_0, \quad (11)$$

where  $\Delta\sigma$  is the stress drop and  $\mu$  is the shear modulus, with the surface-wave magnitude and energy relation [45]

$$\log E_S = 1.5M_S + 11.8, \quad (12)$$

where  $E_S$  and  $M_0$  are expressed in ergs and dyne-cm, respectively. The average value of  $(\Delta\sigma/2\mu)$  is approximately equal to  $1.0 \times 10^{-4}$ . If we use this value in Eq. (11), we obtain

$$\log M_0 = 1.5M_S + 16.1. \quad (13)$$

It is known that  $M_S$  values saturate for great earthquakes ( $M_0$  about  $10^{29}$  dyne-cm or more) and, therefore, that Eqs. (12) and (13) do not hold for such great earthquakes. If a new moment-magnitude scale using the notation  $M_W$  is defined by

$$\log M_0 = 1.5M_W + 16.1 \quad (14)$$

then  $M_W$  is equivalent to  $M_S$  below saturation and provides a reasonable estimate for great earthquakes without the saturation problem [58]. The subscript letter  $W$  stands for the work at an earthquake fault, but soon  $M_W$  became known as the *moment magnitude*. Determining earthquake magnitude using seismic moment is clearly a better approach because it has a physical basis.

The concept of seismic moment led to the development of moment tensor solutions for quantifying the earthquake source, including its focal mechanism [35,36]; the seismic moment is just the scalar value of the moment tensor. Since the 1980s, Centroid-Moment-Tensor (CMT) solutions have been produced routinely for events with moment magnitudes ( $M_W$ ) greater than about 5.5. The CMT methodology is described by Dziewonski et al. [22] and Dziewonski and Woodhouse [20]; a comprehensive review is given in Dziewonski and Woodhouse [21]. These CMT solutions are published yearly

in the journal *Physics of the Earth and Planetary Interiors*, and the entire database is accessible online. This useful service is now performed by the Global CMT Project (<http://www.globalcmt.org/>), and more than 25,000 moment tensors have been determined for large earthquakes from 1976 to 2007. In the most recent decade, Quick CMT solutions [23] determined in near-real time have been added and are distributed widely via e-mail (<http://www.seismology.harvard.edu/projects/CMT/QuickCMTs/>).

### Limitations of Earthquake Catalogs

In addition to international efforts to catalog earthquakes on a global scale, observatories and government agencies issue more-detailed earthquake catalogs at local, regional, and national scales. However, earthquake catalogs from local to global scales vary greatly in spatial and temporal coverage and in quality, with respect to completeness and accuracy, because of the ongoing evolution of instrumentation, data processing procedures, and agency staff. An earthquake catalog, to be used for research, should have at least the following source parameters: origin time, epicenter (latitude and longitude), focal depth, and magnitude.

The International Seismological Summary and its predecessors provided compilations of arrival times and locations of earthquakes determined manually from about 1900 to 1963. Despite their limitations (notably the lack of magnitude estimates), these materials remain valuable. The first global earthquake catalog that contains both locations and magnitudes was published by Gutenberg and Richter in 1949, and was followed by a second edition in 1954 [44]. This catalog contains over 4,000 earthquakes from 1904 to 1951. Unfortunately, its temporal and spatial coverage is uneven as a result of rapid changes in seismic instrumentation, and of the interference of both World Wars. Nevertheless, the procedures used for earthquake location and magnitude estimation were the same throughout, using the arrival-time and amplitude data available to Gutenberg and Richter during the 1940s and early 1950s.

Since 1964, the International Seismological Centre has performed systematic cataloging of earthquakes worldwide by using computers and more modern seismograph networks. The spatial coverage of this catalog is not complete for some areas of the Earth (especially the oceans) because of the paucity of seismographic stations in such areas. By plotting the cumulative numbers of earthquakes above a certain magnitude versus magnitude, and using Eq. (1), the lower limit of completeness of an earthquake catalog may be estimated – it is the magnitude below which the data deviate below a linear fit to Eq. (1).

A *Centennial Earthquake Catalog* covering ISS- and ISC-reported global earthquakes from 1900–1999 was generated using an improved Earth model that takes into account regional variations in seismic wave velocities in the Earth’s crust and upper mantle [24,118]. Engdahl and Villasenor [24] also compiled existing magnitude data from various authors and suggested preferred values. However, these “preferred magnitudes” were not determined by the same procedures. At present, the Global CMT Project (<http://www.globalcmt.org/>) provides the most complete online source parameters for global earthquakes (with  $M_W > 5.5$ ), including Centroid-Moment-Tensor solutions. Although the CMT catalog starts in 1976, the improved global coverage of modern broadband digital seismographs began only in about 1990.

In summary, earthquake catalogs have been used extensively for earthquake prediction research and seismic hazard assessment since the first such catalog was produced. Reservations have been expressed about the reliability of the results and interpretations from these studies because the catalogs cover too little time and have limitations in completeness and accuracy (both random and systematic). Nevertheless, advances have been made in using earthquake catalogs to (1) study the nature of seismicity (e. g., ► [Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space](#)), (2) investigate earthquake statistics (e. g., ► [Earthquake Occurrence and Mechanisms, Stochastic Models for](#)), (3) forecast earthquakes (e. g., ► [Earthquake Forecasting and Verification](#)), (4) predict earthquakes (e. g., ► [Geo-complexity and Earthquake Prediction](#)), (5) assess seismic hazards and risk, and so forth.

### Earthquake Early Warning (EEW) Systems

With increasing urbanization worldwide, earthquake hazards pose ever greater threats to lives, property, and livelihoods in populated areas near major active faults on land or near offshore subduction zones. Earthquake early-warning systems can be useful tools for reducing the impact of earthquakes, provided that cities are favorably located with respect to earthquake sources and their citizens are properly trained to respond to the warning messages. Recent reviews of earthquake early warning systems may be found in Lee and Espinosa-Aranda [73], Kanamori [61], and Allen [6], as well as a monograph on the subject by Gasparini et al. [31].

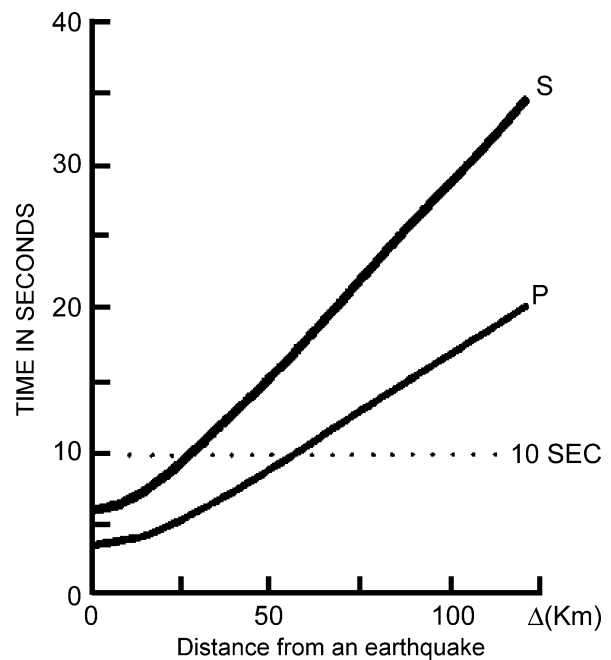
Under favorable conditions, an EEW system can forewarn an urban area of impending strong shaking with lead times that range from a few seconds to a few tens of seconds. A lead time is the time interval between issuing

a warning and the arrival of the S-waves, which are the most destructive seismic waves. Even a few seconds of advanced warning is useful for pre-programmed emergency measures at various critical facilities, such as the deceleration of rapid-transit vehicles and high-speed trains, the orderly shutoff of gas pipelines, the controlled shutdown of some high-technological manufacturing operations, the safe-guarding of computer facilities (e. g., disk-head parking), and bringing elevators to a stop at the nearest floor.

### Physical Basis and Limitations of EEW Systems

The physical basis for earthquake early warning is simple: damaging strong ground shaking is caused primarily by shear (S) and subsequent surface waves, both of which travel more slowly than the primary (P) waves, and seismic waves travel much more slowly than electromagnetic signals transmitted by telephone or radio. However, certain physical limitations must be considered, as shown by Fig. 12.

Figure 12 is a plot of the travel time for the P-wave and S-wave as a function of distance from an earthquake. We make the following assumptions about a typical destructive earthquake: (1) focal depth at  $\sim 20$  km, (2) P-wave velocity  $\sim 8$  km/s, and (3) S-wave velocity  $\sim 4.5$  km/s. If an earthquake is located 100 km from



Earthquake Monitoring and Early Warning Systems, Figure 12  
Travel time of P-waves and of S-waves versus distance for a typical earthquake



a city, the *P*-wave arrives at the city after about 13 s, and the *S*-waves in about 22 s (Fig. 12). If we deploy a dense seismic network near the earthquake source area (capable of locating and determining the size of the event in about 10 s), we will have about 3 s to issue the warning before the *P*-wave arrives, and about 12 s before the more destructive *S*-waves and surface waves arrive at the city. We have assumed that it takes negligible time to send a signal from the seismic network to the city via electromagnetic waves, which travel at about one-third the velocity of light or faster (between about 100,000 and 300,000 km/s depending on the method of transmission).

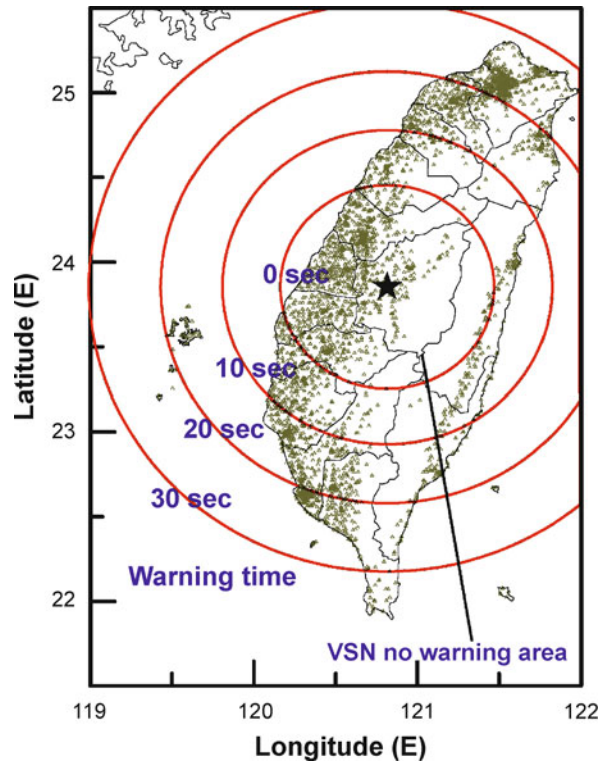
From Fig. 12 it is clear that this strategy may work for earthquakes located at least about 60 km from the urban area. For earthquakes at shorter distances ( $\sim 20$  to  $\sim 60$  km), we must reduce the time needed to detect the event and issue a warning to well under 10 s. This requirement implies that we must deploy a very dense seismic network very close to the fault and estimate the necessary parameters very fast. However, such dense networks are not economical to deploy using existing seismic instruments.

For earthquakes within 20 km of a city, there is little one can do other than installing motion-sensitive automatic shut-off devices at critical facilities (natural gas, for example) and hope that they are either very quick when responding to *S*-waves or are triggered by the onset of the *P*-wave. Normally an earthquake rupture more than  $\sim 100$  km from an urban area does not commonly pose a large threat (seismic waves would be attenuated and spread out farther). There are exceptions caused either by unusual local site conditions, such as Mexico City, or by earthquakes with large rupture zones which therefore radiate efficiently to greater distances.

### Design Considerations for EEW Systems

In the above discussion, we have assumed that one implements an earthquake early warning system with a traditional seismic network. Such EEW systems have limitation as illustrated by Fig. 13, which shows the expected early warning times for a repeat of the 1999 Chi-Chi earthquake. However, Nakamura and his colleagues have been successful in applying a single-station approach [84,99], where seismic signals are recorded and processed locally by the seismograph and an earthquake warning is issued whenever ground motions there exceed some trigger threshold. We will next discuss these two basic approaches, regional versus on-site in designing an earthquake early warning system.

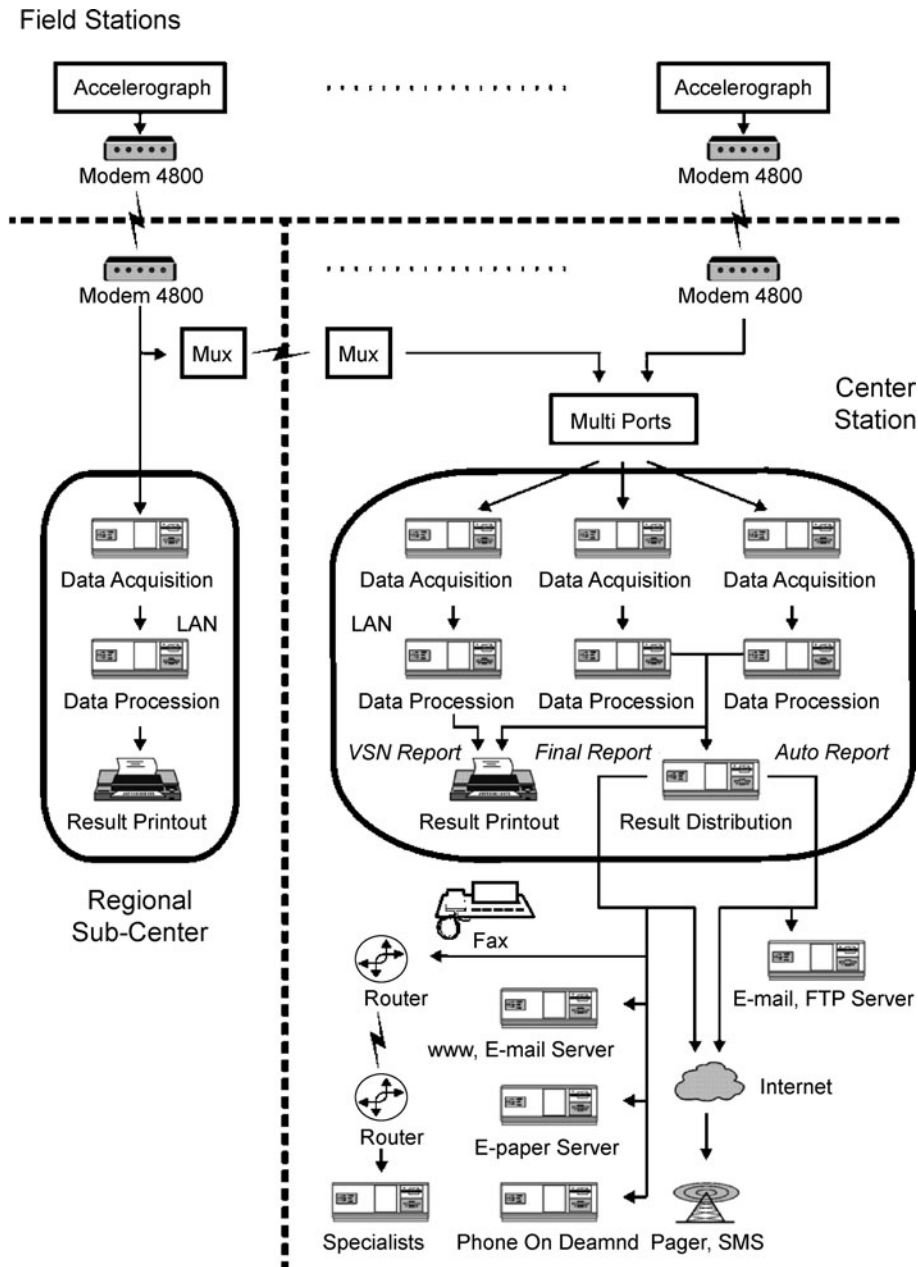
Earthquake early warning capability can be implemented through a rapid reporting system (RRS) from



Earthquake Monitoring and Early Warning Systems, Figure 13 Expected EWS early warning times (indicated by circles) in Taiwan with respect to the occurrence of an event similar to the Chi-Chi earthquake of 20 September 1999. Triangles are locations of elementary schools, which can be regarded as a good indicator for the population density of Taiwan

a traditional network, assuming real-time telemetry into the network's central laboratory. This type of system provides, to populated areas and other sensitive locations, primary event information (hypocenter, magnitude, ground shaking intensities, and potential damage) about one minute after the earthquake begins. The RRS transmits this critical information electronically to emergency response agencies and other interested organizations and to individuals. Each recipient can then take action (some of which may be pre-programmed) shortly after the earthquake begins. Response measures can include the timely dispatch of rescue equipment and emergency supplies to the likely areas of damage.

California's ShakeMap [119,120], Taiwan's CWB, and Japan's JMA systems are typical examples of RSS. In the case of the Taiwan RRS, the CWB has, since 1995, provided intensity maps, hypocenters, and magnitudes within one minute of the occurrence of  $M > 4$  earthquakes [110,128]. This system's reliability, documented by electronic messages to government agencies and scientists,



Earthquake Monitoring and Early Warning Systems, Figure 14  
 A block diagram showing the hardware of the Taiwan Earthquake Rapid Reporting System

has been close to perfect, particularly for large, damaging earthquakes. Figure 14 shows a block diagram of the Taiwan RRS, and details may be found in [128].

Using a set of empirical relationships derived from the large data set collected during the 1999 Chi-Chi earthquake, CWB now releases, within a few minutes of an event, the estimated distributions of PGA and PGV, refined magnitudes, and damage estimates [129]. This near-

real-time damage assessment is useful for rapid post-disaster emergency response and rescue missions.

### Regional Warning Versus Onsite Warning

Two approaches have been adopted for earthquake early warning systems: (1) regional warning, and (2) on-site warning. The first approach relies on traditional seismo-

logical methods in which data from a seismic network are used to locate an earthquake, determine the magnitude, and estimate the ground motion in the region involved. In the second approach, the initial ground motions (mainly *P* wave) observed at a site are used to predict the ensuing ground motions (mainly *S* and surface waves) at the same site.

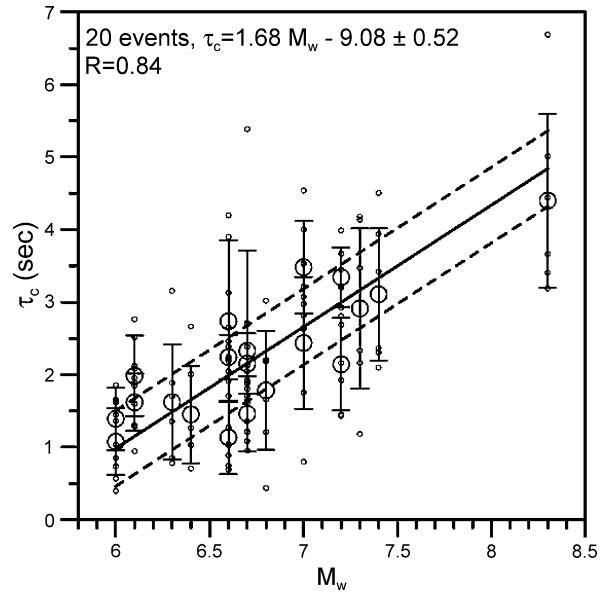
The regional approach is more comprehensive, but takes a longer time to issue an earthquake warning. An advantage of this approach is that estimates of the timing of expected strong motions throughout the affected region can be predicted more reliably. The early warning system in Taiwan is a typical example and it uses a regional warning system called virtual sub-network approach (VSN) that requires an average of 22 s to determine earthquake parameters with magnitude uncertainties of  $\pm 0.25$ . It provides a warning for areas beyond about 70 km from the epicenter (Fig. 13). This system has been in operation since 2002 with almost no false alarms [129]. With the advancement of new methodology and more dense seismic networks, regional systems are beginning to be able to provide early warnings to areas closer to the earthquake epicenter.

The regional approach has also been used in other areas. The method used in Mexico [25] is slightly different from the traditional seismological method. It is a special case of EEW system due to the relatively large distance (about 300 km in this case) between the earthquake source region (west coast of Central America) and the warning site (Mexico City). However, the warning is conceptually “regional”.

In Japan, various EEW techniques have been developed and deployed by the National Research Institute for Earth Science and Disaster Prevention (NIED) and Japan Meteorological Agency (JMA) since 2000 [49,57,89], **► Tsunami Forecasting and Warning**. In particular, JMA has started sending early warning messages to potential users responsible for emergency responses [50]. The potential users include railway systems, construction companies, and others; and they are familiar with the implications of early warning messages, as well as the technical limitations of EEW [57].

**Some Recent EEW Advances**

Allen and Kanamori [7] proposed the Earthquake Alarm System (ElarmS) to issue an earthquake warning based on information determined from the *P*-wave arrival only. Kanamori [61] extended the method of Nakamura [84] and Allen and Kanamori [7] to determine a period parameter,  $\tau_c$ , from the initial 3 s of the *P* wave.  $\tau_c$  is defined as



Earthquake Monitoring and Early Warning Systems, Figure 15  $\tau_c$  estimates from 20 events using the nearest 6 stations of the K-NET. Small open circles show single-record results, and large circles show event-average values with one standard deviation bars. Solid line shows the least squares fit to the event-average values, and the two dashed lines show the range of one standard deviation

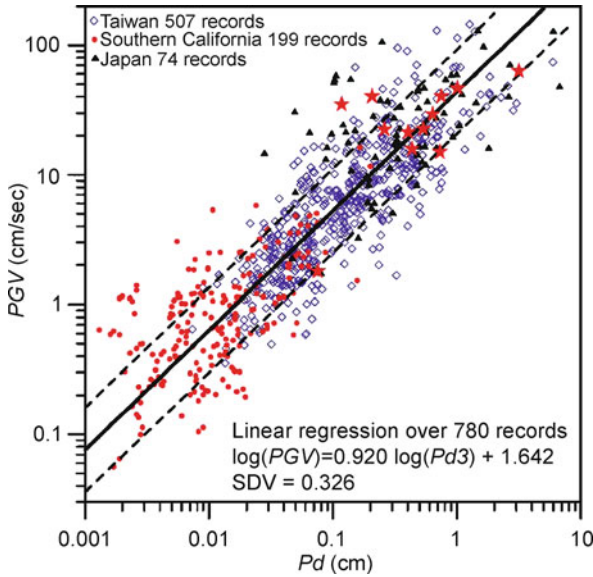
$$\tau_c = 2\pi / \sqrt{r} \tag{15}$$

where

$$r = \frac{\int_0^{\tau_0} \dot{u}^2(t) dt}{\int_0^{\tau_0} u^2(t) dt} \tag{16}$$

$u(t)$  is the ground-motion displacement;  $\tau_0$  is the duration of record used (usually 3 s), and  $\tau_c$ , which represents the size of an earthquake, can be computed from the incoming data sequentially.

The  $\tau_c$  method was used for earthquake early warning in southern California, Taiwan, and Japan by Wu and Kanamori [124,125,126] and Wu et al. [130]. At a given site, the magnitude of an event is estimated from  $\tau_c$  and the peak ground-motion velocity (PGV) from  $P_d$  (the peak amplitude of displacement in the first 3 s after the arrival of the *P* wave). The incoming 3-component signals are recursively converted to ground acceleration, velocity and displacement. The displacements are recursively filtered using an accusal Butterworth high-pass filter with a cut-off frequency of 0.075 Hz, and a *P*-wave threshold trigger is constantly monitored. When a trigger occurs,  $\tau_c$  and  $P_d$  are computed. The relationships between  $\tau_c$  and magnitude ( $M$ ), and  $P_d$  and peak ground velocity (PGV) for



Earthquake Monitoring and Early Warning Systems, Figure 16 Relationship between peak initial displacement amplitude ( $P_d$ ) measurements and peak ground velocity (PGV) for the records with epicentral distances less than 30 km from the epicenter in Southern California (red solid circles), Taiwan (blue diamonds) and Japan (black solid triangles). Solid line shows the least squares fit and the two dashed lines show the range of one standard deviation

southern California, Taiwan, and Japan were investigated. Figure 15 shows a good correlation between  $\tau_c$  and  $M_W$  from the K-NET records in Japan, and Fig. 16 shows the  $P_d$  versus PGV plot for southern California, Taiwan, and Japan. These relationships may be used to detect the occurrence of a large earthquake and provide onsite warning in the area immediately around the station where the onset of strong ground motion is expected within a few seconds after the arrival of the  $P$ -wave. When the station density is high, the onsite warning methods may be applied to data from multiple stations to increase the robustness of an onsite early warning, and to complement the regional warning approach. In an ideal situation, such warnings would be available within 10 s of the origin time of a large earthquake whose subsequent ground motion may last for tens of seconds.

Wu and Zhao [127] investigated the attenuation of  $P_d$  with the hypocentral distance  $R$  in southern California as a function of magnitude  $M$ , and obtained the following relationships:

$$M_{P_d} = 4.748 + 1.371 \times \log(P_d) + 1.883 \times \log(R) \quad (17)$$

and

$$\log(P_d) = -3.463 + 0.729 \times M - 1.374 \times \log(R). \quad (18)$$

For the regional warning approach, when an earthquake location is determined by the  $P$ -wave arrival times at stations close to the epicenter, this relationship can be used to estimate the earthquake magnitude. Their result shows that for earthquakes in southern California the  $P_d$  magnitudes agree with the catalog magnitudes with a standard deviation of 0.18 for events less than magnitude 6.5. They concluded that  $P_d$  is a robust measurement for estimating the magnitudes of earthquakes for regional early warning purposes in southern California. This method has also applied to Italian region by Zollo et al. [132] with a very good performance.

Because the on-site approach is faster than the regional approach, it can provide useful early warning to sites at short distances from the earthquake epicenter where early warning is most needed. Onsite early warning can be generated by either a single station or by a dense array. For a single station operation, signals from  $P$ -waves are used for magnitude and hypocenter determination to predict strong ground shaking. Nakamura [83] first proposed this concept, developed the Urgent Earthquake Detection and Alarm System or UrEDAS [86], and introduced a simple strong-motion index for onsite EEW [85]. However, the reliability of on-site earthquake information is generally less than that obtained with the regional warning system. There currently is a trade-off between warning time and the reliability of the earthquake information. Generally, an information updating procedure is necessary for any EEW system. On-site warning methods can be especially useful in regions where a dense seismic network is deployed.

The Japan Meteorological Agency (JMA) began distribution of earthquake early warning information to the public in October 1, 2007 through several means, such as TV and radio [50] (<http://www.jma.go.jp/jma/en/Activities/eeew.html>). The JMA system was successfully activated during the recent Noto Hanto and Niigata Chuetsu–Oki earthquakes in 2007, and provided accurate information of hypocenter, magnitude, and intensity about 3.8 s after the arrival of  $P$ -waves at nearby stations. The warning message reached sites further than about 30 km from the epicenter as an early warning alert (i. e., information arrived before shaking started at the site). This is a remarkable performance of the system for damaging earthquakes and gives promise of an early warning system as a practical means for earthquake damage mitigation. Although warning alert is most needed within 30 km of the epicenter, it is not feasible with the current density and configuration of the JMA network.

Lawrence and Cochran [68] proposed a collaborative project for rapid earthquake response and early warning by using the accelerometers that are already installed in-

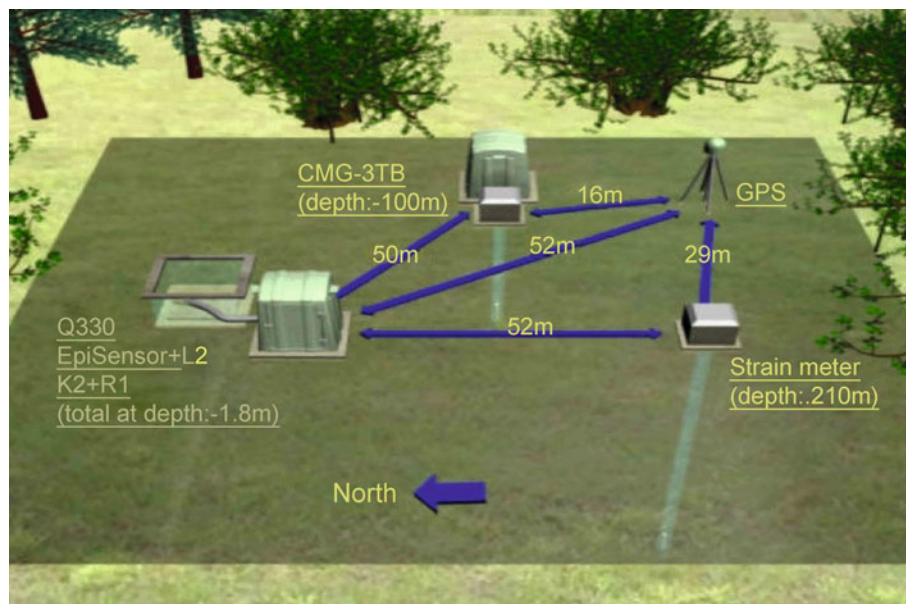
side many laptop computers. Their Quake-Catcher Network (QCN) will employ existing laptops, which have accelerometers already installed, and desktops outfitted with inexpensive (under \$ 50) USB accelerometers to form the world's largest high-density, distributed computing seismic network for monitoring strong ground motions (<http://qcn.stanford.edu/>). By freely distributing the necessary software, anyone having a computer with an Internet connection can join the project as a collaborative member. The Quake-Catcher Network also has the potential to provide better understanding of earthquakes, and the client-based software is also intended to be educational, with instructive material displaying the current seismic signal and/or recent earthquakes in the region. It is an effective way to bring earthquake awareness to students and the general public.

### Future Directions

To be successful, monitoring earthquakes requires large, stable funding over a long period of time. The most direct argument for governments to support long-term earthquake monitoring is to collect scientific data for hazard mitigation. In the past two decades about half a million of

human lives have been lost due to earthquakes, and economic losses from earthquake damage total about \$ 200 billion. Future losses will be even greater as rapid urbanization is taking place worldwide. For example, the recent Japanese Fundamental Seismic Survey and Observation Plan (costing several hundred million US dollars) is a direct response to the economic losses of about \$ 100 billion due to the 1995 Kobe earthquake. In addition to scientific and technological challenges in monitoring earthquakes, seismologists must pay attention to achieve (1) stable long-term funding, (2) effective management and execution, and (3) delivery of useful products to the users.

Seismologists benefit greatly from scientific and technological advances in other fields. For example, Global Positioning Systems (GPS) open a new window for monitoring crustal deformation which is important to understand the driving forces that generate earthquakes (► [GPS: Applications in Crustal Deformation Monitoring](#), ► [Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of](#)). Under the US Earth Scope Program (<http://www.earthscope.org/>) the Plate Boundary Observatory (PBO) is covering the western Northern America and Alaska with a network of high precision GPS and strain-meter stations in order to mea-

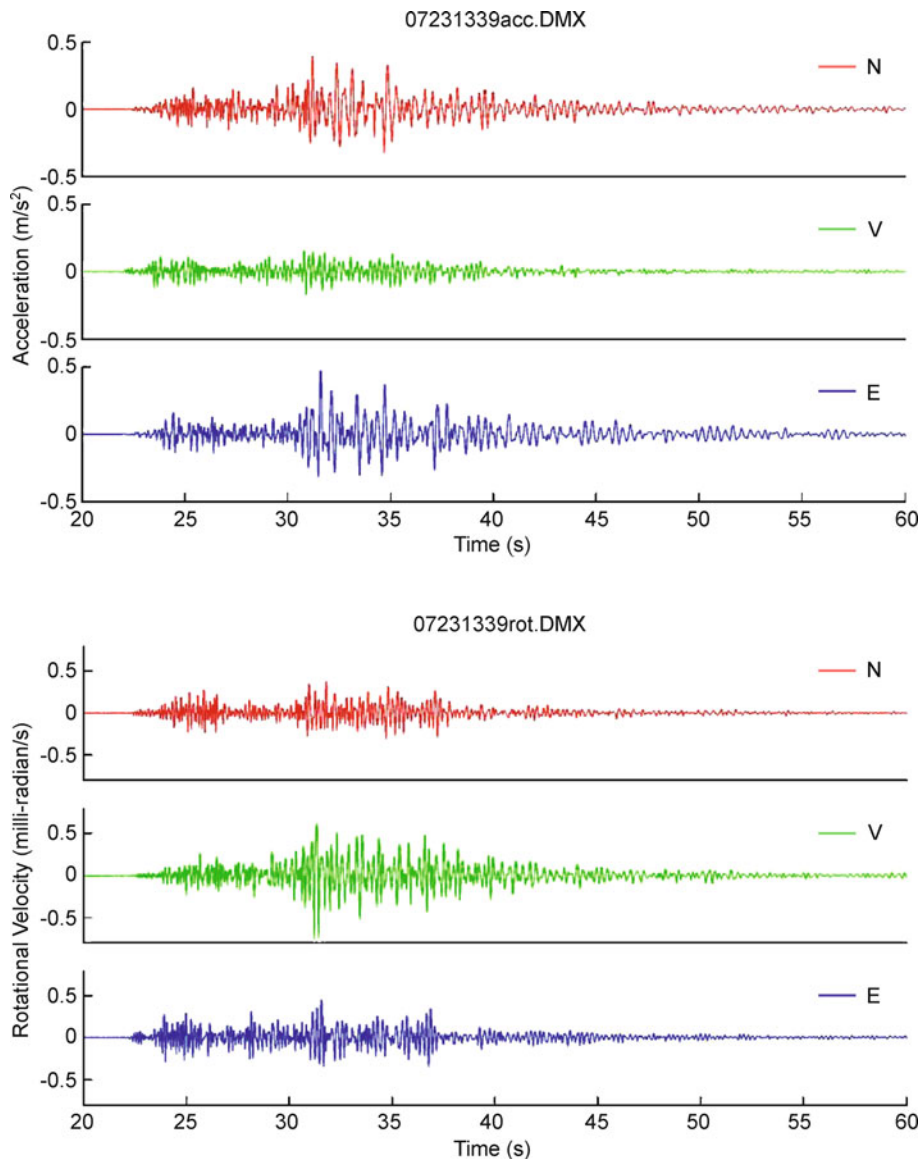


Earthquake Monitoring and Early Warning Systems, Figure 17

Instruments deployed at the HGSD station in eastern Taiwan. Clockwise from the top: (1) A broadband seismometer (Model CMG-3TB) installed at a depth of 100 m), (2) A continuous GPS instrument, (3) A strain-meter installed at a depth of 210 m), (4) A Model Q330 6-channel recorder with an accelerometer (Model EpiSensor) and a short-period seismometer (Model L2), and (5) A Model K2 6-channel accelerometer with an internal accelerometer and a rotational sensor (Model R-1)

sure deformation across the active boundary between the Pacific and North America plates (<http://www.earthscope.org/observatories/pbo>). As the sampling rate of GPS data increases, they can provide time histories of displacement during an earthquake. Monitoring earthquakes with multiple types of instruments and sensors is now increasingly popular, and “integrated” or “super” stations are increasingly common. Figure 17 shows an example of an integrated station (HGSD) in eastern Taiwan. Instruments de-

ployed at the HGSD station in eastern Taiwan include: a broadband seismometer, a continuous GPS instrument, a strain-meter, and a 6-channel accelerograph (Model K2 by Kinemetrics) with an internal accelerometer and a rotational sensor (Model R-1 by eentec). A digital seismogram recorded at the HGSD station from an earthquake ( $M_W = 5.1$ ) of July 23, 2007 at a distance of 34 km is shown in Fig. 18. The importance of rotational seismology and its current status are given in the [Appendix](#).



Earthquake Monitoring and Early Warning Systems, Figure 18

A digital seismogram recorded at the HGSD station from an earthquake ( $M_W = 5.1$ ) of July 23, 2007 at a distance of 34 km. *Top frame:* 3-component translational accelerations. *Bottom frame:* 3-component rotation velocity motions. N = North-South; V = Vertical, and E = East-West

A radically different design of seismographic networks (and earthquake early warning system in particular) is now possible using the “Sensor Network” developed by Intel Research. Intel is working with the academic community and industry collaborators to actively explore the potential of wireless sensor networks. This research is already demonstrating the potential for this new technology to enhance public safety, reduce the cost of doing business, and bring a host of other benefits to business and society ([http://www.intel.com/research/exploratory/wireless\\_sensors.htm](http://www.intel.com/research/exploratory/wireless_sensors.htm)).

It has been very difficult historically to obtain adequate and stable funding for long-term earthquake monitoring, largely because disastrous earthquakes occur infrequently. Since there are many pressing problems facing modern societies, almost all governments react to earthquake (and tsunami) disasters only after the fact, and even then for relatively short periods of time. To advance earthquake prediction research and to develop effective earthquake warning systems will require continuous earthquake monitoring with extensive instrumentations in the near-field for decades and even centuries. Therefore, innovative approaches must be developed and perseverance is needed.

### Acknowledgments

We thank John Evans, Fred Klein, Woody Savage, and Chris Stephens for reviewing the manuscript, their comments and suggestions greatly improved it. We are grateful to Lind Gee and Bob Hutt for information about the Global Seismographic Network (GSN) and for providing a high-resolution graphic file of an up-to-date GSN station map.

### Appendix:

#### A Progress Report on Rotational Seismology

Seismology is based primarily on the observation and modeling of three orthogonal components of translational ground motions. Although effects of rotational motions due to earthquakes have long been observed (e. g., [80]), Richter (see, p. 213 in [97]) stated that:

*Perfectly general motion would also involve rotations about three perpendicular axes, and three more instruments for these. Theory indicates, and observation confirms, that such rotations are negligible.*

However, Richter provided no references for this claim, and the available instruments at that time did not have the sensitivity to measure the very small rotation motions that the classical elasticity theory predicts.

Some theoretical seismologists (e. g., [4,5]) and earthquake engineers have argued for decades that the rotational part of ground motions should also be recorded. It is well known that standard seismometers and accelerometers are profoundly sensitive to rotations, particularly tilt, and therefore subject to rotation-induced errors (see e. g., [39,40,41,93]). The paucity of instrumental observations of rotational ground motions is mainly the result of the fact that, until recently, the rotational sensors did not have sufficient resolution to measure small rotational motions due to earthquakes.

Measurement of rotational motions has implications for: (1) recovering the complete ground-displacement history from seismometer recordings; (2) further constraining earthquake rupture properties; (3) extracting information about subsurface properties; and (4) providing additional ground motion information to engineers for seismic design.

In this Appendix, we will first briefly review elastic wave propagation that is based on the linear elasticity theory of simple homogeneous materials under infinitesimal strain. This theory was developed mostly in the early nineteenth century: the differential equations of the linear elastic theory were first derived by Louis Navier in 1821, and Augustin Cauchy gave his formulation in 1822 that remains virtually unchanged to the present day [103]. From this theory, Simeon Poisson demonstrated in 1828 the existence of longitudinal and transverse elastic waves, and in 1885, Lord Rayleigh confirmed the existence of elastic surface waves. George Green put this theory on a physical basis by introducing the concept of strain energy, and, in 1837, derived the basic equations of elasticity from the principle of energy conservation. In 1897, Richard Oldham first identified these three types of waves in seismograms, and linear elasticity theory has been embedded in seismology ever since.

In the following we summarize recent progress in rotational seismology and the need to include measurements of rotational ground motions in earthquake monitoring. The monograph by Teisseyre et al. [109] provides a useful summary of rotational seismology.

#### Elastic Wave Propagation

The equations of motion for a homogeneous, isotropic, and initially unstressed elastic body may be obtained using the conservation principles of continuum mechanics (e. g., [30]) as

$$\rho \frac{\partial^2 u_i}{\partial t^2} = (\lambda + \mu) \frac{\partial \theta}{\partial x_i} + \mu \nabla^2 u_i, \quad i = 1, 2, 3 \quad (A1)$$

and

$$\theta = \sum_j \partial u_j / \partial x_j \quad (\text{A2})$$

where  $\theta$  is the dilatation,  $\rho$  is the density,  $u_i$  is the  $i$ th component of the displacement vector  $\vec{u}$ ,  $t$  is the time, and  $\lambda$  and  $\mu$  are the elastic constants of the media. Eq. (A1) may be rewritten in vector form as

$$\rho(\partial^2 \vec{u} / \partial t^2) = (\lambda + \mu) \nabla(\nabla \cdot \vec{u}) + \mu \nabla^2 \vec{u}. \quad (\text{A3})$$

If we differentiate both sides of Eq. (A1) with respect to  $x_i$ , sum over the three components, and bring  $\rho$  to the right-hand side, we obtain

$$\partial^2 \theta / \partial t^2 = [(\lambda + 2\mu) / \rho] \nabla^2 \theta. \quad (\text{A4})$$

If we apply the curl operator ( $\nabla \times$ ) to both sides of Eq. (A3), and note that

$$\nabla \cdot (\nabla \times \vec{u}) = 0 \quad (\text{A5})$$

we obtain

$$\partial^2 (\nabla \times \vec{u}) / \partial t^2 = (\mu / \rho) \nabla^2 (\nabla \times \vec{u}). \quad (\text{A6})$$

Now Eqs. (A4) and (A6) are in the form of the classical wave equation

$$\partial^2 \Psi / \partial t^2 = v^2 \nabla^2 \Psi, \quad (\text{A7})$$

where  $\Psi$  is the wave potential, and  $v$  is the wave-propagation velocity (a pseudovector; wave slowness is a proper vector). Thus a dilatational disturbance  $\theta$  (or a compressional wave) may be transmitted through a homogenous elastic body with a velocity  $V_p$  where

$$V_p = \sqrt{[(\lambda + 2\mu) / \rho]} \quad (\text{A8})$$

according to Eq. (A4), and a rotational disturbance  $\nabla \times \vec{u}$  (or a shear wave) may be transmitted with a wave velocity  $V_s$  where

$$V_s = \sqrt{\mu / \rho} \quad (\text{A9})$$

according to Eq. (A6). In seismology, and for historical reasons, these two types of waves are called the primary ( $P$ ) and the secondary ( $S$ ) waves, respectively.

For a heterogeneous, isotropic, and elastic medium, the equation of motion is more complex than Eq. (A3), and is given by Karal and Keller [65] as

$$\begin{aligned} \rho(\partial^2 \vec{u} / \partial t^2) &= (\lambda + \mu) \nabla(\nabla \cdot \vec{u}) + \mu \nabla^2 \vec{u} \\ &+ \nabla \lambda (\nabla \cdot \vec{u}) + \nabla \mu \times (\nabla \times \vec{u}) + 2(\nabla \mu \cdot \nabla) \vec{u}. \end{aligned} \quad (\text{A10})$$

Furthermore, the compressional wave motion is no longer purely longitudinal, and the shear wave motion is no longer purely transverse. A review of seismic wave propagation and imaging in complex media may be found in the entry by Igel et al. ► [Seismic Wave Propagation in Media with Complex Geometries, Simulation of.](#)

A significant portion of seismological research is based on the solution of the elastic wave equations with the appropriate initial and boundary conditions. However, explicit and unique solutions are rare, except for a few simple problems. One approach is to transform the wave equation to the eikonal equation and seek solutions in terms of wave fronts and rays that are valid at high frequencies. Another approach is to develop through specific boundary conditions a solution in terms of normal modes [77]. Although ray theory is only an approximation [17], the classic work of Jeffreys and Bullen, and Gutenberg used it to determine Earth structure and locate earthquakes that occurred in the first half of the 20th century. It remains a principal tool used by seismologists even today. Impressive developments in normal mode and surface wave studies (in both theory and observation) started in the second half of the 20th century, leading to realistic quantification of earthquakes using moment tensor methodology [21].

### Rotational Ground Motions

Rotations in ground motion and in structural responses have been deduced indirectly from accelerometer arrays, but such estimates are valid only for long wavelengths compared to the distances between sensors (e.g., [16,34,52,88,90,104]). The rotational components of ground motion have also been estimated theoretically using kinematic source models and linear elastodynamic theory of wave propagation in elastic solids [14,69,70,111].

In the past decade, rotational motions from teleseismic and small local earthquakes were also successfully recorded by sensitive rotational sensors, in Japan, Poland, Germany, New Zealand, and Taiwan (e.g., [53,55,56,105,106,107,108]). The observations in Japan and Taiwan show that the amplitudes of rotations can be *one to two orders of magnitude greater than expected* from the classical linear theory. Theoretical work has also suggested that, in granular materials or cracked continua, asymmetries of the stress and strain fields can create rotations in addition to those predicted by the classical elastodynamic theory for a perfect continuum (► [Earthquake Source: Asymmetry and Rotation Effects](#)).

Because of lack of instrumentation, rotational motions have not yet been recorded in the near-field (within ~ 25 km of fault ruptures) of strong earthquakes (magnitude



> 6.5), where the discrepancy between observations and theoretical predictions may be the largest. Recording such ground motions will require extensive seismic instrumentation along some well-chosen active faults and luck. To this end, several seismologists have been advocating such measurements, and a current deployment in southwestern Taiwan by its Central Weather Bureau is designed to “capture” a repeat of the 1906 Meishan earthquake (magnitude 7.1) with both translational and rotational instruments.

Rotations in structural response, and the contributions to the response from the rotational components of the ground motion, have also been of interest for many decades (e.g., [78,87,98]). Recent reviews on rotational motions in seismology and on the effects of the rotational components of ground motion on structures can be found, for examples, in Cochard et al. [18] and Pillet and Virieux [93], and Trifunac [112], respectively.

### Growing Interest – The IWGoRS

Various factors have led to spontaneous organization within the scientific and engineering communities interested in rotational motions. Such factors include: the growing number of successful direct measurements of rotational ground motions (e.g., by ring laser gyros, fiber optic gyros, and sensors based on electro-chemical technology); increasing awareness about the usefulness of the information they provide (e.g., in constraining the earthquake rupture properties, extracting information about subsurface properties, and about deformation of structures during seismic and other excitation); and a greater appreciation for the limitations on information that can be extracted from the translational sensors due to their sensitivity to rotational motions e.g., computation of permanent displacements from accelerograms (e.g., [13,39,40,41,93,113]).

A small workshop on Rotational Seismology was organized by W.H.K. Lee, K. Hudnut, and J.R. Evans of the USGS on 16 February 2006 in response to grassroots interest. It was held at the USGS offices in Menlo Park and in Pasadena, California, with about 30 participants from about a dozen institutions participating via teleconferencing and telephone [27]. This event led to the formation of the *International Working Group on Rotational Seismology* in 2006, inaugurated at a luncheon during the AGU 2006 Fall Meeting in San Francisco.

The *International Working Group on Rotational Seismology* (IWGoRS) aims to promote investigations of rotational motions and their implications, and the sharing of experience, data, software and results in an open web-based environment (<http://www.rotational-seismology.org>).

It consists of volunteers and has no official status. H. Igel and W.H.K. Lee currently serve as “co-organizers”. Its charter is accessible on the IWGoRS web site. The Working Group has a number of active members leading task groups that focus on the organization of workshops and scientific projects, including: testing and verifying rotational sensors, broadband observations with ring laser systems, and developing a field laboratory for rotational motions. The IWGoRS web site also contains the presentations and posters from related meetings, and eventually will provide access to rotational data from many sources.

The IWGoRS organized a special session on *Rotational Motions in Seismology*, convened by H. Igel, W.H.K. Lee, and M. Todorovska during the 2006 AGU Fall Meeting [76]. The goal of that session was to discuss rotational sensors, observations, modeling, theoretical aspects, and potential applications of rotational ground motions. A total of 21 papers were submitted for this session, and over 100 individuals attended the oral session.

The large attendance at this session reflected common interests in rotational motions from a wide range of geophysical disciplines, including strong-motion seismology, exploration geophysics, broadband seismology, earthquake engineering, earthquake physics, seismic instrumentation, seismic hazards, geodesy, and astrophysics, thus confirming the timeliness of IWGoRS. It became apparent that to establish an effective international collaboration within the IWGoRS, a larger workshop was needed to allow sufficient time to discuss the many issues of interest, and to draft research plans for rotational seismology and engineering applications.

### First International Workshop

The *First International Workshop on Rotational Seismology and Engineering Applications* was held in Menlo Park, California, on 18–19 September 2007. This workshop was hosted by the US Geological Survey (USGS), which recognized this topic as a new research frontier for enabling a better understanding of the earthquake process and for the reduction of seismic hazards. The technical program consisted of three presentation sessions: plenary (4 papers) and oral (6 papers) held during the first day, and poster (30 papers) held during the morning of the second day. A post-workshop session was held on the morning of September 20, in which scientists of the Laser Interferometer Gravitational-wave Observatory (LIGO) presented their work on seismic isolation of their ultra-high precision facility, which requires very accurate recording of translational and rotational components of ground motions (3 papers). Proceedings of this Workshop were re-

leased in Lee et al. [75] with a DVD disc that contains all the presentation files and supplementary information.

One afternoon of the workshop was devoted to in-depth discussions on the key outstanding issues and future directions. The participants could join one of five panels on the following topics: (1) theoretical studies of rotational motions (chaired by L. Knopoff), (2) measuring far-field rotational motions (chaired by H. Igel), (3) measuring near-field rotational motions (chaired by T.L. Teng), (4) engineering applications of rotational motions (chaired by M.D. Trifunac), and (5) instrument design and testing (chaired by J.R. Evans). The panel reports on key issues and unsolved problems, and on research strategies and plans, can be found in Appendices 2.1 through 2.5 in Lee et al. [75]. Following the in-depth group discussions, the panel chairs reported on the group discussions in a common session, with further discussions among all the participants.

## Discussions

Since rotational ground motions may play a significant role in the near-field of earthquakes, rotational seismology has emerged as a new frontier of research. During the Workshop discussions, L. Knopoff asked: Is there a quadratic rotation-energy relation, in the spirit of Green's strain-energy relation, coupled to it or independent of it? Can we write a rotation-torque formula analogous to Hooke's law for linear elasticity in the form

$$L_{ij} = d_{ijkl}\omega_{kl} \quad (\text{A11})$$

where  $\omega_{kl}$  is the rotation,

$$\omega_{kl} = \frac{1}{2}(u_{k,l} - u_{l,k}). \quad (\text{A12})$$

$L_{ij}$  is the torque density; and  $d_{ijkl}$  are the coefficients of rotational elasticity? How are the  $d$ 's related to the usual  $c$ 's of elasticity? If we define the rotation vector as

$$\vec{\Omega} = \frac{1}{2}(\nabla \times \vec{u}) \quad (\text{A13})$$

we obtain

$$-\nabla_s^2 \nabla \times (\nabla \times \vec{\Omega}) = \partial^2 \vec{\Omega} / \partial t^2 - \frac{1}{2} \rho^{-1} (\nabla \times \vec{f}) \quad (\text{A14})$$

where the torque density is  $\nabla \times \vec{f}$ ,  $\vec{f}$  is the body force density, and  $\rho$  is density of the medium. This shows that rotational waves propagate with S-wave velocity and that it may be possible to store torques. Eq. (15) is essentially an extension using the classical elasticity theory.

Lakes [67] pointed out that the behavior of solids can be represented by a variety of continuum theories. In particular, the elasticity theory of the Cosserat brothers [19]

incorporates (1) a local rotation of points as well as the translation motion assumed in the classical theory, and (2) a couple stress (a torque per unit area) as well as the force stress (force per unit area). In the constitutive equation for the classical elasticity theory, there are two independent elastic constants, whereas for the Cosserat elastic theory there are six. Lakes (personal communication, 2007) advocates that there is substantial potential for using generalized continuum theories in geo-mechanics, and any theory must have a strong link with experiment (to determine the constants in the constitutive equation) and with physical reality.

Indeed some steps towards better understandings of rotational motions have taken place. For example, Twiss et al. [114] argued that brittle deformation of the Earth's crust (► **Brittle Tectonics: A Non-linear Dynamical System**) involving block rotations is comparable to the deformation of a granular material, with fault blocks acting like the grains. They realized the inadequacy of classical continuum mechanics and applied the Cosserat or micropolar continuum theory to take into account two separate scales of motions: macro-motion (large-scale average motion composed of macrostrain rate and macrospin), and micro-motion (local motion composed of microspin). A theoretical link is then established between the kinematics of crustal deformation involving block rotations and the effects on the seismic moment tensor and focal mechanism solutions.

Recognizing that rotational seismology is an emerging field, the *Bulletin of Seismological Society of America* will be publishing in 2009 a special issue under the guest editorship of W.H.K. Lee, M. Çelebi, M.I. Todorovska, and H. Igel.

## Bibliography

### Primary Literature

1. Ahern TK (2003) The FDSN and IRIS Data Management System: providing easy access to terabytes of information. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part B*. Academic Press, Amsterdam pp 1645–1655
2. Aki K (1966) Generation and propagation of G waves from the Niigata Earthquake of June 16, 1964: Part 1. A statistical analysis. *Bull Earthq Res Inst* 44:23–72; Part 2. Estimation of earthquake moment, released energy, and stress-strain drop from the G wave spectrum. *Bull Earthq Res Inst* 44:73–88
3. Aki K (1969) Analysis of the seismic coda of local earthquakes as scattered waves. *J Geophys Res* 74:6215–6231
4. Aki K, Richards PG (1980) *Quantitative Seismology*. W.H. Freeman, San Francisco
5. Aki K, Richards PG (2002) *Quantitative Seismology: Theory and Methods*, 2nd edn. University Science Books, Sausalito

6. Allen RM (2007) Earthquake hazard mitigation: New directions and opportunities. In: Kanamori H (ed) *Earthquake Seismology. Treatise on Geophysics*, vol 4. Elsevier, Amsterdam, pp 607–648
7. Allen RM, Kanamori H (2003) The potential for earthquake early warning in Southern California. *Science* 300:786–789
8. Ambraseys NN, Jackson JA, Melville CP (2002) Historical seismicity and tectonics: The case of the Eastern Mediterranean and the Middle East. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part A*. Academic Press, Amsterdam, pp 747–763
9. Anderson JG (2003) Strong-motion seismology. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part B*. Academic Press, Amsterdam pp 937–965
10. Berger J, Davis P, Ekström G (2004) Ambient earth noise: A survey of the Global Seismographic Network. *J Geophys Res* 109:B11307
11. Borchardt RD (ed) (1997) Vision for the future of the US National Strong-Motion Program, The committee for the future of the US National Strong Motion Program. US Geol Surv Open-File Rept B97530
12. Bormann P (ed) (2002) *New Manual of Seismological Observatory Practice*. GeoForschungsZentrum Potsdam [http://www.gfz-potsdam.de/bib/nmsop\\_formular.html](http://www.gfz-potsdam.de/bib/nmsop_formular.html)
13. Boroschek R, Legrand D (2006) Tilt motion effects on the double-time integration of linear accelerometers: an experimental approach. *Bull Seism Soc Am* 96:2072–2089
14. Bouchon M, Aki K (1982) Strain, tilt, and rotation associated with strong ground motion in the vicinity of earthquake faults. *Bull Seism Soc Am* 72:1717–1738
15. Burridge R, Knopoff L (1964) Body force equivalents for seismic dislocations. *Bull Seism Soc Am* 54:1875–1888
16. Castellani A, Boffi G (1986) Rotational components of the surface ground motion during an earthquake. *Earthq Eng Struct Dyn* 14:751–767
17. Chapman CH (2002) Seismic ray theory and finite frequency extensions. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part A*. Academic Press, Amsterdam, pp 103–123
18. Cochard A, Igel H, Schuberth B, Suryanto W, Velikoseltsev A, Schreiber U, Wassermann J, Scherbaum F, Vollmer D (2006) Rotational motions in seismology: theory, observation, simulation. In: Teisseyre R, Takeo M, Majewski M (eds) *Earthquake Source Asymmetry, Structural Media and Rotation Effects*. Springer, Heidelberg, pp 391–411
19. Cosserat E, Cosserat F (1909) *Theorie des Corps Deformables*. Hermann, Paris
20. Dziewonski AM, Woodhouse JH (1983) An experiment in the systematic study of global seismicity: centroid-moment tensor solutions for 201 moderate and large earthquakes of 1981. *J Geophys Res* 88:3247–3271
21. Dziewonski AM, Woodhouse JH (1983) Studies of the seismic source using normal-mode theory. In: Kanamori H, Boschi B (eds) *Earthquakes: Observation, Theory, and Interpretation*. North-Holland, Amsterdam, pp 45–137
22. Dziewonski AM, Chou TA, Woodhouse JH (1981) Determination of earthquake source parameters from waveform data for studies of global and regional seismicity. *J Geophys Res* 86:2825–2852
23. Ekström G (1994) Rapid earthquake analysis utilizes the Internet. *Comput Phys* 8:632–638
24. Engdahl ER, Villasenor A (2002) Global seismicity: 1900–1999. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part A*. Academic Press, Amsterdam, pp 665–690
25. Espinosa-Aranda JM, Jimenez A, Ibarrola G, Alcantar F, Aguilar A, Inostroza M, Maldonado S (1995) Mexico City Seismic Alert System. *Seism Res Lett* 66(6):42–53
26. Evans JR, Hamstra RH, Kündig C, Camina P, Rogers JA (2005) TREMOR: A wireless MEMS accelerograph for dense arrays. *Earthq Spectr* 21(1):91–124
27. Evans JR, Cochard A, Graizer V, Huang B-S, Hudnut KW, Hutt CR, Igel H, Lee WHK, Liu C-C, Majewski E, Nigbor R, Safak E, Savage WU, Schreiber U, Teisseyre R, Trifunac M, Wassermann J, Wu C-F (2007) Report of a workshop on rotational ground motion. US Geol Surv Open File Rep 20:2007–1145 <http://pubs.usgs.gov/of/2007/1145/>
28. Feigl KL (2002) Estimating earthquake source parameters from geodetic measurements. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part A*. Academic Press, Amsterdam, pp 607–620
29. Field EH, Hough SE (1997) The variability of PSV response spectra across a dense array deployed during the Northridge aftershock sequence. *Earthq Spectr* 13:243–257
30. Fung YC (1965) *Foundations of Solid Mechanics*. Prentice-Hall, Englewood Cliffs
31. Gasparini P, Manfredi G, Zschau J (eds) (2007) *Seismic Early Warning Systems*. Springer, Berlin
32. Gee L, Neuhauser D, Dreger D, Pasyanos M, Uhrhammer R, Romanowicz B (2003) The Rapid Earthquake Data Integration Project. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part B*. Academic Press, Amsterdam, pp 1261–1273
33. Geiger LC (1912) Probability method for the determination of earthquake epicenters from the arrival time only. *Bull St Louis Univ* 8:60–71
34. Ghayamghamian MR, Nouri GR (2007) On the characteristics of ground motion rotational components using Chiba dense array data. *Earthq Eng Struct Dyn* 36(10):1407–1429
35. Gilbert F (1971) Excitation of the normal modes of the Earth by earthquake sources. *Geophys J R Astron Soc* 22:223–226
36. Gilbert F, Dziewonski AM (1975) Application of normal mode theory to the retrieval of structural parameters and source mechanisms from seismic spectra. *Phil Trans Roy Soc Lond A* 278:187–269
37. Goltz JD, Flores PJ, Chang SE, Atsumi T (2001) Emergency response and early recovery. In: 1999 Chi-Chi, Taiwan, Earthquake Reconnaissance Report. *Earthq Spectra Suppl A* 17:173–183
38. Goodstein JR (1991) *Millikan's School: A History of the California Institute of Technology*. Norton, New York
39. Graizer VM (1991) Inertial seismometry methods. *Izv Earth Phys Akad Nauk SSSR* 27(1):51–61
40. Graizer VM (2005) Effect of tilt on strong motion data processing. *Soil Dyn Earthq Eng* 25:197–204

41. Graizer VM (2006) Tilts in strong ground motion. *Bull Seis Soc Am* 96:2090–2106
42. Guidoboni E (2002) Historical seismology: the long memory of the inhabited world. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part A*. Academic Press, Amsterdam, pp 775–790
43. Gutenberg B (1945) Amplitudes of surface waves and magnitudes of shallow earthquakes. *Bull Seism Soc Am* 35:3–12
44. Gutenberg B, Richter CF (1954) *Seismicity of the Earth*, 2nd edn. Princeton University Press, Princeton
45. Gutenberg B, Richter CF (1956) Magnitude and energy of earthquakes. *Ann Geofis* 9:1–15
46. Hanks TC, Wyss M (1972) The use of body wave spectra in the determination of seismic source parameters. *Bull Seism Soc Am* 62:561–589
47. Hauksson E, Jones LM, Shakal AF (2003) TriNet: a modern ground motion seismic network. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part B*. Academic Press, Amsterdam, pp 1275–1284
48. Havskov J, Alguacil G (2004) *Instrumentation in Earthquake Seismology*. Springer, Berlin
49. Horiuchi S, Negishi H, Abe K, Kamimura A, Fujinawa Y (2005) An automatic processing system for broadcasting earthquake alarms. *Bull Seism Soc Am* 95:708–718
50. Hoshiaba M, Kamigaichi O, Saito M, Tsukada S, Hamada N (2008) Earthquake early warning starts nationwide in Japan. *EOS. Trans Am Geophys Un* 89(8):73–74
51. Housner GW (2002) Historical view of earthquake engineering. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part A*. Academic Press, Amsterdam, pp 13–18
52. Huang BS (2003) Ground rotational motions of the 1991 Chi-Chi, Taiwan earthquake as inferred from dense array observations. *Geophys Res Lett* 30(6):1307–1310
53. Huang BS, Liu CC, Lin CR, Wu CF, Lee WHK (2006) Measuring mid- and near-field rotational ground motions in Taiwan. Poster, presented at 2006 Fall AGU Meeting, San Francisco
54. Hutt CR, Bolton HF, Holcomb LG (2002) US contribution to digital global seismograph networks. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part A*. Academic Press, Amsterdam, pp 319–322
55. Igel H, Cochard A, Wassermann J, Schreiber U, Velikoseltsev A, Dinh NP (2007) Broadband observations of rotational ground motions. *Geophys J Int* 168(1):182–197
56. Igel H, Schreiber U, Flaws A, Schuberth B, Velikoseltsev A, Cochard A (2005) Rotational motions induced by the M8.1 Tokachi-oki earthquake, September 25, 2003. *Geophys Res Lett* 32:L08309. doi:10.1029/2004GL022336
57. Kamigaichi O (2004) JMA Earthquake Early Warning. *J Japan Assoc Earthq Eng* 4:134–137
58. Kanamori H (1977) The energy release in great earthquakes. *J Geophys Res* 82:2921–2987
59. Kanamori H (1978) Quantification of earthquakes. *Nature* 271:411–414
60. Kanamori H (2003) Earthquake prediction: an overview. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part B*. Academic Press, Amsterdam, pp 1205–1216
61. Kanamori H (2005) Real-time seismology and earthquake damage mitigation. *Annual Rev Earth Planet Sci* 33:195–214
62. Kanamori H, Brodsky EE (2000) *The physics of earthquakes*. *Phys Today* 54(6):34–40
63. Kanamori H, Rivera L (2006) Energy partitioning during an earthquake. In: Abercrombie R, McGarr A, Kanamori H, Di Toro G (eds) *Earthquakes: Radiated Energy and the Physics of Faulting*. Geophysical Monograph, vol 170. Am Geophys Union, Washington DC, pp 3–13
64. Kanamori H, Hauksson E, Heaton T (1997) Real-time seismology and earthquake hazard mitigation. *Nature* 390:461–464
65. Karal FC, Keller JB (1959) Elastic wave propagation in homogeneous and inhomogeneous media. *J Acoust Soc Am* 31:694–705
66. Kisslinger C, Howell BF (2003) Seismology and physics of the Earth's interior in the USA, 1900–1960. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part B*. Academic Press, Amsterdam, p 1453
67. Lakes RS (1995) Experimental methods for study of Cosserat elastic solids and other generalized continua. In: Mühlhaus H (ed) *Continuum Models for Materials with Micro-structure*. Wiley, New York, pp 1–22
68. Lawrence JF, Cochran ES (2007) The Quake Catcher Network: Cyberinfrastructure bringing seismology into schools and homes. *American Geophysical Union, Fall Meeting 2007*, abstract #ED11C-0633
69. Lee VW, Trifunac MD (1985) Torsional accelerograms. *Int J Soil Dyn Earthq Eng* 4(3):132–139
70. Lee VW, Trifunac MD (1987) Rocking strong earthquake accelerations. *Int J Soil Dyn Earthq Eng* 6(2):75–89
71. Lee WHK (2002) Challenges in observational seismology. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part A*. Academic Press, Amsterdam, 269–281
72. Lee WHK, Benson RB (2008) Making non-digitally-recorded seismograms accessible online for studying earthquakes. In: Fréchet J, Meghraoui M, Stucchi M (eds) *Modern Approach in Historical Seismology: Interdisciplinary studies of past and recent earthquakes*. Springer, Berlin, pp 403–427
73. Lee WHK, Espinosa-Aranda JM (2003) Earthquake early warning systems: Current status and perspectives. In: Zschau J, Koppers AN (eds) *Early Warning Systems for Natural Disaster, Reduction*. Springer, Berlin, pp 409–423
74. Lee WHK, Stewart SW (1981) *Principles and Applications of Microearthquake Networks*. Academic Press, New York
75. Lee WHK, Celebi M, Todorovska MI, Diggles MF (2007) Rotational seismology and engineering applications: Proceedings for the First International Workshop, Menlo Park, California, USA, 18–19 September. *US Geol Surv Open File Rep 2007–1144*. <http://pubs.usgs.gov/of/2007/1144/>
76. Lee WHK, Igel H, Todorovska MI, Evans JR (2007) Rotational Seismology: AGU Session, Working Group, and Website. *US Geol Surv Open File Rep 2007–1263*. <http://pubs.usgs.gov/of/2007/1263/>
77. Lognonne P, Clevede E (2002) Normal modes of the Earth and planets. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part A*. Academic Press, Amsterdam, pp 125–147
78. Luco JE (1976) Torsional response of structures to obliquely incident seismic SH waves. *Earthq Eng Struct Dyn* 4:207–219

79. Mallet R (1858) Fourth report on the facts of earthquake phenomena. *Ann Rep Brit Assn Adv Sci* 28:1–136
80. Mallet R (1862) Great Neapolitan Earthquake of 1857, vol I, II. Chapman and Hall, London
81. Maruyama T (1963) On the force equivalents of dynamical elastic dislocations with reference to the earthquake mechanism. *Bull Earthq Res Inst* 41:467–486
82. Michelini A, De Simoni B, Amato A, Boschi E (2005) Collecting, digitizing and distributing historical seismological data. *EOS* 86(28) 12 July 2005
83. Nakamura Y (1984) Development of the earthquake early-warning system for the Shinkansen, some recent earthquake engineering research and practical in Japan. The Japanese National Committee of the International Association for Earthquake Engineering, pp 224–238
84. Nakamura Y (1988) On the urgent earthquake detection and alarm system, UrEDAS. *Proc Ninth World Conf Earthq Eng* 7:673–678
85. Nakamura Y (2004) On a rational strong motion index compared with other various indices. 13th World Conf Earthq Eng, Paper No 910
86. Nakamura Y, Saita J (2007) UrEDAS, The Earthquake Warning System: today and tomorrow. In: Gasparini P, Manfredi G, Zschau J (eds) *Earthquake Early Warning Systems*. Springer, Berlin, pp 249–281
87. Newmark NM (1969) Torsion in symmetrical buildings. *Proc. Fourth World Conference on Earthquake Eng*, vol II. pp A3/19–A3/32
88. Niazi M (1987) Inferred displacements, velocities and rotations of a long rigid foundation located at El-Centro differential array site during the 1979 Imperial Valley, California, earthquake. *Earthq Eng Struct Dyn* 14:531–542
89. Odaka T, Ashiya K, Tsukada S, Sato S, Ohtake K, Nozaka D (2003) A new method of quickly estimating epicentral distance and magnitude from a single seismic record. *Bull Seism Soc Am* 93:526–532
90. Oliveira CS, Bolt BA (1989) Rotational components of surface strong ground motion. *Earthq Eng Struct Dyn* 18:517–526
91. Oliver J, Murphy L (1971) WWNSS: Seismology's global network of observing stations. *Science* 174:254–261
92. Peterson J (1993) Observations and modeling of seismic background noise. *US Geol Surv Open File Rep*, 93–322
93. Pillet R, Virieux J (2007) The effects of seismic rotations on inertial sensors. *Geophys J Int*. doi:10.1111/j.1365-246X.2007.03617.x
94. Reid HF (1910) The California Earthquake of 18 April 1906, vol 2. The Mechanics of the Earthquake. Carnegie Inst, Washington DC
95. Richards PG (2002) Seismological methods of monitoring compliance with the Comprehensive Nuclear Test-Ban Treaty. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part A*. Academic Press, Amsterdam, pp 369–382
96. Richter CF (1935) An instrumental earthquake magnitude scale. *Bull Seis Soc Am* 25:1–32
97. Richter CF (1958) *Elementary Seismology*. Freeman, San Francisco
98. Rutenberg A, Heidebrecht AC (1985) Rotational ground motion and seismic codes. *Can J Civ Eng* 12(3):583–592
99. Saita J, Nakamura Y (2003) UrEDAS: the early warning system for mitigation of disasters caused by earthquakes and tsunamis. In: Zschau J, Koppers AN (eds) *Early Warning Systems for Natural Disaster, Reduction*. Springer, Berlin, pp 453–460
100. Savage JC (1978) Dislocation in seismology. In: Nabarro FRN (ed) *Dislocation in Solids*. North-Holland, Amsterdam, 251–339
101. Schweitzer J, Lee WHK (2003) Old seismic bulletins: a collective heritage from early seismologists. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part B*. Academic Press, Amsterdam, pp 1665–1717 (with CD-ROM)
102. Shin TC, Tsai YB, Yeh YT, Liu CC, Wu YM (2003) Strong-motion instrumentation programs in Taiwan. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part B*. Academic Press, Amsterdam, pp 1057–1062
103. Sokolnikoff IS (1956) *Mathematical Theory of Elasticity*, 2nd edn. McGraw-Hill, New York
104. Spudich P, Steck LK, Hellweg M, Fletcher JB, Baker LM (1995) Transient stresses at Park-field, California, produced by the m 7.4 Landers earthquake of 28 June 1992: Observations from the UPSAR dense seismograph array. *J Geophys Res* 100:675–690
105. Suryanto W, Igel H, Wassermann J, Cochard A, Schubert B, Vollmer D, Scherbaum F (2006) Comparison of seismic array-derived rotational motions with direct ring laser measurements. *Bull Seism Soc Am* 96(6):2059–2071
106. Takeo M (1998) Ground rotational motions recorded in near-source region. *Geophys Res Lett* 25(6):789–792
107. Takeo M, Ito HM (1997) What can be learned from rotational motions excited by earthquakes? *Geophys J Int* 129:319–329
108. Teisseyre R, Suchcicki J, Teisseyre KP, Wiszniowski J, Palangio P (2003) Seismic rotation waves: basic elements of theory and recording. *Annali Geofis* 46:671–685
109. Teisseyre R, Takeo M, Majewski E (eds) (2006) *Earthquake Source Asymmetry, Structural Media and Rotation Effects*. Springer, Berlin
110. Teng TL, Wu L, Shin TC, Tsai YB, Lee WHK (1997) One minute after: strong motion map, effective epicenter, and effective magnitude. *Bull Seism Soc Am* 87:1209–1219
111. Trifunac MD (1982) A note on rotational components of earthquake motions on ground surface for incident body waves. *Soil Dyn Earthq Eng* 1:11–19
112. Trifunac MD (2006) Effects of torsional and rocking excitations on the response of structures. In: Teisseyre R, Takeo M, Majewski M (eds) *Earthquake Source Asymmetry, Structural Media and Rotation Effects*. Springer, Heidelberg, pp 569–582
113. Trifunac MD, Todorovska MI (2001) A note on the useable dynamic range of accelerographs recording translation. *Soil Dyn Earthq Eng* 21(4):275–286
114. Twiss R, Souter B, Unruh J (1993) The effect of block rotations on the global seismic moment tensor and the patterns of seismic P and T axes. *J Geophys Res* 98(B1):645–674
115. Utsu T (2002) A list of deadly earthquakes in the world (1500–2000). In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part A*. Academic Press, Amsterdam, pp 691–717
116. Utsu T (2002) Relationships between magnitude scales. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part A*. Academic Press, Amsterdam, pp 733–746

117. Uyeda S (2002) Continental drift, sea-floor spreading, and plate/plume tectonics. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part A*. Academic Press, Amsterdam, pp 51–67
118. Villasenor A, Engdahl ER (2007) Systematic relocation of early instrumental seismicity: Earthquakes in the International Seismological Summary for 1960–1963. *Bull Seism Soc Am* 97:1820–1832
119. Wald DJ, Quitoriano V, Heaton TH, Kanamori H (1999) Relationships between peak ground acceleration, peak ground velocity, and modified Mercalli intensity in California. *Earthq Spectr* 15:557–564
120. Wald DJ, Quitoriano V, Heaton TH, Kanamori H, Scrivner CW, Worden CB (1999) TriNet “ShakeMaps”: Rapid generation of peak ground motion and intensity maps for earthquakes in Southern California. *Earthq Spectr* 15:537–555
121. Webb SC (2002) Seismic noise on land and on the seafloor. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part A*. Academic Press, Amsterdam, pp 305–318
122. Wielandt E (2002) Seismometry. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part A*. Academic Press, Amsterdam, pp 283–304
123. Willemann RJ, Storchak DA (2001) Data collection at the International Seismological Centre. *Seism Res Lett* 72:440–453
124. Wu YM, Kanamori H (2005) Experiment on an onsite early warning method for the Taiwan early warning system. *Bull Seism Soc Am* 95:347–353
125. Wu YM, Kanamori H (2005) Rapid assessment of damaging potential of earthquakes in Taiwan from the beginning of P Waves. *Bull Seism Soc Am* 95:1181–1185
126. Wu YM, Kanamori H (2008) Exploring the feasibility of on-site earthquake early warning using close-in records of the 2007 Noto Hanto earthquake. *Earth Planets Space* 60:155–160
127. Wu YM, Zhao L (2006) Magnitude estimation using the first three seconds P-wave amplitude in earthquake early warning. *Geophys Res Lett* 33:L16312. doi:10.1029/2006GL026871
128. Wu YM, Chen CC, Shin TC, Tsai YB, Lee WHK, Teng TL (1997) Taiwan Rapid Earthquake Information Release System. *Seism Res Lett* 68:931–943
129. Wu YM, Hsiao NC, Lee WHK, Teng TL, Shin TC (2007) State of the art and progresses of early warning system in Taiwan. In: Gasparini P, Manfredi G, Zschau J (eds) *Earthquake Early Warning Systems*. Springer, Berlin, pp 283–306
130. Wu YM, Kanamori H, Allen R, Hauksson E (2007) Determination of earthquake early warning parameters,  $\tau_c$  and  $P_d$ , for southern California. *Geophys J Int* 169:667–674
131. Wu YM, Lee WHK, Chen CC, Shin TC, Teng TL, Tsai YB (2000) Performance of the Taiwan Rapid Earthquake Information Release System (RTD) during the 1999 Chi-Chi (Taiwan) earthquake. *Seism Res Lett* 71:338–343
132. Zollo A, Lancieri M, Nielsen S (2006) Earthquake magnitude estimation from peak amplitudes of very early seismic signals on strong motion records. *Geophys Res Lett* 33:L23312. doi:10.1029/2006GL027795

### Books and Reviews

- Abercrombie R, McGarr A, Kanamori H, Di Toro G (2006) *Earthquakes: Radiated Energy and the Physics of Faulting*. Geophysical Monograph, vol 170. American Geophysical Union, Washington DC
- Bolt BA (1993) *Earthquakes*. W.H. Freeman, New York
- Chen YT, Panza GF, Wu ZL (2004) *Earthquake Hazard, Risk, and Strong Ground Motion*. Seismological Press, Beijing
- Kanamori H (ed) (2007) *Earthquake Seismology, Treatise on Geophysics, vol 4*. Elsevier, Amsterdam
- Kanamori H, Boschi E (1983) *Earthquakes: Observation, Theory and Interpretation*. North-Holland, Amsterdam
- Keilis-Borok VI, Soloviev AA (eds) (2003) *Nonlinear Dynamics of the Lithosphere and Earthquake Prediction*. Springer, Berlin
- Lee WHK, Meyers H, Shimazaki K (eds) (1988) *Historical Seismograms and Earthquakes of the World*. Academic Press, San Diego
- Lee WHK, Kanamori H, Jennings JC, Kisslinger C (eds) (2002) *International Handbook of Earthquake and Engineering Seismology, Part A*. Academic Press, San Diego, pp 933 (and 1 CD-ROM)
- Lee WHK, Kanamori H, Jennings JC, Kisslinger C (eds) (2003) *International Handbook of Earthquake and Engineering Seismology, Part B*. Academic Press, San Diego, pp 1009 (and 2 CD-ROMs)
- Pujol J (2003) *Elastic Wave Propagation and Generation in Seismology*. Cambridge Univ Press, Cambridge
- Zschau J, Koppers AN (eds) (2003) *Early Warning Systems for Natural Disaster Reduction*. Springer, Berlin

## Earthquake Networks, Complex

SUMIYOSHI ABE<sup>1,2</sup>, NORIKAZU SUZUKI<sup>3</sup>

<sup>1</sup> Department of Physical Engineering, Mie University, Tsu, Japan

<sup>2</sup> Institut Supérieur des Matériaux et Mécaniques, Le Mans, France

<sup>3</sup> College of Science and Technology, Nihon University, Chiba, Japan

### Article Outline

Glossary

Definition of the Subject

Introduction

Construction of an Earthquake Network

Scale-free Nature of Earthquake Network

Small-World Nature of Earthquake Network

Hierarchical Structure

Mixing Property

Period Distribution

Future Directions

Addendum

Bibliography

### Glossary

**Network or graph** A network (or a graph) [28] consists of vertices (or nodes) and edges (or links) connecting them. In general, a network contains loops (i. e., edges with both ends attached to the same vertices) and multiple edges (i. e., edges more than one that connect two different vertices). If edges have their directions, such a network is called directed. A simple graph is a network, in which loops are removed and each multiple edge is replaced by a single edge. In a stochastic network, each connection is inherently probabilistic. A classical random graph is a simple example, in which each two vertices are connected by an edge with probability  $p$  and unconnected with probability  $1 - p$  ( $0 < p < 1$ ).

**Connectivity distribution or degree distribution** The connectivity distribution (or the degree distribution),  $P(k)$ , is the probability of finding vertices with  $k$  edges in a stochastic network. In a directed network, the number of incoming/outgoing edges is called the in-degree/out-degree. Connectivity of a classical random graph obeys the Poissonian distribution in the limit of the large number of vertex [11,14,20],  $P(k) = e^{-\lambda} \lambda^k / k!$  ( $\lambda$ : a positive parameter,  $k = 0, 1, 2, \dots$ ), whereas a scale-free network [11,12,14,20]

has a power-law shape,  $P(k) \sim k^{-\gamma}$  ( $\gamma$ : a positive exponent), for large  $k$ .

**Preferential attachment rule** This is a concept relevant to a growing network, in which the number of vertices increases. Preferential attachment [11,12,14,20] implies that a newly created vertex tends to link to pre-existing vertices with the probability  $\Pi(k_i) = k_i / \sum_j k_j$ , where  $k_i$  stands for the connectivity of the  $i$ th vertex. That is, the larger the connectivity of a vertex is, the higher the probability of getting linked to a new vertex is.

**Clustering coefficient** The clustering coefficient [27] is a quantity characterizing an undirected simple graph. It quantifies the adjacency of two neighboring vertices of a given vertex, i. e., the tendency of two neighboring vertices of a given vertex to be connected to each other. Mathematically, it is defined as follows. Assume the  $i$ th vertex to have  $k_i$  neighboring vertices. There can exist at most  $k_i(k_i - 1)/2$  edges between the neighbors. Define  $c_i$  as the ratio

$$c_i = \frac{\text{actual number of edges between the neighbors of the } i\text{th vertex}}{k_i(k_i - 1)/2} . \quad (1)$$

Then, the clustering coefficient is given by the average of this quantity over the network:

$$C = \frac{1}{N} \sum_{i=1}^N c_i , \quad (2)$$

where  $N$  is the total number of vertices contained in the network. The value of the clustering coefficient of a random graph,  $C_{\text{random}}$ , is much smaller than unity, whereas a small-world network has a large value of  $C$  which is much larger than  $C_{\text{random}}$ .

**Hierarchical organization** Many complex networks are structurally modular, that is, they are composed of groups of vertices that are highly interconnected to each other but weakly connected to outside groups. This hierarchical structure [22] can conveniently be characterized by the clustering coefficient at each value of connectivity,  $c(k)$ , which is defined by

$$c(k) = \frac{1}{NP_{SG}(k)} \sum_{i=1}^N c_i \delta_{k_i, k} , \quad (3)$$

where  $c_i$  is given by (1),  $N$  the total number of vertices, and  $P_{SG}(k)$  the connectivity distribution of an undirected simple graph. Its average is the clustering coefficient in (2):  $C = \sum_k c(k)P_{SG}(k)$ . A network is said to

be hierarchically organized if  $c(k)$  varies with respect to  $k$ , typically due to a power law,  $c(k) \sim k^{-\beta}$ , with a positive exponent  $\beta$ .

**Assortative mixing and disassortative mixing** Consider the conditional probability,  $P(k'|k)$ , of finding a vertex with connectivity  $k'$  linked to a given vertex with connectivity  $k$ . Then, the nearest-neighbor average connectivity of vertices with connectivity  $k$  is defined by [20,21,26]

$$\bar{k}_{nn}(k) = \sum_{k'} k' P(k'|k). \quad (4)$$

If  $\bar{k}_{nn}(k)$  increases/decreases with respect to  $k$ , mixing is termed assortative/disassortative. A simple model of growth with preferential attachment is known to possess no mixing. That is,  $\bar{k}_{nn}(k)$  does not depend on  $k$ . The above-mentioned linking tendency can be quantified by the correlation coefficient [17] defined as follows. Let  $e_{kl}(= e_{lk})$  be the joint probability distribution for an edge to link with a vertex with connectivity  $k$  at one end and a vertex with connectivity  $l$  at the other. Calculate its marginal,  $q_k = \sum_l e_{kl}$ . Then, the correlation coefficient is given by

$$r = \frac{1}{\sigma_q^2} \sum_{k,l} kl (e_{kl} - q_k q_l), \quad (5)$$

where  $\sigma_q^2 = \sum_k k^2 q_k - (\sum_k k q_k)^2$  stands for the variance of  $q_k$ .  $r \in [-1, 1]$ , and if  $r$  is positive/negative, mixing is assortative/disassortative [17,20].

### Definition of the Subject

Complexity is an emergent collective property, which is hardly understood by the traditional approach in natural science based on reductionism. Correlation between elements in a complex system is strong, no matter how largely they are separated both spatially and temporally, therefore it is essential to treat such a system in a holistic manner, in general.

Although it is generally assumed that seismicity is an example of complex phenomena, it is actually nontrivial to see how and in what sense it is complex. This point may also be related to the question of primary importance of why it is so difficult to predict earthquakes.

Development of the theory of complex networks turns out to offer a peculiar perspective on this point. Construction of a complex earthquake network proposed here consists of mapping seismic data to a growing stochastic graph. This graph, or network, turns out to exhibit a number of remarkable behaviors both physically and mathematically, which are in common with many other complex

systems. The scale-free and small-world natures are typical examples. In this way, one will be able to obtain a novel viewpoint of seismicity.

### Introduction

Seismicity is a field-theoretical phenomenon. Released energy of each earthquake may be regarded as a field amplitude defined at a discrete spacetime point. However, in contrast to a familiar field theory such as the electromagnetic theory, both amplitudes and locations are intrinsically probabilistic. The fault distribution may geometrically be fractal [18], and the stress distribution superposed upon it often has a complex landscape. Accordingly, seismicity is characterized by extremely rich phenomenology, which attracts physicists' attention from the viewpoint of science of complex systems.

There are at least two celebrated empirical laws known in seismology. One is the Gutenberg–Richter law [16], which states that the frequency of earthquakes obeys a power law with respect to released energy. This power-law nature makes it difficult or even meaningless to statistically distinguish earthquakes by their values of magnitude because of the absence of typical energy scales. The other is the Omori law [19], which states that the rate of the frequency of aftershocks following a main shock algebraically decays with respect to time elapsed from the main shock. This slow relaxation reminds one of complex glassy dynamics [13]. Such a viewpoint is supported by the discovery of the aging phenomenon and the scaling law for aftershocks [2].

Another point, which seems less noticed, is that correlation of two successive events is strong, no matter how large their spatial separation is. There is, in fact, an observation [23] that an earthquake can be triggered by a fore-going one more than 1000 km away. The reason why two successive events are indivisibly related can also be found in another observation [3,6] that both spatial distance and time interval between two successive events obey the  $q$ -exponential distributions in nonextensive statistics [1,15,24], which offers a statistical-mechanical framework for describing complex systems. Thus, the correlation length can be enormously large and long-wave-length modes of seismic waves play an important role. This has a strong similarity to phase transitions and critical phenomena. Accordingly, it may not be appropriate to use spatial windows in analysis of seismicity. Furthermore, all of the data in a relevant area (ideally the whole globe, though still not satisfactorily available) should be treated based on the nonreductionistic standpoint.



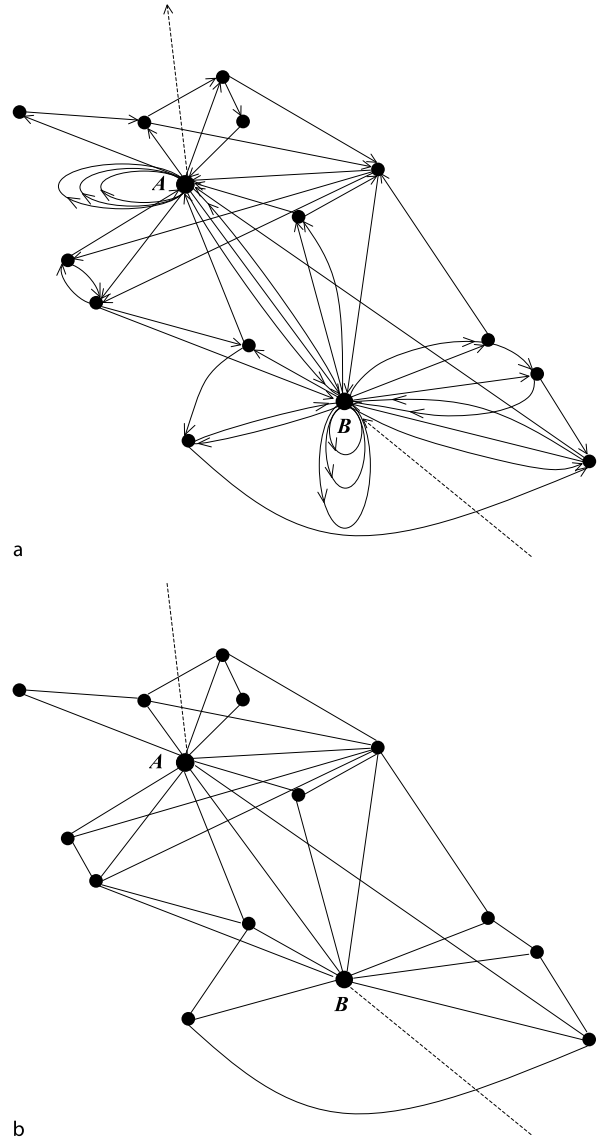
The network approach is a powerful tool for analyzing kinematical and dynamical structures of complex systems in a holistic manner. Such a concept was introduced to seismology by the present authors in 2004 [4] in order to represent complexity of seismicity. The procedure described in Sect. “Construction of an Earthquake Network” allows one to map a seismic time series to a growing stochastic network in an unambiguous way. Vertices and edges of such a network correspond to coarse-grained events and event-event correlations, respectively. Yet unknown microscopic dynamics governing event-event correlations and fault-fault interactions are replaced by these edges. Global physical properties of seismicity can then be explored by examining its geometric (e. g., topological etc.) and dynamical features. It turns out that earthquake networks have a number of intriguing properties, some of which are shared by many other natural as well as artificial systems including metabolic networks, food webs, the Internet, the world-wide web, and so on [11,14,20]. This, in turn, enables seismologists to study seismicity in analogy with such relatively better understood complex systems. Thus, the network approach offers a novel way of analyzing seismic time series and casts fresh light on the physics of earthquakes.

In this article, only the data taken from California is utilized. However, it has been ascertained that the laws and trends discussed here are universal and hold also in other geographical regions including Japan.

### Construction of an Earthquake Network

An earthquake network is constructed as follows [4]. A geographical region under consideration is divided into small cubic cells. A cell is regarded as a vertex if earthquakes with any values of magnitude above a certain detection threshold occurred therein. Two successive events define an edge between two vertices. If they occur in the same cell, a loop is attached to that vertex. This procedure enables one to map a given interval of the seismic data to a growing probabilistic graph, which is referred to as an earthquake network (see Fig. 1a).

Several comments are in order. Firstly, this construction contains a single parameter: cell size, which is a scale of coarse graining. Once cell size is fixed, an earthquake network is unambiguously defined. However, since there exist no a priori operational rule to determine cell size, it is important to notice how the properties of an earthquake network depend on this parameter. Secondly, as mentioned in Sect. “Introduction”, edges and loops efficiently represent event-event correlation. Thirdly, an earthquake network is a directed graph in its nature. Directedness does



Earthquake Networks, Complex, Figure 1

**a** A schematic description of earthquake network. The *dashed lines* correspond to the initial and final events. The vertices, *A* and *B*, contain main shocks and play roles of hubs of the network. **b** The undirected simple graph reduced from the network in **a**.

not bring any difficulties to statistical analysis of connectivity (degree, i. e., the number of edges attached to the vertex under consideration) since, by construction, the in-degree and out-degree are identical for each vertex except the initial and final vertices in analysis. Therefore, the in-degree and out-degree are not distinguished from each other in the analysis of the connectivity distribution (see Sections “Scale-free Nature of Earthquake Network”

and “Mixing Property”). However, directedness becomes essential when the path length (i. e., the number of edges) between a pair of connected vertices, i. e., the degree of separation between the pair, is considered. This point is explicitly discussed in the analysis of the period distribution in Sect. “Period Distribution”. Finally, a full directed earthquake network has to be reduced to a simple undirected graph, when its small-worldness and hierarchical structure are examined (see Sections “Small-World Nature of Earthquake Network” and “Hierarchical Structure”). There, loops are removed and each multiple edge is replaced by a single edge (see Fig. 1b). The path length in this case is the smallest value among the possible numbers of edges connecting a pair of vertices.

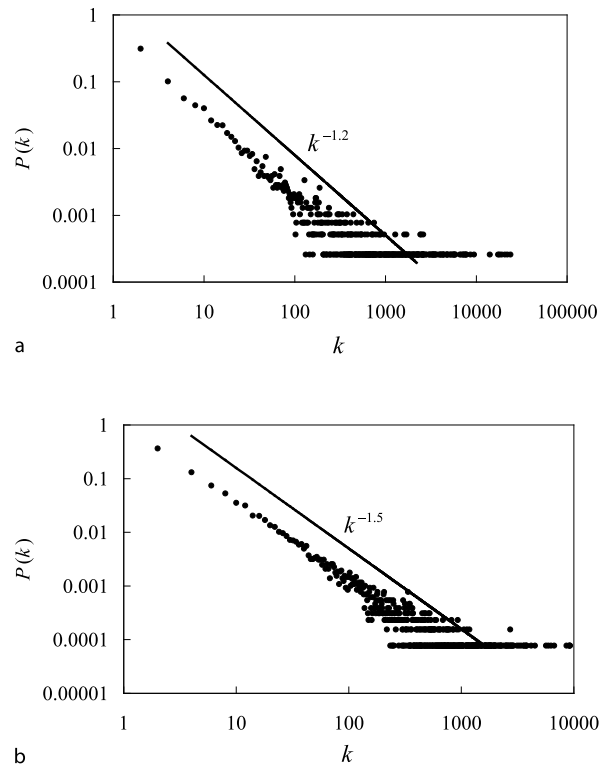
**Scale-free Nature of Earthquake Network**

An earthquake network contains some special vertices which have large values of connectivity. Such “hubs” turn out to correspond to cells with main shocks. This is due to a striking fact discovered from real data analysis that aftershocks associated with a main shock tend to return to the locus of the main shock, geographically. This is the primary reason why a vertex containing a main shock becomes a hub. The situation is analogous to the preferential attachment rule for a growing network [11,14,20]. According to this rule, a newly created vertex tends to be connected to the (already existing)  $i$ th vertex with connectivity  $k_i$  with probability,  $\Pi(k_i) = k_i / \sum_j k_j$ . It can generate a scale-free network characterized by the power-law connectivity distribution [11,12]:

$$P(k) \sim k^{-\gamma}, \tag{6}$$

where  $\gamma$  is a positive exponent.

In Fig. 2, the connectivity distribution of the full earthquake network with loops and multiple edges is presented [4]. From it, one appreciates that the earthquake network in fact possesses connectivity of the form in (6) and is therefore scale-free. The smaller the cell size is, the larger the exponent,  $\gamma$ , is, since the number of vertices with large values of connectivity decreases as cell size becomes smaller. The scale-free nature may be interpreted as follows. As mentioned above, aftershocks associated with a main shock tend to be connected to the vertex of the main shock, satisfying the preferential attachment rule. On the other hand, the Gutenberg–Richter law states that frequency of earthquakes decays slowly as a power law with respect to released energy. This implies that there appear to be quite a few giant components, and accordingly the network becomes highly inhomogeneous.



Earthquake Networks, Complex, Figure 2  
 The log-log plots of the connectivity distributions of the earthquake network constructed from the seismic data taken in California [the Southern California Earthquake Data Center (<http://www.data.scec.org/>)]. The time interval analyzed is between 00:25:8.58 on January 1, 1984 and 22:21:52.09 on December 31, 2003. The region covered is 29°06.00’N–38°59.76’N latitude and 113°06.00’W–122°55.59’W longitude with the maximum depth 175.99 km. The total number of events is 367 613. The data contains no threshold for magnitude (but “quarry blasts” are excluded from the analysis). Two different values of cell size are examined: a 10 km × 10 km × 10 km and b 5 km × 5 km × 5 km. All quantities are dimensionless.

**Small-World Nature of Earthquake Network**

The small-world nature is an important aspect of complex networks. It shows how a complex network is different from both regular and classical random graphs [27]. A small-world network resides in-between regularity and randomness, analogous to the edge of chaos in nonlinear dynamics.

To study the small-world nature of an earthquake network, a full network has to be reduced to a simple undirected graph: that is, loops are removed and each multiple edge is replaced by a single edge (see Fig. 1b). This is because in the small-world picture one is concerned only with simple linking pattern of vertices.

Earthquake Networks, Complex, Table 1

The small-world properties of the undirected simple earthquake network. The values of the number of vertices,  $N$ , the clustering coefficient,  $C$ , (compared with those of the classical random graphs,  $C_{\text{random}}$ ) and the average path length,  $L$  are presented. The data employed is the same as that in Fig. 2.

Cell size	10 km × 10 km × 10 km	5 km × 5 km × 5 km
Number of vertices	$N = 3869$	$N = 12913$
Clustering coefficient	$C = 0.630$ ( $C_{\text{random}} = 0.014$ )	$C = 0.317$ ( $C_{\text{random}} = 0.003$ )
Average path length	$L = 2.526$	$L = 2.905$

A small-world network is characterized by a large value of the clustering coefficient in (2) and a small value of the average path length [27]. The clustering coefficient quantifies the tendency of two neighboring vertices of a given vertex to be connected to each other. A small-world network has a large value of the clustering coefficient, whereas the value for the classical random graph is very small [11,14,20,27]:  $C_{\text{random}} = \langle k \rangle / N \ll 1$ , where  $N$  and  $\langle k \rangle$  are the total number of vertices and the average value of connectivity, respectively.

In Table 1, the results are presented for the clustering coefficient and the average path length [5,9]. One finds that the values of the clustering coefficient are in fact much larger than those of the classical random graphs and the average path length is short. Thus, the earthquake network sportant features of small-world network.

### Hierarchical Structure

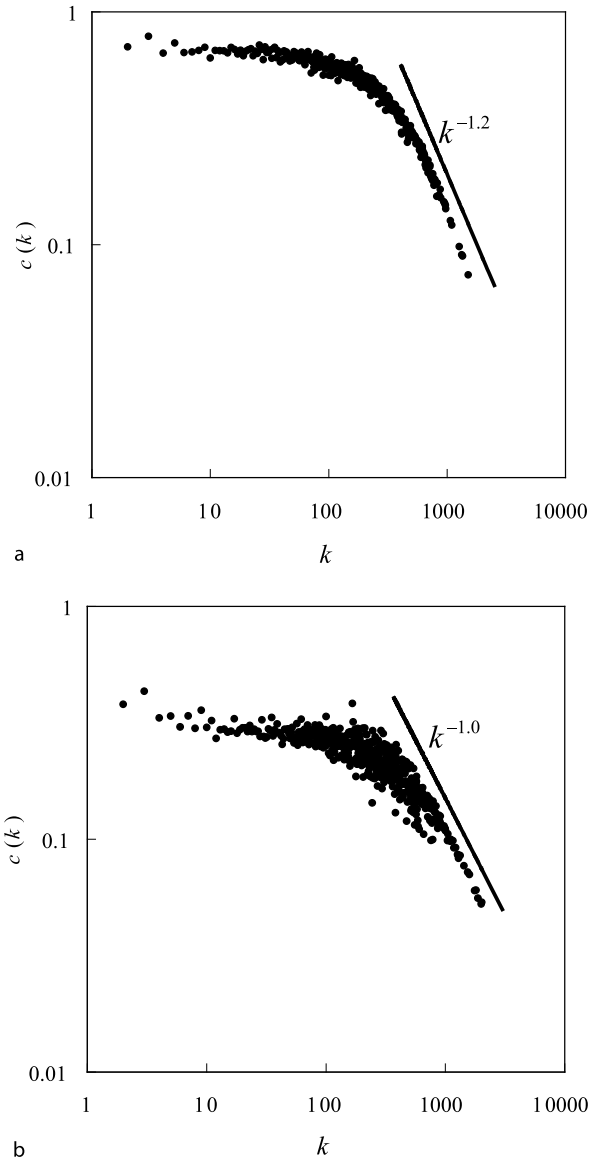
As seen above, seismicity generates a scale-freeexity of an earthquake network further, one may examine if it is hierarchically organized [8]. The hierarchical structure can be revealed by analyzing the clustering coefficient as a function of connectivity. The connectivity-dependent clustering coefficient,  $c(k)$ , is defined in (3). This quantifies the adjacency of two vertices connected to a vertex with connectivity,  $k$ , and gives information on hierarchical organization of a network.

In Fig. 3, the plots of  $c(k)$  are presented [8]. As can be clearly seen, the clustering coefficient of the undirected simple earthquake network asymptotically follows the scaling law

$$c(k) \sim k^{-\beta} \quad (7)$$

with a positive exponent  $\beta$ . This highlights hierarchical organization of the earthquake network.

Existence of the hierarchical structure is of physical importance. The earthquake networkment [11,12,14,20].



Earthquake Networks, Complex, Figure 3

The log-log plots of the connectivity-dependent clustering coefficient for two different values of cell size: a 10 km × 10 km × 10 km and b 5 km × 5 km × 5 km. The analyzed period is between 00:25:8.58 on January 1, 1984 and 22:50:49.29 on December 31, 2004, which is taken from the same catalog as in Fig. 2. The region covered is 28°36.00'N–38°59.76'N latitude and 112°42.00'W–123°37.41'W longitude with the maximal depth 175.99 km. The total number of the events is 379728. All quantities are dimensionless.

However, the standard preferential-attachment-model is known to fail at generating hierarchical organization [22]. To mediate between growth with preferential attachment and the presence of hierarchical organization, the concept

of vertex deactivation has been introduced in the literature [25]. According to this concept, in the process of network growth, some vertices deactivate and cannot acquire new edges any more. This has a natural physical implication for an earthquake network: active faults may be deactivated through the process of stress release. In addition, the fitness model [26] is also known to generate hierarchical organization. This model generalizes the preferential attachment rule in such a way that not only connectivity but also “charm” of vertices (i. e., attracting a lot of edges) are taken into account. Seismologically, fitness is considered to describe intrinsic properties of faults such as geometric configuration and stiffness. Both of these two mechanisms can explain a possible origin of the complex hierarchical structure, by which relatively new vertices have chances to become hubs of the network. In the case of an earthquake network, it seems plausible to suppose that the hierarchical structure may be due to both deactivation and fitness.

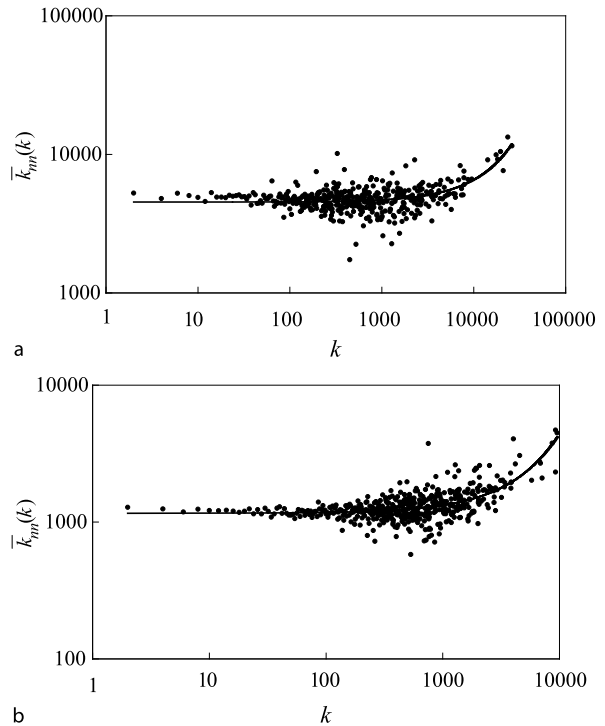
A point of particular interest is that the hierarchical structure disappears if weak earthquakes are removed. For example, setting a lower threshold for earthquake magnitude, say  $M_{th} = 3$ , makes it difficult to observe the power-law decay of the clustering coefficient hierarchical structure of an earthquake network is largely supported by weak shocks.

**Mixing Property**

The scale-free nature, small-worldness, growth with preferential attachment, and hierarchical organization all indicate that earthquake networks are very similar to other known networks, for example, the Internet. However, there is at least one point which shows an essential difference between the two. It is concerned with the mixing property, which is relevant to the concept of the nearest-neighbor average connectivity  $\bar{k}_{nn}(k)$  [in (4)], of a full network with loops and multiple edges.

The plots of this quantity are presented in Fig. 4. There, the feature of assortative mixing [8] is observed, since  $\bar{k}_{nn}(k)$  increases with respect to connectivity  $k$ . Therefore, vertices with large values of connectivity tend to be linked to each other. That is, vertices containing stronger shocks tend to be connected among themselves with higher probabilities.

To quantify this property, the correlation coefficient in (5) is evaluated [8]. The result is summarized in Table 2. The value of the correlation coefficient is in fact positive, confirming that the earthquake network has assortative mixing. On the other hand, the Internet is of disassortative mixing [17,20,21,26]. That is, the mixing properties



**Earthquake Networks, Complex, Figure 4**  
 The log-log plots of the nearest-neighbor average connectivity for two different values of cell size: **a** 10 km × 10 km × 10 km and **b** 5 km × 5 km × 5 km. The data employed is the same as that in Fig. 3. The solid lines show the trends depicted by the exponentially increasing functions. All quantities are dimensionless.

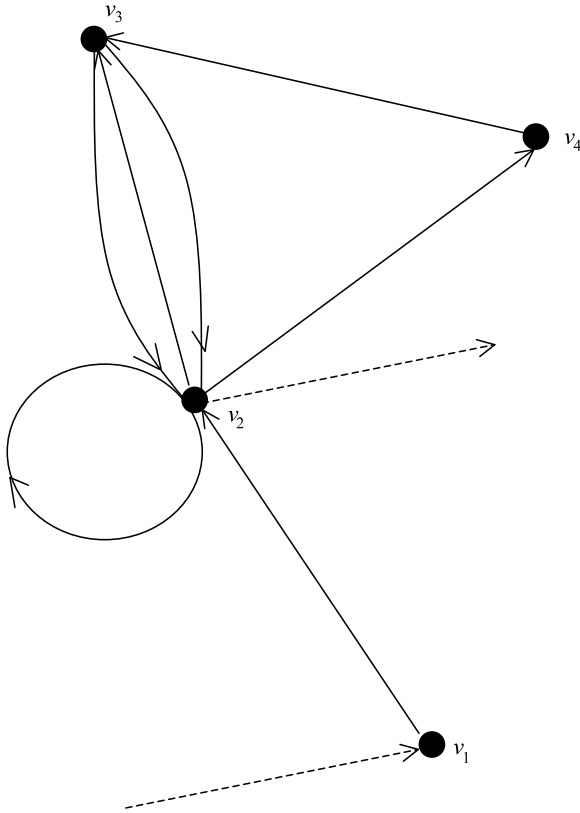
**Earthquake Networks, Complex, Table 2**  
 The values of the dimensionless correlation coefficient. The data employed is the same as that in Fig. 3. Positivity of the values implies that mixing is assortative.

10 km × 10 km × 10 km	5 km × 5 km × 5 km
$r = 0.285$	$r = 0.268$

of the earthquake network and the Internet are opposite to each other. It is noticed however that the loops and multiple edges play essential roles for the assortative mixing: an undirected simple graph obtained by reducing a full earthquake network turns out to exhibit disassortative mixing. These are purely the phenomenological results, and their physical origins still have yet to be clarified.

**Period Distribution**

So far, directedness of an earthquake network has been ignored. The full directed network picture is radically different from the small-world picture for a simple undirected graph. It enables one to consider interesting dynamical



Earthquake Networks, Complex, Figure 5  
 A full directed network:  $\dots \rightarrow v_1 \rightarrow v_2 \rightarrow v_3 \rightarrow v_2 \rightarrow v_2 \rightarrow v_4 \rightarrow v_3 \rightarrow v_2 \rightarrow \dots$ . The period associated with  $v_3$  is 4, whereas  $v_2$  has 1, 2 and 3.

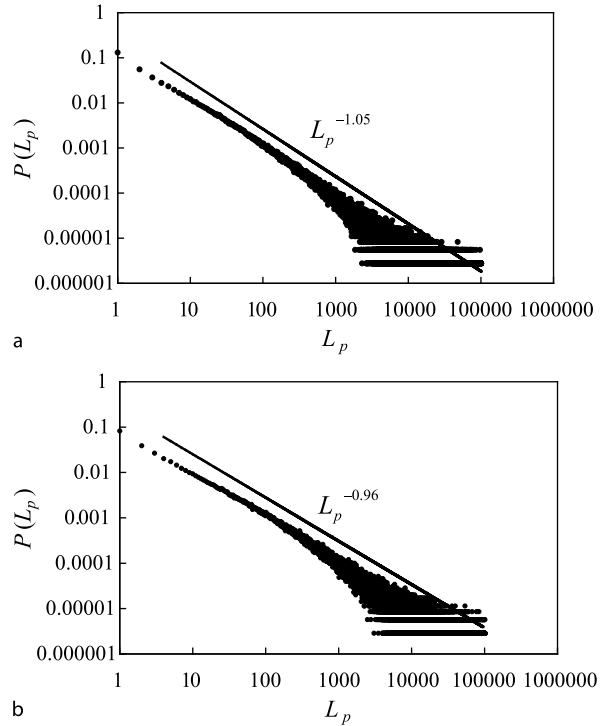
features of an earthquake network. As an example, here the concept of period [7] is discussed. This is relevant to the question “after how many earthquakes does an event return to the initial cell, statistically?” It is therefore of obvious interest for earthquake prediction.

Period in a directed network is defined as follow. Given a vertex of a network, there are various closed routes starting from and ending at this vertex. The period,  $L_p$ , of a chosen closed route is simply the number of edges forming the route (see Fig. 5).

The period distribution,  $P(L_p)$ , is defined as the number of closed routes. The result is presented in Fig. 6 [7]. As can be seen there,  $P(L_p)$  obeys a power law

$$P(L_p) \sim (L_p)^{-\alpha}, \tag{8}$$

where  $\alpha$  is a positive exponent. This implies that there exist a number of closed routes with significantly long periods in the network. This fact makes it highly nontrivial to statistically estimate the value of period.



Earthquake Networks, Complex, Figure 6  
 The log-log plots of the period distribution for two different values of cell size: a  $10 \text{ km} \times 10 \text{ km} \times 10 \text{ km}$  and b  $5 \text{ km} \times 5 \text{ km} \times 5 \text{ km}$ . The data employed is the same as that in Fig. 2. All quantities are dimensionless.

### Future Directions

In the above, the long-time statistical properties of an earthquake network have mainly been considered. On the other hand, given the cell size, an earthquake network represents all the dynamical information contained in a seismic time series, and therefore the study of its time evolution may give a new insight into seismicity. This, in turn, implies that it may offer a novel way of monitoring seismicity.

For example, it is of interest to investigate how the clustering coefficient changes in time as earthquake network dynamically evolves. According to the work in [10], the clustering coefficient remains stationary before a main shock, suddenly jumps up at the main shock, and then slowly decays to become stationary again following the power-law relaxation. In this way, the clustering coefficient successfully characterizes aftershocks in association with main shocks.

A question of extreme importance is if precursors of a main shock can be detected through monitoring dynamical evolution of earthquake network. Clearly, further de-

velopments are needed in science of complex networks to address to this question.

### Addendum

Some authors (e.g., Sornette and Werner) raised questions about the applicability of complex earthquake network using online accessible earthquake catalog data and on the validity of our results. A preprint of an article by the authors discussing these questions can be found in the e-print available at <http://arxiv.org/abs/0708.2203>.

### Bibliography

1. Abe S, Okamoto Y (eds) (2001) *Nonextensive statistical mechanics and its applications*. Springer, Heidelberg
2. Abe S, Suzuki N (2003) Aging and scaling of earthquake aftershocks. *Physica A* 332:533–538
3. Abe S, Suzuki N (2003) Law for the distance between successive earthquakes. *J Geophys Res* 108(B2):2113 ESE 19-1-4
4. Abe S, Suzuki N (2004) Scale-free network of earthquakes. *Europhys Lett* 65:581–586
5. Abe S, Suzuki N (2004) Small-world structure of earthquake network. *Physica A* 337:357–362
6. Abe S, Suzuki N (2005) Scale-free statistics of time interval between successive earthquakes. *Physica A* 350:588–596
7. Abe S, Suzuki N (2005) Scale-invariant statistics of period in directed earthquake network. *Eur Phys J B* 44:115–117
8. Abe S, Suzuki N (2006) Complex earthquake networks: Hierarchical organization and assortative mixing. *Phys Rev E* 74:026113-1-5
9. Abe S, Suzuki N (2006) Complex-network description of seismicity. *Nonlin Processes Geophys* 13:145–150
10. Abe S, Suzuki N (2007) Dynamical evolution of clustering in complex network of earthquakes. *Eur Phys J B* 59:93–97
11. Albert R, Barabási A-L (2002) *Statistical mechanics of complex networks*. *Rev Mod Phys* 74:47–97
12. Barabási A-L, Albert R (1999) Emergence of scaling in random networks. *Science* 286:509–512
13. Debenedetti PG, Stillinger FH (2001) Supercooled liquids and the glass transition. *Nature* 410:259–267
14. Dorogovtsev SN, Mendes JFF (2003) *Evolution of networks: from biological nets to the Internet and WWW*. Oxford University Press, Oxford
15. Gell-Mann M, Tsallis C (eds) (2004) *Nonextensive Entropy: Interdisciplinary Applications*. Oxford University Press, Oxford
16. Gutenberg B, Richter CF (1944) Frequency of earthquakes in California. *Bull Seismol Soc Am* 34:185–188
17. Newman MEJ (2002) Assortative mixing in networks. *Phys Rev Lett* 89:208701-1-4
18. Okubo PG, Aki K (1987) Fractal geometry in the San Andreas fault system. *J Geophys Res* 92(B1):345–355
19. Omori F (1894) On the after-shocks of earthquakes. *J Coll Sci Imper Univ Tokyo* 7:111–200
20. Pastor-Satorras R, Vespignani A (2004) *Evolution and structure of the Internet: a statistical physics approach*. Cambridge University Press, Cambridge
21. Pastor-Satorras R, Vázquez A, Vespignani A (2001) Dynamical and correlation properties of the Internet. *Phys Rev Lett* 87:258701-1-4
22. Ravasz E, Barabási A-L (2003) Hierarchical organization in complex networks. *Phys Rev E* 67:026112-1-7
23. Steeples DW, Steeples DD (1996) Far-field aftershocks of the 1906 earthquake. *Bull Seismol Soc Am* 86:921–924
24. Tsallis C (1988) Possible generalization of Boltzmann–Gibbs statistics. *J Stat Phys* 52:479–487
25. Vázquez A, Boguñá M, Moreno Y, Pastor-Satorras R, Vespignani A (2003) Topology and correlations in structured scale-free networks. *Phys Rev E* 67:046111-1-10
26. Vázquez A, Pastor-Satorras R, Vespignani A (2002) Large-scale topological and dynamical properties of the Internet. *Phys Rev E* 65:066130-1-12
27. Watts DJ, Strogatz SH (1998) Collective dynamics of ‘small-world’ networks. *Nature* 393:440–442
28. Wilson RJ (1996) *Introduction to graph theory*, 4th edn. Prentice Hall, Harlow

## Earthquake Nucleation Process

YOSHIHISA IIO

Disaster Prevention Research Institute,  
Kyoto University, Kyoto, Japan

### Article Outline

Glossary

Definition of the Subject

Introduction

Contribution from the Development  
of Earthquake Early-Warning Systems

Observations of Initial Rupture Processes

Discussion

Future Directions

Acknowledgments

Bibliography

### Glossary

**Nucleation process** The process in which rupture velocity accelerates from quasi-static to dynamic. The dynamic rupture velocity almost equals the shear wave velocity.

**Nucleation zone** The portion of the fault where rupture velocity accelerates from quasi-static to dynamic.

**Initial rupture process** The rupture process that precedes the largest slip. This term is used when the acceleration of rupture velocity is not clear. This is a wider concept that includes the earthquake nucleation process. The area where the initial rupture process occurs is called the initial rupture fault.

**Slip velocity** The dislocation velocity at a point on the fault. The rupture velocity is the velocity at which the rupture front is expanding.

**Preslip model** An earthquake source model having a detectable size nucleation zone.

**Cascade model** An earthquake source model in which smaller sub-events successively trigger larger sub-events. A sub-event is the same as a small earthquake if it does not trigger a successive sub-event.

**Stress drop (static stress drop)** The amount of shear stress change at a point on the fault before and after an earthquake. It is proportional to the strain released on the fault.

**Dynamic stress drop** The difference between the initial shear stress and the minimum frictional stress at a point on the fault during fault slip.

**Fault strength** The shear stress level necessary to initiate slip at a point on the fault.

**$M$**  Magnitude. Earthquake size computed basically from waveform amplitudes and focal distances.

**Seismic moment** The most reliable measure of earthquake size which is determined from the products of the rigidity near the fault, the amount of slip, and the area of the fault surface.

**$M_w$**  Moment magnitude. Earthquake magnitude derived from the seismic moment.

### Definition of the Subject

Earthquake prediction in the long, intermediate, and short terms is essential for the reduction of earthquake disasters. However, it is not practical at present, in particular, for the intermediate and short time scales of a few days to years. This is mainly because we do not know exactly how and why earthquakes begin and grow larger or stop. Theoretical and laboratory studies have confirmed that earthquakes do not begin abruptly with dynamic rupture propagation. They show that a quasi-static rupture growth precedes dynamic rupture. Thus, if we can detect the quasi-static rupture growth, we could forecast the following dynamic rupture. A key issue is how natural earthquakes initiate. To solve this issue, a first approach would be to investigate the very beginning parts of observed waveforms of earthquakes, since they can reflect the earthquake nucleation process from a quasi-static to a dynamic rupture. This paper reviews the studies analyzing the beginning parts of observed waveforms, and shows what we presently understand about the earthquake nucleation process.

### Introduction

Earthquakes initiate at a small portion on a fault. Then, their rupture fronts expand outward until they stop. Some large earthquakes have a rupture extent greater than 1000 km (e. g., the 2004 Sumatra Earthquake), while fault lengths of very small microearthquakes ( $M = 0$ ) are estimated to be a few meters [13]. Fault lengths of earthquakes range at least over 6 orders of magnitude. Surprisingly, the concept that earthquakes are self-similar is widely accepted in spite of the difference in fault length (e. g., [40]). One example of the similarity is that the average fault slip is proportional to the fault length. In other words, the static stress drop is constant independent of earthquake size.

The similarity law raises fundamental questions: Why do large earthquakes grow larger? What is the difference between large and small earthquakes? An end member model represents earthquakes as ruptures that grow randomly, and terminate in an earlier stage for smaller earthquakes while continuing longer for larger earthquakes.

This type of model has been proposed mainly to explain the frequency–size distribution of earthquakes (e. g., [8]) and implies that it is impossible to forecast the final size of earthquakes at the time of the rupture initiations. However, another end member model predicts that larger earthquakes have a larger “seed” than smaller earthquakes and that large and small earthquakes are different even at their beginnings (e. g., [66]).

The key issue is how earthquakes initiate and grow larger.

All the theoretical models of earthquake sources predict that earthquake shear failures begin with a quasi-static rupture growth on a small portion on the fault (e. g., [5]). The shear failures become unstable and the rupture growth begins to accelerate after the energy released by the rupture growth equals to or becomes larger than the work necessary for producing new fracture surfaces. Finally, the rupture velocity, the velocity at which the rupture front is expanding, reaches a constant value comparable to the shear wave velocity. We call the process from the rupture initiation to rupture growth acceleration the earthquake nucleation process, and the portion of the fault where the rupture growth accelerates the nucleation zone. The important point is that the rupture velocity (and slip velocity) accelerates during the nucleation process, since it is a transient process from quasi-static to dynamic. One of the major differences between various models is whether the size of the nucleation zone is much smaller than the final fault length and is not detected by observations, and whether the size of the nucleation zone is different between large and small earthquakes.

If the size of the nucleation zone is extremely small and can be approximated as a point, the waveforms radiated from the earthquake are shown in Fig. 1a. Figure 1a displays the initial rise of the P-wave velocity pulse at a far-field (distant relative to the fault length) station, when the waveform propagates in a purely elastic homogenous medium with no attenuation and scattering under the following assumption: the rupture initiates at a point, the circular rupture front expands at a constant velocity, the stress on the fault abruptly drops from an initial level to a final level, and the stress drop is constant over the fault, as modeled by Sato and Hirasawa [60]. In this case, the initial portion of the P-wave velocity pulse is characterized by a linear rise (and the initial rise of the displacement pulse is quadratic with time, due to increasing circular fault area with time). The tangent of the linear rise is proportional to the stress drop and rupture velocity.

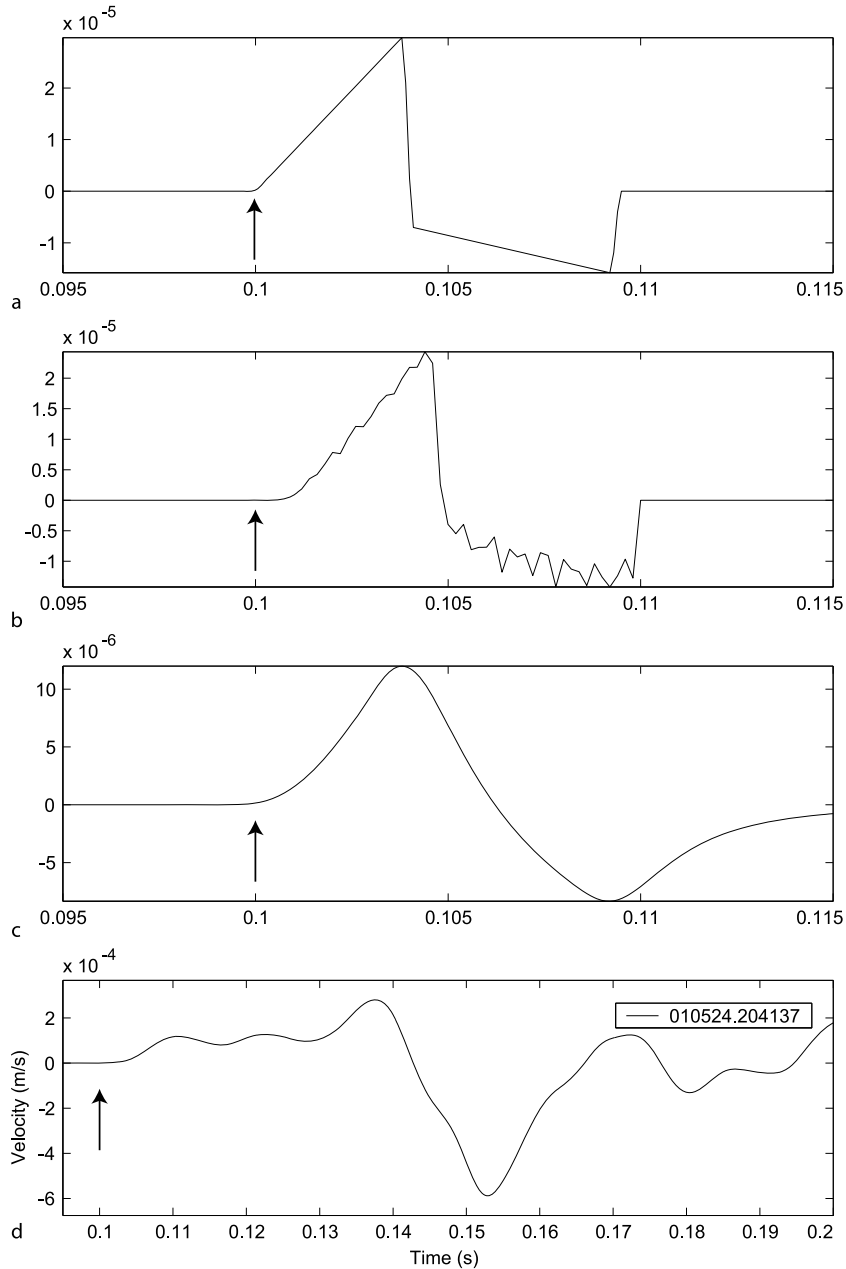
On the other hand, if the size of the nucleation zone is relatively large, the velocity pulse can be approximated as

shown in Fig. 1b. It is seen that the linear rise is delayed and the slope of the initial portion gradually increases. Theoretical models explain such a slow rise as follows. Shibazaki and Matsu’ura [64] demonstrated that the slow rise could be explained by gradually accelerating rupture and slip velocities in a relatively large nucleation zone. The size of the nucleation zone is controlled by the critical slip distance,  $D_c$ , that is the amount of slip necessary to drop the peak shear strength down to the dynamic frictional level (e. g., [19]). This frictional behavior is called slip-weakening, which is theoretically predicted (e. g., [15]) and is observed in laboratory experiments (e. g., [52,56]). According to their theory, since large slips are necessary to decrease the friction on faults with a large  $D_c$ , the rupture and slip velocities are not accelerated in the beginning on the faults. Sato and Kanamori [61] modeled the slow rise by Griffith’s fracture criterion based on the energy balance indicated above, as the rupture velocity gradually increases under the assumptions of a large pre-existing fault and small trigger factor (instantaneous small stress increment) on the fault. The pre-existing fault is often called the initial crack. In this model, large initial cracks result from large fracture energies on the fault and the rupture and slip velocities are not accelerated right after the rupture onset owing to the large fracture energies. In this paper, the model having a detectable size of the nucleation zone is called the preslip model, following Beroza and Ellsworth [9].

It is important to examine whether initial rises of observed velocity pulses are linear or gradually increasing, since the above two models predict different initial rises. Although the examination seems to be very easy at first glance, it is actually very difficult. The path effect is an inevitable obstacle to the examination, in particular for small earthquakes. The effects of anelastic attenuation and/or scattering contaminate observed waveforms and can reproduce a slow rise of observed waveforms even though the waveforms show a linear rise at the source. Figure 1c shows the initial rise of velocity pulses calculated by considering anelastic attenuation by a convolution of a Q operator [7]. It is found that the linear rise is delayed as shown in Fig. 1b.

Another serious obstacle is the complexity of observed waveforms. They are not often as simple and smooth as those shown in Figs. 1b and c, but complicated as shown in Fig. 1d. If several small patches in a relatively large nucleation zone break during the nucleation process [9,65] as shown in Fig. 2a, it is difficult to detect an acceleration of the rupture velocity from the complicated waveforms that include several small phases radiated from the breaks of the small patches. In other words, it is difficult to infer



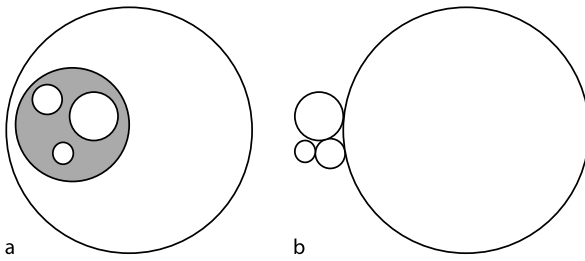


Earthquake Nucleation Process, Figure 1

Examples of P-wave velocity pulses. Arrows indicate onsets of P-waves. **a** Source pulse from the circular fault model [60]. Stress drop: 1 MPa, fault radius: 17 m, rupture velocity:  $0.8 V_s$ , take-off angle: 61 degrees. **b** Source pulse from a circular fault model with a rupture velocity accelerating with time for first 20% of the total duration time. The other parameters are the same as a. **c** Q convolved velocity pulse at a distance of 3.58 km from the source. Q is set as 300. The other parameters are the same as a. **d** Velocity waveform of a complicated shape observed in the Western Nagano Prefecture region

the nucleation process from the complicated waveforms. On the other hand, the model with small nucleation zones claims that the complicated waveforms are radiated only from successive breaks of small sub-events and a delayed

break of the largest sub-event, as shown in Fig. 2b. This concept is known as the cascade model (e. g., [1,9,23,78]). In this model, a former sub-event triggers successive larger sub-events and the final event is the largest among de-



Earthquake Nucleation Process, Figure 2

Schematic source models in which small amplitude phases are generated by breaking of small fault patches within the nucleation zone (*shaded portion*) **a**, and as successive small subevents **b**. *Large and small circles* are the mainshock fault and small subevent fault patches, respectively

tectable sub-events. Although it is not always clear how larger sub-events are delayed from former smaller sub-events, the growth of the rupture can raise the possibility that larger sub-events are generated. For small earthquakes also, it is possible that very small sub-events trigger successive sub-events. In this case, their waveforms produced by the source are complicated, but the observed waveforms are likely to show a smooth initial rise as shown in Figs. 2b and c, due to the path effects. Thus, in this paper, both smooth initial rises of small earthquakes and complicated initial portions of large earthquakes are discussed in the same manner.

By the way, if the size of the nucleation zone is relatively large, it could be possible to detect a quasi-static rupture growth at the very beginning of the earthquake nucleation process by near-field broadband instruments, such as strain meters. However, it has not, to present, succeeded except for a few very large earthquakes, such as the 1964 Chile earthquake (e. g., [41]). This fact suggests the possibility that the size of the nucleation zone is too small to be recorded by strain meters or the duration of the quasi-static rupture process is much longer than a practical frequency range of strain meters. Furthermore, the observations for very large earthquakes were not explained by the nucleation process that occurs near the hypocenter but by slow slips on the downward extension of the seismogenic faults (e. g., [33,41]). Thus, we will examine data obtained by seismometers.

It is very important to carefully analyze observed waveforms recorded at a short focal distance by a wide-dynamic range and frequency range. This paper reviews various studies related to the earthquake nucleation process, indicates the problems about these studies, and summarizes the current observations. First, in Sect. “**Introduction**”, contributions from earthquake early-warning systems are

reviewed, since they are directly influenced by the problems cited above. In Sect. “**Contribution from the Development of Earthquake Early-Warning Systems**”, important studies about the initial rupture process are reviewed basically in the order in which they were published. In Sect. “**Observations of Initial Rupture Processes**”, a probable model for the process will be proposed based on the reviews in the former sections.

### Contribution from the Development of Earthquake Early-Warning Systems

To solve the problem of how earthquakes initiate and grow larger, the most straightforward approach is a comparison between initial rises of observed waveforms of large and small earthquakes. Important studies were done for the development of earthquake early-warning systems which hope to forecast, as early as possible, the final size of earthquakes from observed waveforms, from the very beginning parts of seismograms. Olson and Allen [57] applied the method by Allen and Kanamori [3] to a new dataset including many large earthquakes ( $M_w > 6$ ) and claimed that the final size of earthquakes ( $3 < M_w < 8$ ) can be estimated from waveforms of the first several seconds. The method is based on the scaling relationship between the predominant period of waveforms and the final size of earthquakes, which was first derived by Aki [2]. To implement the scaling relationship in earthquake early-warning systems, Allen and Kanamori [3] adopted an algorithm to estimate the period from a short waveform, computing regressively the ratio of velocity and displacement amplitudes (modified after Nakamura [49]).

For smaller earthquakes, it is quite natural to be able to estimate the final size from the first several seconds of the waveforms, since the waveforms cover the entire source duration. The problem is estimating large earthquakes ( $M_w > 5.5$ ). Rydelek and Horiuchi [58] evaluated the statistical significance of the results by Olson and Allen [57] and reported that the results are not clear. Rydelek and Horiuchi [58] analyzed waveforms observed by Hi-net, a nation-wide high gain seismometer network operated by NIED in Japan, and found no trend between the period and earthquake size for larger earthquakes. On the other hand, Wu et al. [77] analyzed many waveforms recorded at the southern California Seismic Network stations and showed that the period increases with magnitude ( $4 < M < 7.5$ ). For earthquake early-warning system applications, where a quick response is essential, a precise waveform analysis is not required. A more careful analysis is needed to answer the question raised by Rydelek and Horiuchi [58].

## Observations of Initial Rupture Processes

### Measurements of Initial Parts of Observed Seismograms

#### How are Initial Portions of the Waveforms Measured?

Observed waveforms of large to small earthquakes are measured in various manners in the studies reviewed in the following sections. Sometimes special names are given to a characteristic phase of the waveforms. First, the problems concerning the measurement of observed waveforms will be discussed.

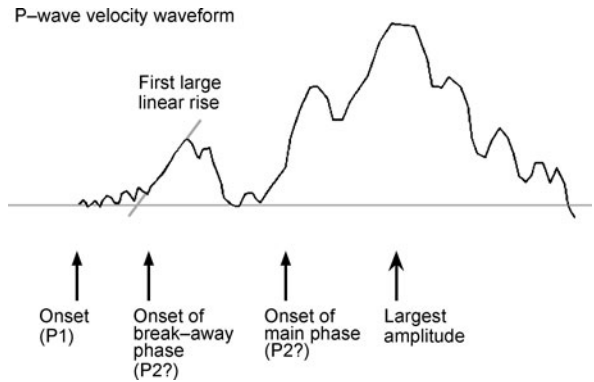
Many studies try to find which phase is radiated from the source process with a constant rupture velocity and slip velocity, since the nucleation process is basically an accelerating rupture process. Since the nucleation process is a beginning rupture process, it is likely that the rupture front just after the nucleation process expands nearly circularly. Consequently, the phase following the nucleation phase is thought to show a linear rise in velocity waveforms. For this reason, the first linear rise in velocity waveforms is extensively investigated in various studies.

Various studies also measure a pulse of the largest amplitude or a group of pulses including the largest amplitude, which is called the main phase in this paper. Amplitudes of waveforms are thought to show the maximum value after they reach terminal velocity, which is the maximum rupture velocity determined by the shear wave velocity. However, the largest amplitude does not necessarily occur just after rupture velocities reach the terminal velocity; all the waveforms preceding the maximum phase do not necessarily reflect the nucleation phase.

The problem is determining which phase is radiated just after rupture velocities reach terminal velocity. Furthermore, since slip velocities can cover a wide range of magnitude, it is important to investigate whether slip velocities are similar to the average values of earthquakes. In the following, when it is necessary to explicitly indicate the phase generated by faulting with a nearly constant rupture velocity and dynamic stress drop that are representative of average values for earthquakes, the phase is called the ordinary phase.

As schematically shown in Fig. 3, for large earthquakes, the first linear rise is not always a part of the main phase, which has the largest amplitude. For small earthquakes, the main phase often shows a linear rise. However, it is possible that the main phase is not the ordinary phase, as we discuss in a later section.

**Large Earthquakes** One of the first studies that indicated the importance of small amplitude waveforms preceding the main pulse is Furumoto and Nakanishi [24],

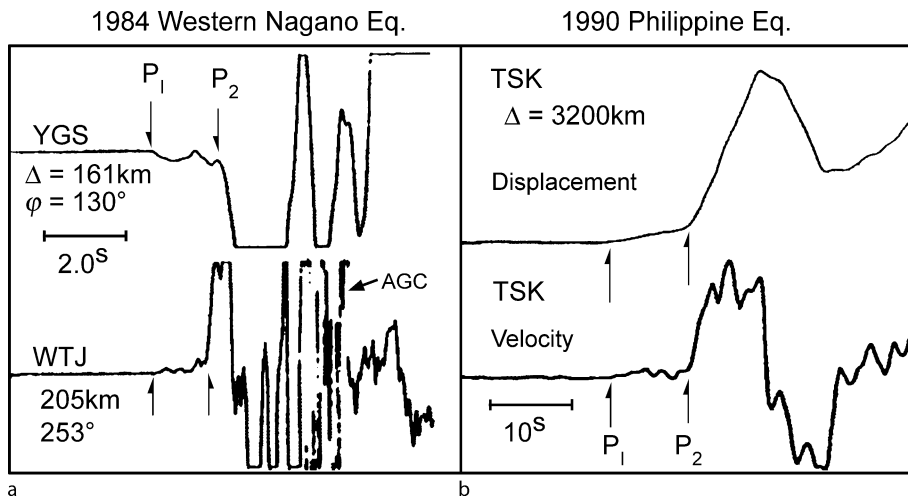


Earthquake Nucleation Process, Figure 3  
Schematic illustration of large and small earthquakes to show the first linear rise and the main phase

which analyzed long-period waveforms from a worldwide network; this was followed by many similar studies (e. g., [4,14,68]). Umeda [72,73] analyzed broadband seismograms at local, regional and teleseismic distances and found that small amplitude waveforms preceded large amplitude waveforms, as shown in Fig. 3. Umeda [72] called the onsets of the P-waves and large amplitude waveforms P1 and P2, respectively. He originally thought that the large amplitude waveforms were generated by somewhat special processes that characterized large earthquakes, e. g., breaks of many small faults triggered by both static and dynamic stress concentrations due to the preceding fault motion and/or a dynamic connection of separated faults [74], while the smaller waveforms were radiated from ordinary smooth rupture propagation. His important result is the relationship between the duration of P2-P1 and magnitude ( $3 < M < 7$ ), suggesting earthquake sizes are scaled with P2-P1, the duration of small amplitude waveforms.

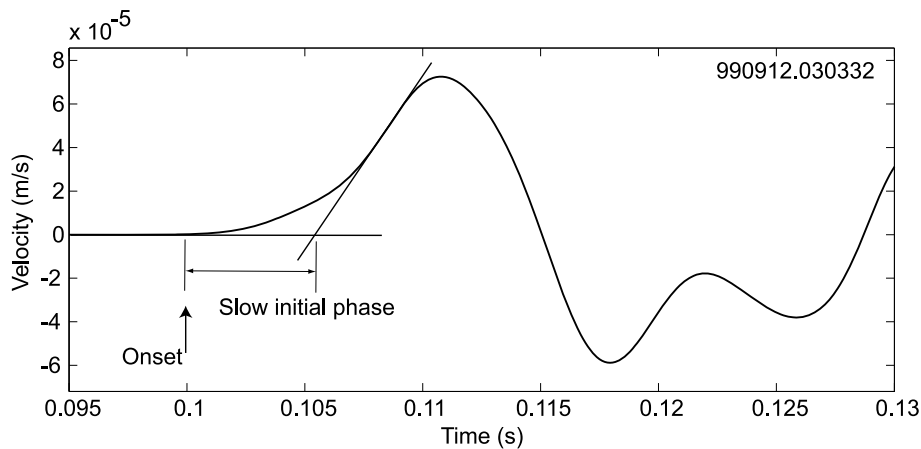
The determinations of P2 seem to be reasonable from his figures. However, the criterion of P2 determination is not necessarily clear and quantitative compared to the studies cited in the following sections. In many cases it seems that P2 is the onset of the main phase that includes the largest amplitude, in particular for smaller earthquakes, but this point is not clearly shown in his papers. Furthermore, it is seen that absolute velocity amplitudes after P1 do not show an accelerating increase with time, although both the preslip model and the cascade model basically indicate that velocity amplitudes should have a growth that is faster than linear with time.

One of the important problems raised by Umeda [72, 73] is identification of the ordinary phase, the main phase or preceding smaller amplitude waveforms. The ordinary



Earthquake Nucleation Process, Figure 4

Initial parts of observed waveforms of large earthquakes [73]. P1 and P2 indicate the onsets of P-waves and the large amplitude phases, respectively. Velocity waveforms are shown in a and the lower panel of b. Figure 7 in [73], copyright 1992 by Elsevier Science Publishers B.V.



Earthquake Nucleation Process, Figure 5

Slow initial phase of the velocity pulse observed in the Western Nagano Prefecture region and the measurement of the duration of the slow initial phase relative to the main phase that shows a linear increase of velocity amplitudes. The vertical arrow indicates the onset of P-waves, determined by their amplitudes. The portion indicated by the horizontal arrow is defined as the slow initial phase. The inclined line is the tangent at the maximum slope. The same pulse is shown in Fig. 7 by the black line

phase is the phase generated by faulting with a nearly constant rupture velocity and a dynamic stress drop, which are representative of average values for earthquakes. Furumoto and Nakanishi [24] regarded the main phase as the ordinary phase with a special rupture process occurring before the main phase, while in Umeda [72,73] the opposite is the case.

**Small Earthquakes** Iio [30,31] analyzed waveforms of microearthquakes recorded at very short focal distances

(down to a few hundreds of meters) using instruments with a wide-frequency response and found that the initial rises of velocity pulses did not show a linear increase but, rather, a convex downward (or upward) shape as approximated by  $t^n$  ( $t$  is the time measured from the onset and  $2 < n < 4$ ). He termed the initial rise “the slow initial phase”, and measured the duration of the slow initial phase relative to the main phase that shows a linear increase, as shown in Fig. 5. He found that the duration of the slow initial phase is proportional to the earthquake size. Al-

though the slow initial phase of microearthquakes could be the effect of anelastic attenuation as shown in Sect. “**Introduction**”, Iio et al. [32] analyzed velocity pulses recorded at a 10 kHz sampling frequency at numerous stations at short focal distances using an instrument with a wide dynamic range and frequency range, and concluded that the slow initial phase mainly reflects the source process for M2 events.

These observations are unique and important, but their interpretations and implications contain several fundamental problems and might cause some misunderstandings. First, it was intuitively regarded that the main phase showing the linear rise was the ordinary phase radiated from a circular fault with constant rupture and slip velocities of ordinary magnitudes. This was derived from the interpretation that the slow initial phase reflected the nucleation process in which the rupture and/or slip velocities gradually accelerated. However, the interpretation has not been thoroughly examined. As emphasized by Ellsworth and Beroza [21] and Beroza and Ellsworth [9], waveforms recorded by seismometers basically reflect slips on the fault that have accelerated above a certain level.

The second problem is the meaning of the scaling relationship indicating that larger earthquakes have a longer slow initial phase. Although the possibility was suggested from this relationship that larger earthquakes had a larger nucleation zone, this is true only if the main phase is the ordinary phase and the slow initial phase reflects the nucleation process. The relationship could be also misunderstood as larger earthquakes begin more slowly. The observational results mean only that the portion of velocity pulse approximated by  $t^n$  ( $2 < n < 4$ ) is longer for larger earthquakes. Furthermore, Iio [30,31] mentioned nothing about the difference between the slopes right after the onsets of larger and smaller earthquakes. Thus, another possible explanation is that for larger earthquakes, the slope of the velocity pulse increases with time for a longer period and the maximum slope becomes larger, even though the initial slope is the same as that for smaller earthquakes. This possibility implies that the final size of earthquakes is not estimated only from the initial part of velocity pulses, as pointed by Mori and Kanamori [48], although larger earthquakes have a greater dynamic stress drop. Several studies have indicated that the initial rises of large earthquakes ( $M > 6$ ) are similar to small events occurring in the vicinity of their hypocenters (e. g., [12]), while other studies have indicated that larger earthquakes display a larger initial rise (e. g., [35,50]). However, it is possible that the initial parts of these waveforms mainly reflect the initial rise radiated from a small patch within the initial rupture fault, as shown in Sect. “**Introduction**”.

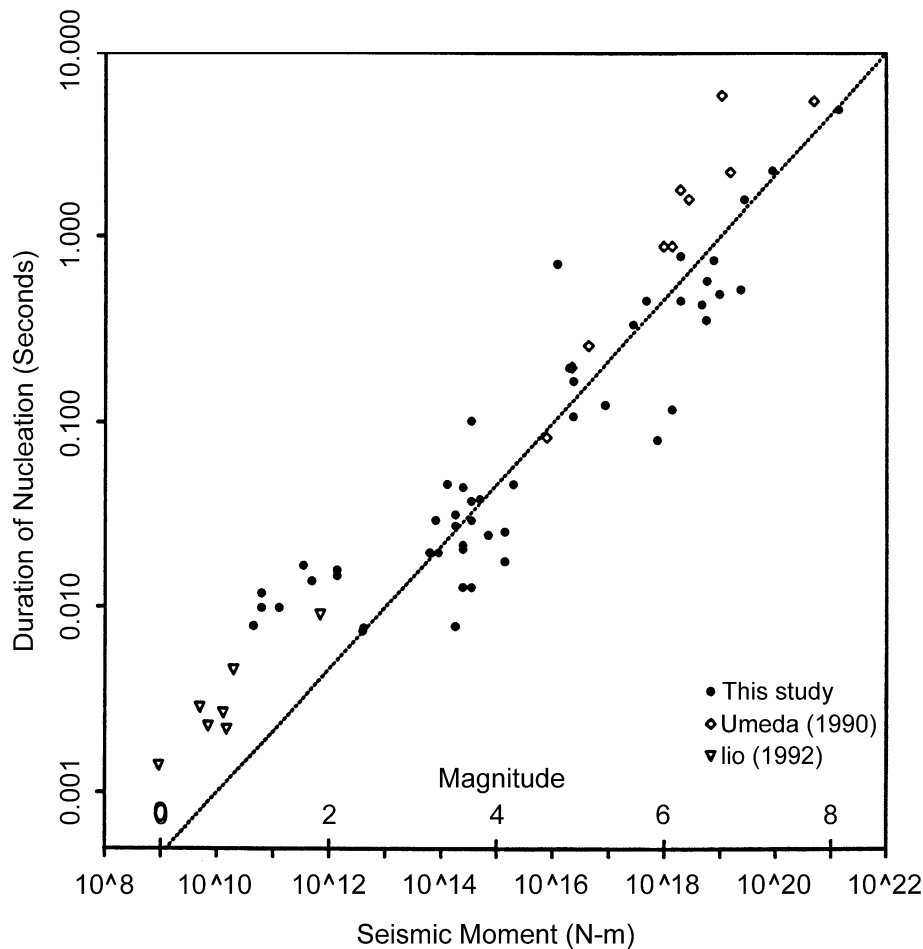
In that case, the results did not negate the existence of an observable earthquake nucleation process for large events. Further, these results were obtained only for a few earthquake sequences and need to be systematically examined for a greater data set.

The third problem is the propagation effects from the source to the observation stations. Iio [30,31] thought that observed waveforms basically reflect the characteristics of the source process in which the rupture and slip velocities gradually increase during the slow initial phase, assuming that source pulses of small earthquakes are simple and consist of a single event. However, it is possible that anelastic attenuation modifies two connected linear trends of small and large slopes from two subevents, creating a smoothed waveform of convex downward shape. This point is qualitatively examined in a later section. Furthermore, it might be possible that waveforms near the source are complex; several very short duration small pulses may precede the main phase, since short duration pulses can be smoothed by the path effect to produce a slow initial phase.

### Estimate of Source Time Functions

Ellsworth and Beroza [21] and Beroza and Ellsworth [9] focused on initial parts of seismograms of a very wide range of earthquake sizes ( $1 < M_w < 8$ ) and estimated source time functions. They paid attention to the first large linear rise of velocity pulses and deduced that the signal before the linear rise reflected the earthquake nucleation process, since a linear rise of velocity pulses is characteristic of the waveforms radiated from an ordinary circular fault model. They termed the linear rise “the breakaway phase” and the portion before “the seismic nucleation phase”. The term “seismic” is added, since waveforms detected by seismometers are radiated from dynamic slip on the fault. The criterion for the detection of the breakaway phase is not necessarily quantitative, but it seems from their figures that their classifications are reasonable. As shown above, they regarded the breakaway phase as the ordinary phase, since they roughly estimated that dynamic stress drops of the breakaway phases were several or several tens of MPa, that is comparable to average values of earthquakes. However, it is possible that the seismic nucleation phase is the ordinary phase, since the rupture velocity and slip velocity during the seismic nucleation phase is not quantitatively estimated.

Kilb and Gomberg [43] analyzed initial portions of the waveforms of the Northridge earthquake, which were also used in Ellsworth and Beroza [21] as a typical example, and claimed that the initial portions are very similar to those from nearby small earthquakes. They inferred from these



Earthquake Nucleation Process, Figure 6

Relationship between the duration of the seismic nucleation phase and the seismic moment, together with the data of the slow initial phases by Iio [30,31] and P2-P1 by Umeda [72,73] (Beroza and Ellsworth [9])

results that the cascade model is plausible. However, their results did not exclude the preslip model in which the rupture and slip velocities are accelerating in the nucleation zone, since they analyzed only the very beginning portions and not the following portions that might result from slow slips in the nucleation zone. Shibazaki et al. [67] discussed the possibility that large earthquakes begin by a breaking of a small patch in the nucleation zone. As discussed in the Introduction, the breaking of small patches can mask the nucleation process.

Ellsworth and Beroza [21] demonstrated the relationship between the duration of the seismic nucleation phase and the seismic moment, together with the data of the slow initial phases by Iio [30,31] and P2-P1 by Umeda [72,73], as shown in Fig. 6, and concluded that the duration of the seismic nucleation phase scaled with the seismic moment. However, it is seen in Fig. 6 that the data of Iio [30,31]

and Umeda [72,73] are shifted upward from the regression line. It seems that their data for small earthquakes also shifted upward from the regression line. The offsets of these shifts are about 1 order of magnitude. For larger earthquakes, this may be because Umeda [72,73] determined P2 as the onset of the main phase, while Ellsworth and Beroza [21] determined onsets of the first large linear rises. For smaller earthquakes, Iio [30,31] also detected the onsets of the main phases. They regarded the phases as the first large linear rises, but it is possible that the onsets of the first large linear rises are earlier than the main phases, as discussed in the following section.

### Estimate of Source Processes

**Small Earthquakes** In order to solve the problems cited in the previous sections, several studies tried to estimate

the source process that produces the initial phases. Hiramatsu et al. [26] analyzed rising parts of seismograms using the model of Sato and Kanamori [61], in which larger initial cracks generate a slower initial phase under a small triggering factor.

It was inferred from deep borehole (1800 m) seismograms by Hiramatsu et al. [26] that initial rises of velocity pulses of 5 events were explained by the ordinary circular fault model [60], namely the initial crack is too small to be detected, while the other 7 events needed a large initial crack. They first estimated source processes of the beginning of such small earthquakes; however, some basic problems remained. The first is that they modeled the first half cycle of velocity pulses. Since even portions of the first half cycle of the waveform can be affected by rupture arrest, it is not reasonable to fit the first half cycle of the waveform by using a model that does not include a reasonable rupture stopping process. Furthermore, a closer inspection of their results (Fig. 5 of [26]) reveals that some of the waveforms modeled with an initial crack do not display a smooth increase of the rising slope but appear to consist of two linear phases with different slopes. It is possible that these earthquakes result from a cascade rupture of a first and second sub-events with smaller and larger dynamic stress drops, respectively. For the data analyzed by Iio [30,31], waveforms with a predominantly longer slow initial phase appear to show a similar feature. Furthermore, it is not clear which is the ordinary phase, the first or second linear phase.

The problem of identifying the “ordinary phase”, the initial or main phase, is again pointed out. Recently, Iio et al. [34] analyzed the 10 kHz sampling data [32,36] and obtained results that suggest the initial rise is the ordinary phase.

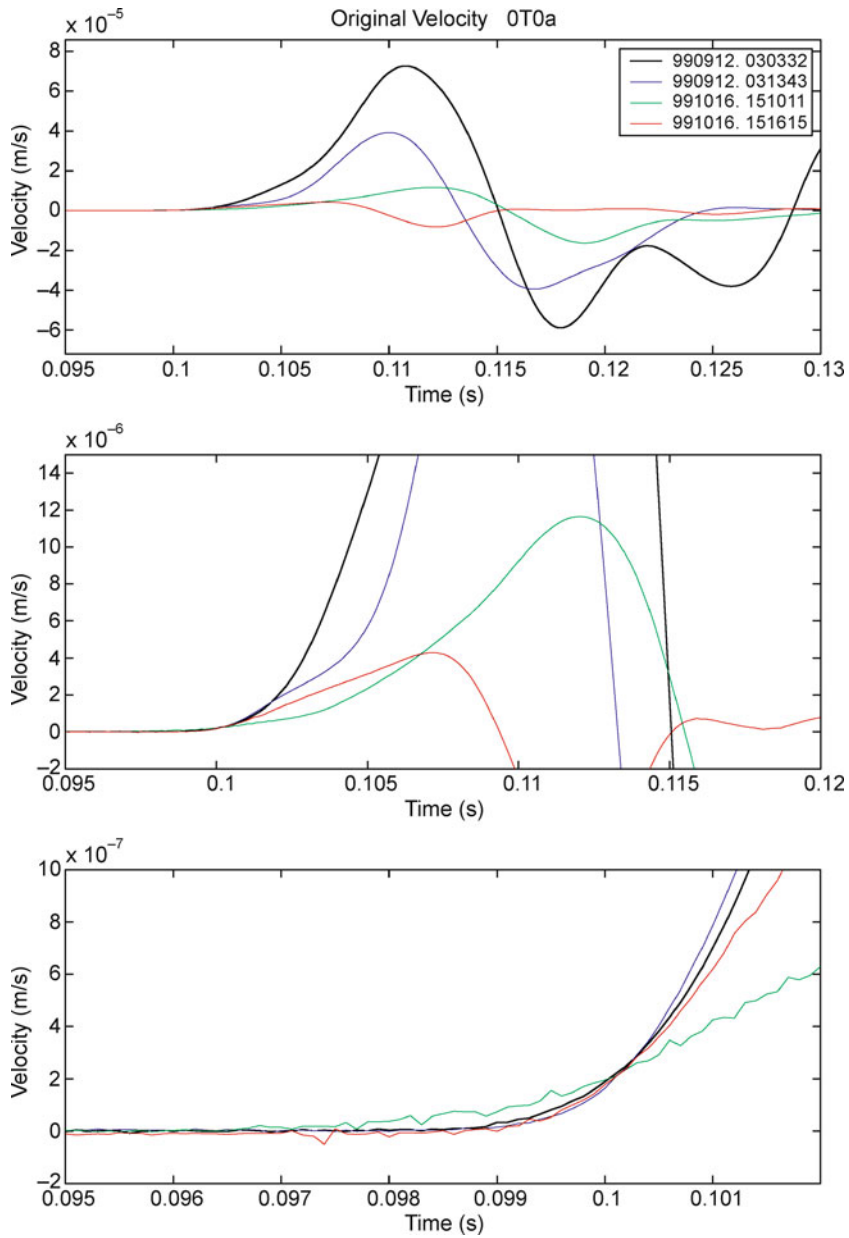
Figure 7 shows velocity pulses of microearthquakes ( $0.5 < M < 2.0$ ) for which relative locations are estimated within 100 m and fault plane solutions are similar [34]. These pulses were recorded at a borehole station at a depth of 800 m in the Western Nagano prefecture region [37,76]. The focal distances are about 3.6 km. It is seen that the smallest event (shown in red) shows a linear rise except for the first 1 ms, while slopes of the other events increase with time. Thus, if the linear rise is explained by a circular fault model of the Sato and Hirasawa type with an ordinary rupture velocity and stress drop, the increasing slopes reflect increasing slip velocities accelerated beyond ordinary values.

The linear rise of the waveforms was modeled by various kinematic fault models. Figure 8 displays the comparison of synthesized and observed velocity pulses for three borehole stations at depths of 800, 150, and 100 m [37].

The fault plane was determined from focal mechanisms and hypocentral distributions around these events. Since all three waveforms were not explained simultaneously by a Sato and Hirasawa type of circular fault model, they used a fan fault of various angles. Further, they did not assume that the slip terminated simultaneously over the fault, but that it begins to stop at a point on the fault edge and the stopping phase propagates circularly at a constant velocity. The stress drop, rupture and stopping phase expanding velocities, fault radius, fan angle and fan fault geometry were determined by a grid search technique.  $Q$  values are set as 300 by trial and error, to fit the very beginning initial rises with duration of about 1 ms.

It is seen that the observed waveforms in the first half cycle are well explained by the fan fault model, except for the middle trace, which might be contaminated by a surface reflection. The rupture velocity was estimated as  $0.8 V_s$  (the shear wave velocity), which is similar to that for large earthquakes. The stress drop was estimated as about 5 MPa by the formula for the circular fault of the equivalent fault area. Although they modeled only two events, they found that a few percent of events that occurred within about 5 km from the 800 m borehole station also showed such a linear rise. Furthermore, the other events generally show a steeper initial rise than the linear rise events. Consequently, it is likely that small earthquakes occurring in the Western Nagano prefecture region have a rupture velocity equivalent to those of large earthquakes even in the initial rupture process. This suggests that the slow initial phase of small earthquakes does not reflect the nucleation process that is characterized by accelerating rupture velocity, since the rupture velocity has already accelerated to a shear wave velocity in an early stage of rupture growth.

What do steeper main phases of larger earthquakes in the Western Nagano region reflect? In order to clarify this problem, Miura et al. [46] analyzed waveforms of earthquakes ( $0.0 < M < 4.0$ ) occurring from 1996 to 2003 in the Western Nagano prefecture region. They selected M3 events ( $3.0 < M < 4.0$ ) and identified 21 earthquake clusters that consist of M3 events and earthquakes occurring within 500 m of the hypocenters of M3 events. They investigated P-wave velocity pulses observed at three borehole stations for each cluster, in particular the difference in pulse shapes of large and small earthquakes. They found that waveforms of half of the clusters displayed complicated shapes, as shown in Fig. 1d, which are characterized by more inflection points than simple pulses [63]. The other half showed simple waveforms that enabled them to clarify the difference between large and small earthquakes. One example of waveforms is shown in Fig. 9. The ob-



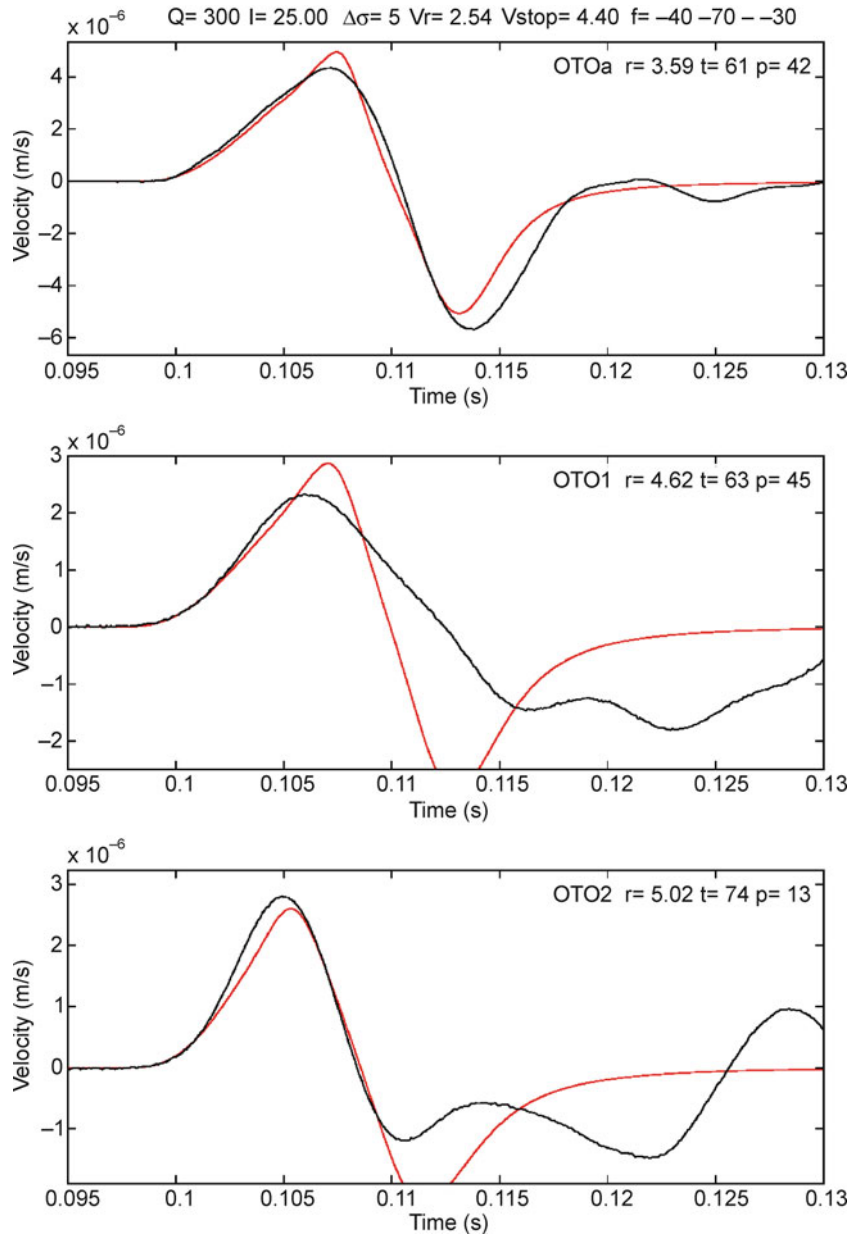
Earthquake Nucleation Process, Figure 7

Comparison of velocity pulses of microearthquakes ( $0.5 < M < 2.0$ ) of which relative locations are estimated within 100 m and fault plane solutions are similar, recorded at a borehole station at a depth of 800 m in the Western Nagano prefecture region [34]. Focal distances are about 3.6 km

served P-wave velocity pulses of  $0.5 < M < 3.1$  are displayed in different magnifications. It is seen that initial rises are similar for the first 1 ms but that slopes of larger events increase with time. The theoretical pulses from the circular fault model [60] are synthesized at the same focal distances for a similar magnitude range in Fig. 9b, assum-

ing a constant stress drop independent of earthquake size, and a  $Q$  value of 230, which was obtained by modeling the waveform of the smallest event. Responses of seismometers are also included. The synthesized pulse shapes of larger events do not display a linear rise but a slight convex upward shape. It is found that a distinct difference between



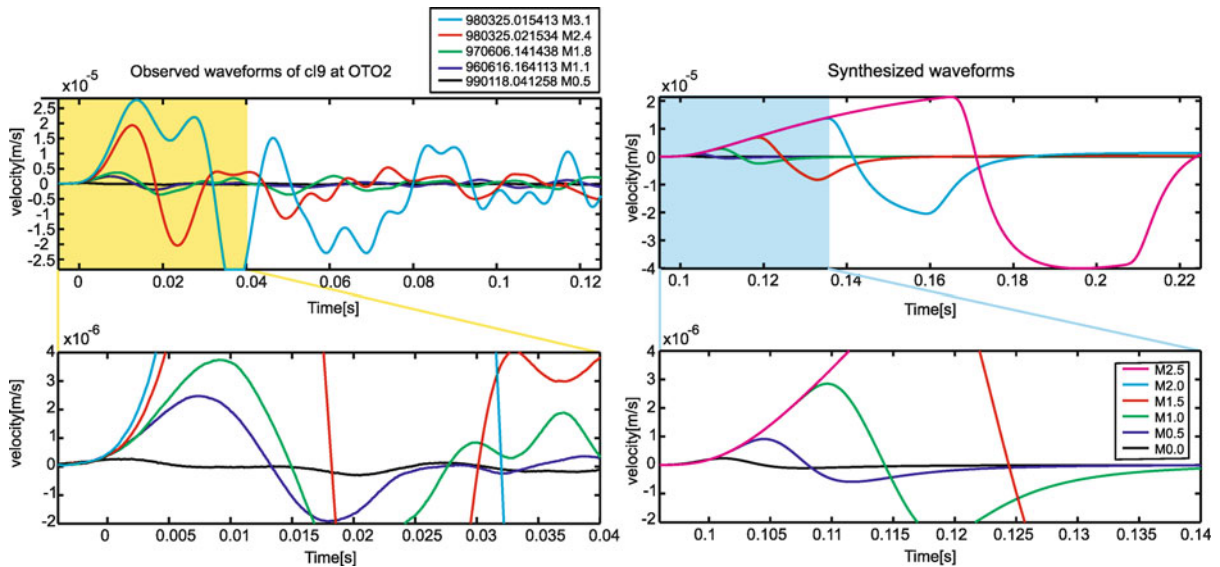


Earthquake Nucleation Process, Figure 8

Comparison of observed and synthesized waveforms for an event that displays a linear rise, shown in Fig. 7 by the red line [34]. Waveforms observed at three borehole stations of depths of 800, 150, and 100 m are shown in the top, middle and bottom panels, respectively. The synthesized waveforms are calculated by a fan fault model in the homogenous half space using the following parameters: The stress drop is estimated as about 5 MPa. The rupture and stopping phase expanding velocities are  $0.8 V_s$  and  $0.8 V_p$  (the P-wave velocity), respectively. Fault radius: 25 m, fan angle: 40. A Q value is 300

theoretical and observed waveforms is seen in the maximum slope of larger events. This figure clearly demonstrates that larger earthquakes have larger dynamic stress drops than smaller events. Similar features are also seen in the other clusters that show simple waveforms. These re-

sults imply that the similarity law does not hold for a small range of magnitudes in the Western Nagano prefecture region, as indicated by Venkataraman et al. [58]. The similarity law predicts pulses as shown in the right hand side of the figure.



Earthquake Nucleation Process, Figure 9

Comparison of observed (*left*) and synthesized (*right*) waveforms for events of an earthquake cluster that shows smooth waveform shapes [46]. The cluster consists of the M3 class events and earthquakes occurring within 500 m from their hypocenters. The synthesized waveforms are calculated by an ordinary circular fault model [60] at the same focal distance. A  $Q$  value of 230 was obtained by modeling the waveform of the smallest event. Responses of the instruments are also convolved

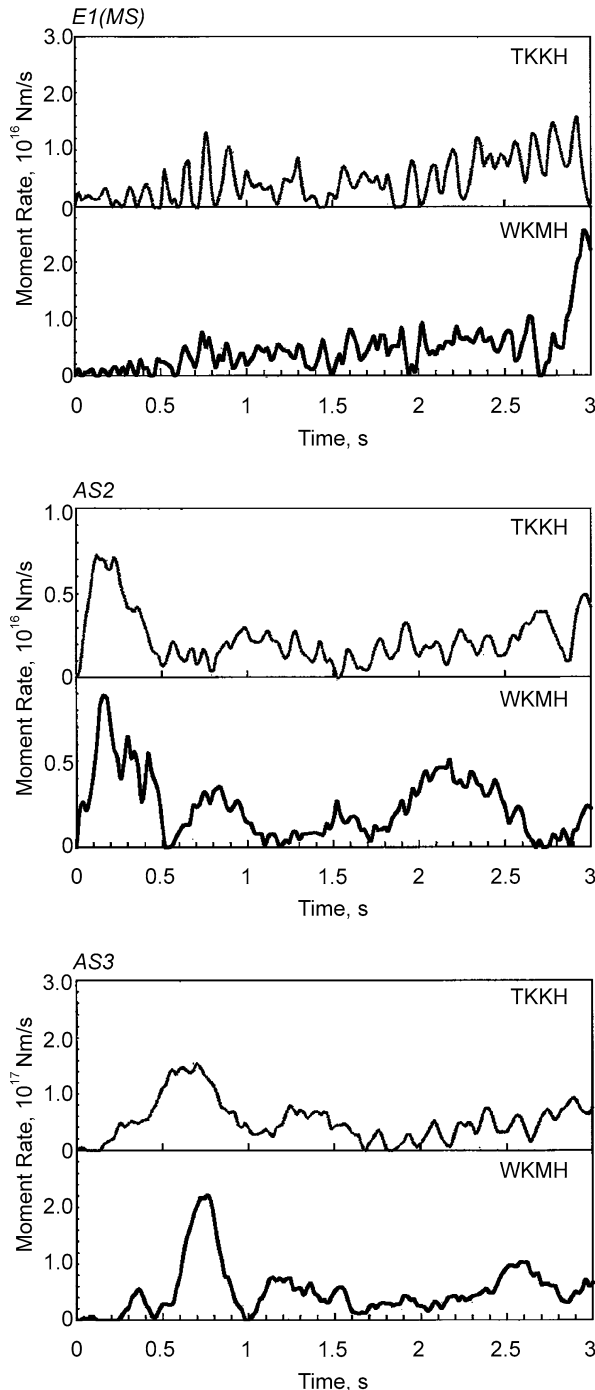
**Large Earthquakes** Sato and Mori [62] used the same method as Hiramatsu et al. [26] and analyzed the very beginnings of waveforms for a very wide range of magnitudes ( $3 < M < 8$ ) recorded by high gain short-period seismometers at local distances. They showed that large initial cracks (a few tens of meters) are necessary to explain the initial rises and that initial crack lengths are almost constant, independent of the eventual earthquake size. Although they might have analyzed only the beginning of the first sub-event in the nucleation zone, it is important that they found that very beginning portions of large earthquakes show an accelerating rise over the first 0.1 s.

For large earthquakes, initial rupture processes can be estimated by near-source broadband seismograms. Shibazaki et al. [67] determined a P2 phase 0.5 s after the P-wave onset (P1) for the Kobe earthquake and performed a waveform inversion using the initial portion of seismograms. They estimated that the slip velocity slowly accelerated for 0.5 s. The average rupture velocity within 0.5 s after the onset of dynamic rupture was about 2 km/s, which is not very low. However, their results possibly detect the earthquake nucleation process, since it is difficult to estimate accelerating rupture velocity in the small nucleation zone of about 500 m, which was inferred from the waveform inversion. It is noted that their estimate of P2-P1 of 0.5 s is consistent with the scaling relationship of Ellsworth and Beroza [21], but is not consistent with that

of Umeda [72,73]. The initial phase of the Kobe earthquake shows an accelerated increase of velocity amplitude with time [67], while a significant part of the data analyzed by the other studies do not show the accelerated increase. These results suggest the possibility that a major part of the initial rises analyzed in the other studies do not reflect the earthquake nucleation process.

Recently, longer durations of P2-P1 were observed for two intraplate earthquakes in Japan, the 2000 Western Tottori Prefecture earthquake (M7.3) and the 2004 West-off Fukuoka earthquake (M7.0). The initial phases of long durations are suitable for analyzing the rupture processes. For the Western Tottori Prefecture earthquake, Hirata [27] measured a duration of 2.5 s for P2-P1, determined the location of the P2 source and estimated the initial rupture process by a waveform inversion. They estimated that during the first 2.5 s, a small slip occurred within the limited area around the hypocenter and later, a large slip began near the location of P2 about 5 km south-east of the hypocenter. The average rupture velocity during the first 2.5 s was estimated to be 1.8 km/s from the waveform inversion of the initial phase. The stress drop in the limited portion was estimated to be small compared with the large slip area.

Since the duration of P2-P1 is long, the above features are also obtained in ordinary waveform inversions. Iwata and Sekiguchi [39] estimated the slip distribution



Earthquake Nucleation Process, Figure 10  
 Source time functions for the initial part of the 2004 West-Off Fukuoka earthquake (M7.0). The waveforms after the onsets are shown. The mainshock (*top*), a M4.5 aftershock (*middle*), and a M5.4 aftershock (*bottom*) estimated by an empirical green function method [75]. Zisin, 59:250, Figure 11, copyright 2007 by the Seismological Society of Japan

using strong ground motion seismograms and geodetic data, and found a large slip area located about 5 km southeast of the hypocenter. They regarded the large slip area as an “asperity” on the fault. It has been well known that slip distributions of large earthquakes are heterogeneous and large slips occur within a limited portion on the fault (e. g., [44]). Although the origin of large slip areas has not yet been clarified, one interpretation is that fault strength is higher in the areas of large slip than in the surrounding areas. This concept is called the asperity model. The observations about P1 and P2 shown above can be explained by the asperity model: The earthquake rupture begins at the weakest portion on the fault, then propagates to stronger portions where large slips occur.

For the 2004 West-off Fukuoka earthquake (M7.0), similar results were obtained. A P2 phase was detected by Yamaguchi et al. [75] and the duration of P2-P1 was long, estimated as 3.38 s. They determined the location of the P2 source as 3.44 km southeast of the hypocenter. The average rupture velocity during the 3.38 s period was computed to be about 1.02 km/s from the distance between the locations of P1 and P2. These results are consistent with those from ordinary waveform inversion studies (e. g., [6]). They also estimated the source time functions of the mainshock and a few large aftershocks by an empirical Green function method, as shown in Fig. 10. It was found that the source time function of the mainshock is not impulsive but has a gradual onset and long duration, while those of the aftershocks are impulsive. Only one aftershock shows a small initial phase, as shown in Fig. 10c. The seismic moment released before the P2 phase is comparable for those of M4 aftershocks. These results clearly show that the initial rise of the West-off Fukuoka Prefecture earthquake is different from those of the small aftershocks and is characterized by a small stress drop. It is suggested that large earthquakes do not accidentally grow larger.

For these two earthquakes with long initial phases, the average rupture velocities and stress drops on the initial rupture faults are small; however, the initial phases do not show a accelerating increase of velocity amplitude with time. These facts suggest the possibility that long initial phases do not reflect the nucleation process.

## Discussion

### Summary of the Observations

The reviews in the previous sections have revealed several important characteristics concerning the initial phases that are commonly seen in many studies. In this section, first, these characteristics are summarized, and then, an inferred initial rupture process will be discussed.

The almost linear initial rises were observed in the Western Nagano region and were well explained by a fan fault model with a constant rupture velocity comparable to the shear wave velocity. On the other hand, nearby larger earthquakes showed an initial rise in which the slope is of the same order of magnitude as that of the linear rise, plus increasing slopes with times of a few to several ms after the onset [34]. Thus, it is thought that the rupture velocity of these larger earthquakes is also comparable to the shear wave velocity during the initial phase. Furthermore, observed waveforms do not necessarily display a gradual increase of rising slope but sometimes show a discrete change in the slope, suggesting that the initial rise of the velocity pulse at the source is not smooth. Consequently, successive sub-events with larger dynamic stress drops possibly produce larger earthquakes for these events occurring in the Western Nagano Prefecture. These inferences are consistent with the results obtained by Hiramatsu et al. [26] that about half of their data are explained by an ordinary circular fault model and longer initial phases are not always smooth. For small earthquakes ( $M < 4$ ), the slow initial phase probably does not reflect the earthquake nucleation process. The nucleation size of small earthquakes is probably small and thus, the very beginning of observed waveforms should be analyzed very carefully considering path effects. This matter is beyond the scope of this paper and is left to future studies.

Larger earthquakes showed a variety of initial phases, as pointed out by Ellsworth and Beroza [21]. It is likely that the observed initial phase of the Kobe earthquakes reflects the earthquake nucleation process [67]. Also, that of the Northridge earthquake probably reflects the earthquake nucleation process, since the very weak initial phase shows an accelerated increase of velocity amplitudes with time [21], although the very beginning part of the initial phase is similar to the waveform of nearby small aftershocks [43]. Further, several data of Ellsworth and Beroza [21], in particular, those shifted downward from the regression line shown in Fig. 6 possibly reflect the earthquake nucleation process. However, it is likely that those of the other large earthquakes do not reflect the earthquake nucleation process, since they do not show an accelerated increase with time but rather are roughly flat. These facts are clear for the two intraplate earthquakes that have a relatively long initial phase (e. g., [27,75]). For these two earthquakes, the average rupture velocity during the initial phases are estimated as 1.02 to 1.8 km/s, slightly smaller than ordinary values, and the stress drops on these large initial rupture faults are estimated to be small (e. g., [27,75]). Ordinary wave-

form inversions for these earthquakes showed that the main phases were generated by the breaking of asperities (e. g., [6,28,39]).

As summarized above, a major part of the initial phases of large earthquakes do not show an accelerating increase of velocity amplitudes with time and it is thought that the rupture and slip velocities of these earthquakes are not accelerating during this time period. Consequently, it is thought that the initial portions of the observed waveforms of these large earthquakes do not reflect the nucleation process that is characterized by the transition from a quasi-static to a dynamic state; instead they represent a part of the dynamic rupture process characterized by a smaller stress drop, before the breaking of large asperities. Small earthquakes possibly show similar characteristics with an initial rupture process that is characterized by an ordinary rupture velocity and much smaller dynamic stress drop than the main phase. Although the nucleation process is probably seen in the very beginning stage of the large earthquakes analyzed by Sato and Mori [62] and Shibasaki et al. [67], it is likely that the scaling relationships between the duration of the initial phases and the rupture size shown by Umeda [72,73], Iio [30,31], and a major part of the data Ellsworth and Beroza [21] do not reflect the nucleation process but a part of the dynamic rupture process before the breaks of relatively large asperities.

### A Possible Model

As summarized in the previous section, since a major part of the observed data do not reflect the earthquake nucleation process, in particular the data from small earthquakes, we cannot examine the preslip model here.

The observed data do not match the cascade model, since the cascade model basically predicts self-similar fault breaks for successive events, not increasing dynamic stress drop events. A variation on the cascade model, the hierarchy fault model explains increasing slopes of initial rises by the abrupt increase of the moment rate function due to a longer fault edge at a higher hierarchy level [23]. However, as shown in Fig. 6, the difference between the slopes of larger and smaller earthquakes can be greater than one order of magnitude, so it is difficult to explain the observed data only by the change in rupture front geometry. It may be necessary to consider differences in dynamic fault parameters.

So the questions to be answered are, why do larger earthquakes have a longer slow initial rupture process? In the first place, why do earthquakes need an initial rupture process to break a stronger portion on the fault? Why don't

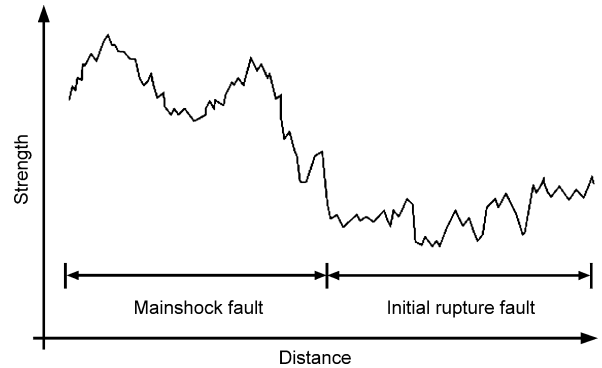
earthquakes initiate as a breaking of an asperity without the initial phase?

The key to solving these problems may lie in the studies of the two intraplate earthquakes with longer initial phases. The important observations about these earthquakes are the geometries of the initial rupture fault and mainshock fault. For the 2000 Tottori earthquake, it is found that azimuths of the initial rupture fault and mainshock fault are estimated as  $N135^\circ E$  and  $N150^\circ E$ , respectively [51]. Since the direction of the maximum compressional stress is estimated as  $N90^\circ-100^\circ E$  [42], it is found that the initial rupture fault is favorably oriented, while the mainshock fault is unfavorably oriented. For the 2004 West-off Fukuoka prefecture earthquake, the azimuth of the initial rupture fault is different from that of the mainshock fault by 20 degrees (e. g., [38,69,71]). A more obvious example is obtained for the Landers earthquake, where the Emerson fault with the largest slip is unfavorably oriented, compared with the faults that had broken before (e. g., [25]). These observations indicate that a larger slip occurred on an unfavorably oriented fault. It is possible that the mainshock fault, which generates the main phase, generally has a higher strength than the initial rupture fault. This may be the reason why larger dynamic stress drops occurred on faults of successive larger events.

If this proposed strength profile on the initial and mainshock faults holds for all the earthquakes, the above questions can be re-written: Why is a larger weak initial rupture fault necessary to break a larger strong mainshock fault?

A similar problem has been discussed by Ohnaka [55]. He derived the relationship between the critical slip distance  $D_c$  and the asperity size from laboratory experiments and the physics of contacts on faults. Since  $D_c$  is related to the size of the nucleation zone (e. g., [52]) and it is inferred from the results of waveform inversions that the asperity size is proportional to the total fault length (e. g., [47]), his relationship might be regarded as the relationship between the lengths of initial rupture fault and the mainshock fault. However, the  $D_c$  used in his relationship is that of the asperity, not of the initial rupture fault. Further, it is inferred from the above discussion that a major part of observed initial phases probably do not reflect the nucleation process (namely,  $D_c$ ).

It is important to clarify strength profiles along faults; however, we do not presently have enough information. In the following, we will assume the fault strength is controlled by the geometry of the fault surface, as estimated for the Tottori earthquake. More concretely, it is assumed that the angle between the tangent of the local fault surface and the direction of the uniform principal stresses



Earthquake Nucleation Process, Figure 11  
Schematic illustration of the strength profile along the fault deduced from a fractal geometry of fault surface. The rupture initiates at the weakest portion on the initial rupture fault and then propagates to a portion of higher strengths

controls the fault strength at each point. In this case, it is deduced that the fault strength shows a fractal-like distribution along the fault as shown in Fig. 11, since the geometry of fault surface is thought to be fractal [45]). The fault strength profile should include various wavelengths, but here we will consider the longest wavelength to investigate the interaction between the asperity and initial rupture fault, since amplitudes of a longer wavelength are thought to be larger on fractal fault surfaces (e. g., [11]).

Under the above assumptions, the asperity and initial rupture fault are attributed to the portions of higher and lower strengths on the strength profile, respectively. Strength profiles of faults of larger earthquakes are thought to have a longer wavelength. Actually, it is empirically derived that the asperity size is proportional to the total fault length [47]. Consequently, it is inferred from these results that larger earthquakes have larger initial rupture faults. On initial rupture faults, the fault strength increases with distance from the hypocenter, since the hypocenter is the weakest point on the fault. In other words, the strength at a rupture front increases with rupture growth. Thus, it is thought that the initial rupture does not expand smoothly and the slip velocity does not significantly accelerate, but a large asperity can break after a breaking of the initial rupture fault. This is only a possible qualitative model for the initial rupture process and it suggests a possibility for explaining the observed data. It should be examined by extensive studies.

We do not presently have any clear answers to the subtitle question, 'Does the initiation of earthquake rupture knows about its termination?'. Even if the above model is correct, we also have to know the geometry of the fault surface and the time, slip and slip velocity dependent stress on

the fault for large and small earthquakes, in order to simulate the rupture propagation and must understand the factors that control the earthquake size.

### Future Directions

This paper reviewed studies analyzing the very beginning portions of observed waveforms of earthquakes and showed what we have presently clarified about the earthquake nucleation process. To make further progress, the most straightforward path is to investigate the initial rupture process by precise waveform inversions for large earthquakes. In this case, new inversion methods, as proposed by Uchide and Ide [70], are useful. Furthermore, it is necessary to use broadband near field data, since it is possible that very slow slips occur on the initial rupture fault in association with the initial rupture process or for a very long time before the initial rupture process. For smaller earthquakes, high resolution data, as those in the Western Nagano Prefecture region, are necessary for investigating their rupture processes. These extensive studies about the initial rupture process can clarify the true feature of the earthquake nucleation process.

### Acknowledgments

The project in the Western Nagano Prefecture is a co-operative study with Shigeki Horiuchi, Shiro Ohmi, Hisao Ito, Yasuto Kuwahara, Eiji Yamamoto, Kentaro Omura, Koichi Miura, Bun'ichiro Shibazaki, and Haruo Sato. We thank James Mori and Masumi Yamada for their critical reviews of the manuscript. This work is partly supported by JSPS.KAKENHI (19204043), Japan. We are grateful for two anonymous reviewers for their critical and thoughtful comments.

### Bibliography

- Abercrombie R, Mori J (1994) Local observations of the onset of a large earthquake, 28 June 1992. Landers, California. *Bull Seismol Soc Am* 84:725–734
- Aki (1967) Scaling law of seismic spectrum. *J Geophys Res* 72:1217–1231
- Allen RM, Kanamori H (2003) The potential for earthquake early warning in Southern California. *Science* 300(5620):786–789
- Anderson JG, Bodin P, Brune JN, Pince J, Singh SK, Quaaas R, Ohnate M (1986) Strong ground motion from the Michoacan, Mexico, earthquake. *Science* 233:1043–1049
- Andrews DJ (1976) Rupture velocity of plane strain shear cracks. *J Geophys Res* 81:5679–5687
- Asano K, Iwata T (2006) Source process and near-source ground motions of the 2005 West Off Fukuoka Prefecture earthquake. *Earth Planet Space* 58:93–98
- Azimi SA, Kalinin AV, Kalinin VV, Pivovarov BL (1968) Impulse and transient characteristics of media with linear quadratic absorption laws. *Izv. Earth Phys* 1968(2):88–93
- Bak P, Teng C (1989) Earthquakes as self-organized critical phenomenon. *J Geophys Res* 94:635–15–637–15
- Beroza GC, Ellsworth WL (1996) Properties of the seismic nucleation phase. *Tectonophysics* 261:209–227
- Boatwright J (1978) Detailed spectral analysis of two small New York State earthquakes. *Bull Seismol Soc Am* 68:1117–1131
- Brown SR, Scholz CH (1985) Broad bandwidth study of the topography of natural rock surfaces. *J Geophys Res* 90:12575–12582
- Brune JN (1979) Implications of earthquake triggering and rupture propagation for earthquake prediction based on premonitory phenomena. *J Geophys Res* 84:2195–2198
- Cheng X, Fenglin Niu, Silver PG, Horiuchi S, Takai K, Ito H, Iio Y Similar microearthquakes observed in western Nagano, Japan and implications to rupture mechanics. *J Geophys Res* 112:B04306. doi:10.1029/2006JB004416
- Christensen DH, Ruff LJ (1986) Rupture process of the Chilean earthquake, 3 March 1985. *Geophys Res Lett* 13:721–724
- Das S, Scholz CH (1982) Theory of time-dependent rupture in the Earth. *J Geophys Res* 86:6039–6051
- Deichman N (1997) Far-field pulse shapes from circular sources with variable rupture velocities. *Bull Seismol Soc Am* 87:1288–1296
- Dieterich JH (1978) Preseismic fault slip and earthquake prediction. *J Geophys Res* 83:3940–3948
- Dieterich JH (1979) Modelling of rock friction: 1 Experimental results and constitutive equations. *J Geophys Res* 84:2161–2168
- Dieterich JH (1986) A model for the nucleation of earthquake slip. In: *Earthquake source mechanics*. Geophysical Monograph. Maurice Ewing Series, vol 6. Am Geophys Union, Washington DC, pp 37,36–47
- Dodge DA, Beroza GC, Ellsworth WL (1996) Detailed observations of California foreshock sequences: Implications for the earthquake initiation process. *J Geophys Res* 101(22):371–392
- Ellsworth WL, Beroza GC (1995) Seismic evidence for an earthquake nucleation phase. *Science* 268:851–855
- Ellsworth WL, Beroza GC (1998) Observation of the seismic nucleation phase in the 1995 Ridgecrest, California sequence. *Geophys Res Lett* 25:401–404
- Fukao Y, Furumoto M (1985) Hierarchy in earthquake size distribution. *Phys Earth Planet Inter* 37:149–168
- Furumoto M, Nakanishi I (1983) Source times and scaling relations of large earthquakes. *J Geophys Res* 88:2191–2198
- Hardebeck JL, Hauksson E (2001) The crustal stress field in southern California and its implications for fault mechanics. *J Geophys Res* 106(21):859–882
- Hiramatsu Y, Furumoto M, Nishigami K, Ohmi S (2002) Initial rupture process of microearthquakes recorded by high sampling borehole seismographs at the Nojima fault, central Japan. *Phys Earth Planet Inter* 132:269–279
- Hirata M (2003) The initial rupture process of the 2000 Western Tottori Earthquake. Master Thesis, Kyoto University
- Horikawa H (2006) Rupture process of the 2005 West Off Fukuoka Prefecture, Japan, earthquake. *Earth Planet Space* 58:87–92
- Ide S, Beroza GC, Prejean SG, Ellsworth WL (2003) Apparent break in earthquake scaling due to path and site effects on

- deep borehole recordings. *J Geophys Res* 108(B5):2271; doi:10.1029/2001JB001617
30. Iio Y (1992) Slow initial phase of the P-wave velocity pulse generated by microearthquakes. *Geophys Res Lett* 19:477–480
  31. Iio Y (1995) Observation of the slow initial phase generated by microearthquakes: Implications for earthquake nucleation and propagation. *J Geophys Res* 100:15333–15349
  32. Iio Y, Ohmi S, Ikeda R, Yamamoto E, Ito H, Sato H, Kuwahara Y, Ohminato T, Shibazaki B, Ando M (1999) Slow initial phase generated by microearthquakes occurred in the Western Nagano prefecture, Japan -the source effect-. *Geophys Res Lett* 26(13):1969–1972
  33. Iio Y, Kobayashi Y, Tada T (2002) Large earthquakes initiate by the acceleration of slips on the downward extensions of seismogenic faults, Earth Planet. *Sci Lett* 202:337–343
  34. Iio Y, Horiuchi S, Ohmi S, Ito H, Kuwahara Y, Yamamoto E, Omura K, Miura K, Shibazaki B, Sato H (2006) Slow initial phase of microearthquakes. Program and abstracts of 2006 fall meeting of the Seismological Society of Japan, A48 (in Japanese)
  35. Ishihara Y, Fukao Y, Yamada I, Aoki H (1992) Rising slope of moment rate functions: the 1989 earthquakes off east coast of Honshu. *Geophys Res Lett* 19:873–876
  36. Ito S (2003) Study for the initial rupture process of microearthquakes in western Nagano, central Japan, estimated from seismograms recorded in three boreholes. Ph D Thesis, Tohoku University (in Japanese)
  37. Ito S, Ito H, Horiuchi S, Iio Y (2004) Local attenuation in western Nagano, central Japan, estimated from seismograms recorded in three boreholes. *Geophys Res Lett* 31:L20604; doi:10.1029/2004GL020745
  38. Ito Y, Obara K, Takeda T, Shiomi K, Matsumoto T, Sekiguchi S, Hori S (2006) Initial-rupture fault, main-shock fault, and after-shock faults: Fault geometry and bends inferred from centroid moment tensor inversion of the 2005 West Off Fukuoka Prefecture earthquake. *Earth Planet Space* 58:69–74
  39. Iwata T, Sekiguchi H (2002) Source process and near-source ground motion during the 2000 Tottori-ken Seibu earthquake ( $M_w$  6.8). Reports on Assessments of Seismic local-site effects at plural test sites. MEXT, pp 231–241
  40. Kanamori H, Anderson DL (1975) Theoretical bases for some empirical relations in seismology. *Bull Seism Soc Am* 65:1073–1095
  41. Kanamori H (1996) Initiation process of earthquakes and its implications for seismic hazard reduction strategy. *Proc Natl Acad Sci* 93:3726–3731
  42. Kawanishi R, Iio Y, Yukutake Y, Katao H, Shibutani T (2006) Estimate of the stress field in the region of the 2000 Western Tottori earthquake. Program and abstracts of 2006 fall meeting of the Seismological Society of Japan, P099 (in Japanese)
  43. Kilb D, Gombert J (1999) The initial subevent of the 1994 Northridge, California, Earthquake – is earthquake size predictable? *J Seismol* 3:409–420
  44. Lay T, Kanamori H, Ruff L (1982) The asperity model and the nature of large subduction zone earthquakes. *Earthq Predict Res* 1:3–71
  45. Mandelbrot BB (1982) *The fractal geometry of nature*. W.H. Freeman, New York
  46. Miura K, Iio Y, Yukutake Y, Takai K, Horiuchi S (2005) The feature of initial motion for waveforms of microearthquakes in Western Nagano, Japan. Program and abstracts of 2005 fall meeting of the Seismological Society of Japan, P103. (in Japanese)
  47. Miyake H, Iwata T, Irikura K (2003) Source characterization for broadband ground motion simulation: Kinematic heterogeneous source model and strong motion generation area. *Bull Seism Soc Am* 93:2531–2545
  48. Mori J, Kanamori H (1996) Initial rupture of earthquake in the 1995 Ridgecrest, California sequence. *Geophys Res Lett* 23:2437–2440
  49. Nakamura Y (1988) *Proc World Conference on Earthquake Engineering*, VII, 6763
  50. Nakatani M, Kaneshima S, Fukao Y (2000) Size-dependent microearthquake initiation inferred from high-gain and low-noise observations at Nikko district, Japan. *J Geophys Res* 105(B12):28095–28110; doi:10.1029/2000JB900255
  51. Ohmi S, Watanabe K, Shibutani T, Hirano N, Nakao S (2002) The 2000 Western Tottori Earthquake—Seismic activity revealed by the regional seismic networks. *Earth Planet Space* 54:819–830
  52. Ohnaka M, Kuwahara Y, Yamamoto K, Hirasawa T (1986) Dynamic breakdown processes and the generating mechanism for high-frequency elastic radiation during stick-slip instabilities. In: Das S, Boatwright J, Scholz CH, AGU (eds) *Earthquake source mechanics*. Geophysical Monograph, vol 37. Maurice Ewing Series, vol 6. American Geophysical Union, Washington DC, pp 13–24
  53. Ohnaka M, Kuwahara Y (1990) Characteristic features of local breakdown near a crack-tip in the transition zone from nucleation to unstable rupture during stick-slip shear failure. *Tectonophysics* 175:197–220
  54. Ohnaka M, Shen L (1999) Scaling of the shear rupture process from nucleation to dynamic propagation: Implications of geometric irregularity of the rupture surfaces. *J Geophys Res* 104:817–844
  55. Ohnaka M (2000) A physical scaling relation between the size of an earthquake and its nucleation zone size. *Pure Appl Geophys* 157:2259–2282
  56. Okubo PG, Dieterich JH (1984) Effects of physical fault properties on frictional instabilities produced on a simulated faults. *J Geophys Res* 89:5817–5827
  57. Olson EL, Allen RM (2006) Is earthquake rupture deterministic? *Nature* 442:E5–E6; doi:10.1038/nature04963
  58. Rydelek P, Horiuchi S (2006) Is earthquake rupture deterministic? (Reply). *Nature* 442:E6; doi:10.1038/nature04964
  59. Sato T (1994) Seismic radiation from circular cracks growing at variable rupture velocity. *Bull Seismol Soc Am* 84:1199–1215
  60. Sato T, Hirasawa T (1973) Body wave spectra from propagating shear cracks. *J Phys Earth* 21:415–431
  61. Sato T, Kanamori H (1999) Beginning of earthquakes modeled with the Griffith's fracture criterion. *Bull Seismol Soc Am* 89:80–93
  62. Sato K, Mori J (2006) Scaling relationship of initiations for moderate to large earthquakes. *J Geophys Res* 111:B05306; doi:10.1029/2005JB003613
  63. Sato K, Mori J (2006) Relation between rupture complexity and earthquake size for two shallow earthquake sequences in Japan. *J Geophys Res* 10.1029/2005JB003613
  64. Shibazaki B, Matsu'ura M (1992) Spontaneous processes for nucleation, dynamic propagation, and stop of earthquake rupture. *Geophys Res Lett* 19:1189–1192
  65. Shibazaki B, Matsu'ura M (1995) Foreshocks and pre-events associated with the nucleation of large earthquakes. *Geophys Res Lett* 22(10):1305–1308; doi:10.1029/95GL01196

66. Shibazaki B, Matsu'ura M (1998) Transition process from nucleation to high-speed rupture propagation: Scaling from stick-slip experiments to natural earthquakes. *Geophys J Int* 132:14–30
67. Shibazaki B, Yoshida Y, Nakamura M, Nakamura M, Katao H (2002) Rupture nucleations in the 1995 Hyogo-ken Nanbu earthquake and its large aftershocks. *Geophys J Int* 149:572–588
68. Spudich P, Cranswick E (1984) Direct observation of rupture propagation during the 1979 Imperial Valley earthquake using a short-baseline accelerometer array. *Bull Seismol Soc Am* 74:2083–2114
69. Takenaka H, Nakamura T, Yamamoto Y, Toyokuni G, Kawase H (2006) Precise location of the fault plane and the onset of the main rupture of the 2005 West Off Fukuoka Prefecture earthquake. *Earth Planets Space* 58:75–80
70. Uchide T, Ide S (2007) Development of multiscale slip inversion method and its application to the 2004 Mid-Niigata Prefecture earthquake. *J Geophys Res* doi:10.1029/2006JB004528
71. Uehira K, Yamada T, Shinohara M, Nakahigashi K, Miyamachi H, Iio Y, Okada T, Takahashi H, Matsuwo N, Uchida K, Kanazawa T, Shimizu H (2006) Precise aftershock distribution of the 2005 West Off Fukuoka Prefecture Earthquake ( $M_j = 7.0$ ) using a dense onshore and offshore seismic network. *Earth Planet Space* 58:1605–1610
72. Umeda Y (1990) High-amplitude seismic waves radiated from the bright spot of an earthquake. *Tectonophysics* 175:81–92
73. Umeda Y (1992) The bright spot of an earthquake. *Tectonophysics* 211:13–22
74. Umeda Y, Yamashita T, Tada T, Kame N (1996) Possible mechanisms of dynamic nucleation and arresting of shallow earthquake faulting. *Tectonophysics* 261:179–192
75. Yamaguchi S, H Kawakata, T Adachi, Y Umeda (2007) Features of initial process of rupture for the 2005 West off Fukuoka Prefecture Earthquake. *Zisin Ser* 2:241–252 (in Japanese)
76. Venkataraman A, Beroza GC, Ide S, Imanishi K, Ito H, Iio Y (2006) Measurements of spectral similarity for microearthquakes in western Nagano, Japan. *J Geophys Res* 111:B03303; doi:10.1029/2005JB003834
77. Wu Y, Kanamori H, Allen R, Hauksson E (2007) Determination of earthquake early warning parameters,  $\tau_c$  and  $P_d$ , for southern California. *Geophys J Int* (OnlineEarly Articles) doi:10.1111/j.1365-246X.2007.03430.x
78. Wyss M, Brune J (1967) The Alaska earthquake of 28 March 1964—a complex multiple rupture. *Bull Seismol Soc Am* 57:1017–1023



# Earthquake Occurrence and Mechanisms, Stochastic Models for

DAVID VERE-JONES

Statistical Research Associates and Victoria University,  
Wellington, New Zealand

## Article Outline

Glossary  
 Definition of the Subject  
 Introduction  
 Historical Overview  
 Stochastic Models for Earthquake Mechanisms  
 Models for Paleoseismological  
 and Historical Earthquakes  
 Point Process Models for Regional Catalogues  
 Stochastic Models with Precursors  
 Further Topics  
 Future Directions  
 Acknowledgments  
 Bibliography

## Glossary

**Stochastic** occurring by chance;

**Stochastic process** physical or other process evolving in time governed in part by chance.

**Earthquake mechanism** physical processes causing the occurrence of an earthquake.

**Independent events** events not affecting each other's probability of occurrence.

**Branching process** process of ancestors and offspring, as in the model of nuclear fission.

**Point process** stochastic process of point-events in time or space.

**Probability forecast** prediction of the probability distribution of the time and other features of some future event, as distinct from a forecast for the time (etc.) of the event itself.

**Model test** a statistical test for the extent to which a stochastic model is supported by the relevant data.

**Precursory signal** observed quantity which affects the occurrence probability of a future event (earthquake).

## Definition of the Subject

Stochastic models for earthquake mechanism and occurrence combine a model for the physical processes generating the observable data (catalog data) with a model for the errors, or uncertainties, in our ability to predict those observables. Such models are essential to properly quantify

the uncertainties in the model, and to develop probability forecasts. They also help to isolate those features of earthquake mechanism and occurrence which can be attributed to mass action effects of a statistical mechanical character. We do not consider in this paper applications of the models to earthquake engineering and insurance.

## Introduction

The complexity of earthquake phenomena, the difficulty of understanding and monitoring the processes involved in their occurrence, and the consequent difficulty of accurately predicting them, are now widely accepted points of view. What are stochastic models, and what role do they play in aiding our understanding of such phenomena?

The present article is an attempt to address these questions. We start from the beginnings, the distinction between stochastic and deterministic models, and the first attempts to model earthquake phenomena in stochastic or statistical terms. We then follow through with a systematic account of some of the main classes of stochastic models that are currently in use, discussing in turn earthquake mechanisms, historical earthquakes, regional catalogs, descriptive patterns, and earthquake precursors.

The focus throughout is on the stochastic modeling aspects, rather than on statistical procedures or associated algorithms. As a result we have given only brief mention to pattern-recognition techniques, or to descriptive procedures such as the estimation of fractal dimensions or of second order properties, which do not lead to fully defined models. Again, although a primary use of stochastic models is in developing probability forecasts, we have limited ourselves to the briefest account of how such forecasts can be produced and assessed. Nor do we directly consider applications to engineering and insurance problems.

The fundamental difference between a physical model and a stochastic model, in broad terms, is that while the physical model attempts to fully describe and predict the process under study, the stochastic model treats some aspects of the physical process as out of range of exact modeling, at least for the time being, and replaces it by some unpredictable and hence random process. The resulting stochastic model should reproduce those aspects of the physical phenomenon which are important and accessible to measurement, but may relegate the rest to dice-tossing or one of its more contemporary avatars such as Brownian motion or the Poisson process.

Across their many different fields of application, two broad roles for stochastic models may be distinguished. The first is epitomized by statistical mechanics. Here the stochastic model plays an integral role in understanding

the physical processes themselves. The macroscopic phenomena that we are able to observe directly – temperature, pressure and the like – are shown to be a consequence, not of the details of the collision processes at the microscopic level, but of their mass interactions, which are governed largely by laws of an essentially statistical character such as the law of large numbers or the central limit theorem. For predicting the macroscopic behavior, it is not necessary to know the details of the complex interactions between individual molecules; it is sufficient to replace them by a simple random process that nonetheless preserves the crucial physical aspects such as mean velocities and angular distributions.

Within seismology such a role is implicit when the fracture processes within the earth's crust are compared to 'frozen turbulence', or in applications of branching process or percolation theory to explain energy distributions, or in discussions of the fracture strength of materials as functions of the density and size distribution of microcracks, or in the use of cellular automata and similar models to explain the appearance of long-range correlations and power-law distributions in the approach to criticality of certain types of complex systems.

In the other, by far more common, type of application, the stochastic model is used as a basis for planning and prediction. In such situations it is vital to know, not just a forecast value, but also something about the reliability of that value. It is also vitally important that the models can be fully fitted to the observable data. Most branches of applied statistics have evolved in response to such requirements. Within seismology, applied models of this type are needed in discussions of earthquake risk for insurance or building codes, in many parts of engineering seismology, and in the development of decision rules for earthquake response and emergency planning. Probability forecasts of any kind, including all forecasts with some associated estimate of precision, necessarily rely on stochastic models of this kind.

Many decades ago, the famous geophysicist and seismologist Sir Harold Jeffreys, who is also regarded as a pioneer in inferential statistics, argued that, to be worthy of its name, every physical theory should contain within itself the means not only of predicting the relevant quantities, but also of predicting their uncertainties [45]. In our terminology, he was arguing that every physical theory should be based on a stochastic model. In the classical studies of physics and astronomy, the uncertainties in the model are assumed to be due to nothing deeper than observational errors. In a subject such as seismology, however, the uncertainties are much more fundamental.

While general patterns of earthquake behavior may

be predicted from physical theories, the predictions do not extend to the times and locations of individual earthquakes. Moreover, the available observational data are rarely more than indirectly relevant to the physical processes controlling the details of earthquake occurrence, as these usually take place many kilometers beneath the surface of the earth, and out of range of direct observation.

Stochastic models of earthquake occurrence that can be used for earthquake prediction must somehow marry the limited physical theory to the limited available data. Attempts to grapple with this central problem have intensified in recent years. They form one factor in the emergence of 'Statistical Seismology' as a new sub-discipline. Another, perhaps dominating, factor, is the enormous improvement in both the quantity and quality of the data that are available, whether from earthquake catalogs, from GPS measurements of ground deformation, or, less commonly, from data on auxiliary quantities such as well levels, electrical signals, ionospheric depression and others thought to have a potential bearing on earthquake occurrence. The high quality data demand a comparable quality in the statistical modeling and analysis.

## Historical Overview

The forerunner of any serious statistical modeling is the availability of reliable and relevant data. For models of earthquake occurrence this means the availability of good quality earthquake catalogs. Broadly speaking, such catalogs had to wait, not only until around the turn of the 20th century, when the first instrumental records became available, but until the development of modern instrumentation and the extensive station networks which came into being after the second World War. Before then, the lack of any consistent measure of the size of an earthquake, and the general unevenness of network coverage, made the records of limited value for statistical purposes.

A turning point was the appearance of the first edition of the classic text [28] by Gutenberg and Richter (1949). For the first time it gave a comprehensive overview of the major features of the seismicity of the earth, and of the key empirical relations governing earthquake occurrence. From that time onwards, the way was open for serious statistical analysis, although recent data is far more comprehensive and detailed. Modern instrumental catalogs, prepared from digital records telemetered to a local center from a dense network of stations, may contain hundreds of thousands of events down to very small magnitudes. Typically such catalogs list for each event the origin time (initiation of rupture), epicenter (latitude

and longitude of place of first motion), depth, magnitude or seismic moment (alternative measures of earthquake size), and often other parameters relating to the fault mechanism (orientation of the fault and direction of first motion).

The availability of these high-quality catalogs, alongside the increasing availability of data from GPS measurements and other earthquake-related phenomena, is a key reason for the recent upsurge of interest and research in statistical seismology. The broad aims of this emerging field may be described as finding statistical models to describe and make use of such data, and to marry it with the existing physical theory.

Even preinstrumental catalogs inspired the investigation of two statistical issues at least: does the occurrence of major earthquakes exhibit some form of periodicity in time? do the numbers of events in time intervals of fixed length follow a Poisson distribution? We comment briefly on these two questions before looking at statistical models more generally.

### Periodicity of Earthquakes

Periodicity of earthquakes, in some more or less regular sense, was the earliest issue to be investigated, and inspired many early studies, including one of Schuster's classic papers on the periodogram [112]. Until the 1930s, however, neither the data nor the statistical techniques were sufficiently developed to allow the question to be properly resolved. On the statistical side, for instance, the periodogram was a new concept, and statistical tests based on it were still in development. Schuster's paper, applied to a special case, contains within itself all the basic elements of point process spectral theory, starting from the finite Fourier transform, calculating the equivalent of rough significance levels using Rayleigh's random flights, and briefly treating the problems caused by binning the data and by clustering. Jeffreys [44] was one of the first to use a modern statistical approach to tackle the question, while Davison [21] reviewed many of the earlier studies and came to the conclusion that most of those studies were inconclusive.

The topic remains controversial, although it is now clear that no obvious periodicities exist. The most important current contenders for small-scale periodicities are in relation to earth tides. It is suggested that the small fluctuations in crustal stress due to the relative movements of the moon and sun around the earth may be large enough to trigger earthquake activity under favorable conditions, for example in regions already under high stress. A careful recent study of lunar tides on microearthquakes, with fur-

ther references, is given in [40]. The possibility of using the response of small-scale seismicity to lunar tides as a possible indicator of regions in some near-critical state, and hence as a precursor for larger events, has been suggested in [139]; a statistical analysis is given in [142].

The possibility of long-term periodicities, of the order of decades or centuries, is unclear because of the shortage of data; substantial fluctuations certainly exist. The problem is still full of potential traps. In testing for periodic effects, for example, it is essential to take into account earthquake clustering, and whether or not the period being tested for is preassigned (as for the lunar cycle) or suggested by the data. Both of these issues are illustrated in the discussion in [134] of Kawasumi's historical data for large earthquakes in the Kanto region of Japan.

### The Poisson Distribution and Process

The distribution

$$p_n = (\mu^n/n!)e^{-\mu}, \quad \mu > 0, \quad n \geq 0,$$

was introduced by Poisson as an approximation to the binomial distribution when the number of trials  $N$  becomes very large but the probability  $p$  of success becomes very small, the two balancing in such a way in such a way that the expected number  $\mu = Np$  of successes remains moderate in size.

Earthquakes were included among the examples studied by von Bortkiewicz [136] in his 1898 compilation of phenomena to which he could apply 'the law of small numbers', the name he gave to the Poisson approximation to the binomial. The question was studied in greater depth by later writers, including Gutenberg and Richter [28], and several important qualifications were noted. In the first instance, the disturbing effect of aftershocks was pointed out, and so the Poisson distribution was supposed to apply just to main shocks. Then other disturbing effects, such as trends and longer-term fluctuations in activity, were noted. In fact almost no catalog fits the Poisson description exactly, and for research purposes its role as a base-line model for 'standard seismicity' has been replaced by the ETAS model (see Sect. "The ETAS Model"), which provides a much better approximation to the clustering properties of smaller earthquakes.

Nevertheless the simple Poisson form is still the principal basis for determining earthquake risk and for earthquake insurance practices. Underlying its continued relevance is the same idea underlying Poisson's original approximation to the binomial: when the data under examination consists of rare 'successes' from many different

and essentially unrelated sources, the Poisson distribution generally emerges as a good approximation.

It is necessary to distinguish between the *Poisson distribution* and the *Poisson process*. The Poisson process refers to an evolutionary model for the occurrence of events in time or space or both. Its principal characteristics, at least in the stationary case, are

1. The number of events within any bounded region (interval in time; area or volume in space) follows a Poisson distribution with parameter  $\mu$  proportional to the size (length, area etc) of the region selected for study;
2. The numbers of events in disjoint regions are independent random variables.

The second condition embodies the famous ‘lack of memory’ property of the Poisson process: the temporal version asserts that the occurrence of one or more events before a certain time has no effect on the occurrence probabilities of subsequent events. It dictates the exponential form of the distribution of the time interval between events, and under simple conditions even dictates the form of the Poisson distribution itself; see, for example, the discussion in Chap. 2 of [20].

### The Empirical Laws of Seismology

The advent of more complete and reliable catalogs saw the recognition of a number of statistical regularities in the occurrence of earthquakes. Two of these are central features of seismicity studies today.

**Omori’s Law** Already by the end of the 19th century, the Japanese pioneer seismologist Omori had made detailed studies of some large Japanese aftershock sequences [92], and suggested that the frequency of aftershock occurrence, say  $\lambda(\tau)$ , decayed approximately hyperbolically with the time  $\tau$  after the main event:

$$\lambda(\tau) \approx A/\tau$$

where  $A$  is a constant characteristic of the particular mainshock and associated with its size. His own and subsequent studies suggested the need for refinements, and the most widely accepted form today is the the Omori–Utsu formula

$$\lambda(\tau) = A/(c + \tau)^p \quad (1)$$

where the parameters  $A$ ,  $c$ ,  $p$  are again peculiar to the individual aftershock sequence,  $c$  is generally small (of the order of seconds to days) and  $p$  is close to 1. A detailed study of the history and other issues associated with the Omori law over the 100 years 1894–1994 is given in [124].

The simplest stochastic model for aftershocks is that suggested by Jeffreys in [44], namely an inhomogeneous Poisson process in which  $\lambda(\tau)$  is interpreted as the current value of the time-varying Poisson intensity; the independence property (2) of the Poisson process of Sect. “**The Poisson Distribution and Process**” is retained, but the mean parameter for the number of events in  $(s, t)$  is now  $\mu = \int_s^t \lambda(\tau) d\tau$ .

**The Gutenberg–Richter (GR) Law** The law was formulated after the definition of earthquake magnitude gave an objective method of quantifying the size of an earthquake. It is a basic component of [28], although a similar relationship, based on the more qualitative maximum intensity concept, had been formulated somewhat earlier by Ishimoto and Iida for Japanese earthquakes [39].

The GR law provides a summary of the magnitude data in a catalog of earthquakes with magnitudes complete above a certain threshold, say  $M_0$ . It is commonly written in the form

$$\begin{aligned} \{\text{Number of events above magnitude } M\} \\ \approx 10^{a-b(M-M_0)} \quad (2) \end{aligned}$$

or equivalently

$$\begin{aligned} \{\text{Proportion of events above Magnitude } M\} \\ = 10^{-b(M-M_0)}. \quad (3) \end{aligned}$$

It is a pity in our view that the former rather than the latter of these two forms has become traditional. The danger then is that  $a$  becomes treated as a separate parameter instead of as a normalizing constant,  $a = \log_{10} N$ , where  $N$  is the total number of events under consideration. One reason for this tradition may have been the common (and incorrect) use of ordinary least squares methods to compute the line of best fit from a graph of the binned numbers. Such an approach will certainly produce a slope as one of the parameters, but the estimate is unstable and distorts the interpretation unless it is especially modified to fit the distribution function context.

The second form makes it clear that what we are looking at is an empirical probability distribution, and that the right hand side could equally and more appropriately be written in the form

$$10^{-b(M-M_0)} = e^{-\beta(M-M_0)}$$

where  $\beta = b \log_e 10 \approx 2.3b$ . Then it is clear that the GR law asserts that, under suitable conditions, the empirical distribution of magnitudes is approximately exponential.

In principle,  $a$  could be regarded as a parameter in an extended model for the space-time-magnitude distribution of events in a given space-time window, but such an interpretation is rarely given.

How valid the exponential distribution remains when examined in greater detail, and whether, and if so by what, it should be replaced for general modeling purposes, is still a subject of debate. The main reservation relates to the possibility of extremely large events, which is physically unreasonable and can lead to misleading conclusions if used in simulation studies of long-term behavior.

Of the many alternatives offered, which include truncated and multi-parameter versions (see [123] for a listing and software), perhaps the most plausible is the ‘tapered Pareto distribution’, or ‘Kagan distribution’ (e.g. [48, 135]), with distribution function written out in terms of seismic moments or energies as (4) below. This distribution arises in branching and similar models for crack propagation (see [123]), and is derived from maximum-entropy considerations in [73]; a further derivation from a critical phase transition in a finite elastic solid is given in [26].

The use of magnitude itself as a basic variable is also open to question. It is not a uniquely or tightly defined quantity. In terms of quantities such as energies or seismic moments with a more direct physical interpretation, the exponential distribution for magnitudes becomes a Pareto (inverse power-law) distribution,

$$\Pr\{E > x\} = (x/x_0)^{-\alpha}, \quad (x > x_0)$$

through a transformation of the form

$$\log_{10}(E) \approx 1.5M + \text{const } n,$$

where  $E$  is the energy. This illustrates the fact that the magnitude scale is essentially a decibel scale, ultimately a consequence of its initial definition in terms of the logarithm of the maximum amplitude of the trace on a seismograph. The tapered Pareto form mentioned earlier is

$$\Pr\{X > x\} \approx Cx^{-\alpha}e^{-\gamma x}, \quad (x \geq x_0), \quad (4)$$

where  $C$  is a normalizing constant, or the variant with a similar form for the density.

**Båth’s Law** The so-called Båth’s Law asserts, loosely, that in an aftershock sequence, the difference between the magnitude of the mainshock and that of the largest aftershock is around 1.2 magnitude units.

Although noted by Båth in 1965, and even earlier by Utsu, this regularity has never enjoyed quite the same status as the other two laws. The question, still an active topic

of debate, is whether it represents a physical phenomenon in its own right, or is merely a consequence of the more general properties of earthquake clustering. It was suggested in [125] (see also the reviews and more extensive studies in [17,63]) that it might be simply a consequence of the statistical properties of the largest and second largest in a sequence of events following the G-R law. More recently, it has been linked to the ‘productivity function’ of earthquake clustering: the expected number of aftershocks increases typically as an exponential function  $Ke^{\alpha M}$  of the magnitude of the main shock, with Båth’s Law resulting when the exponent  $\alpha$  equals the exponent  $\beta$  in the GR law (see [25]). This suggestion is supported by the appearance of a Båth’s law phenomenon in the ETAS model, where there is certainly no explicit model feature relating to Båth’s law, but there is an exponential productivity function [38].

## Stochastic Models for Earthquake Mechanisms

### General Considerations

The earliest model for earthquake mechanism is Reid’s elastic rebound model [98]. It was inspired by studies of large-scale earthquakes, in particular the famous San Francisco earthquake of 1906. The upper part of the crust is deformed elastically by large scale tectonic motions, then ruptures and rebounds when its breaking strength is reached, resulting in an earthquake.

The many attempts to marry this physical picture with simple stochastic ideas lead typically to models based broadly on the renewal process, and will be picked up in the discussion in Sect. “[Background and Data](#)”.

In this section we look rather at models which describe the behavior at the microscopic level, the evolution of the fracture itself, and can be used to explain the basic empirical laws, among other features.

There are strong links with theories on the strength of materials, starting from the classic studies of Griffiths (e.g. [27]) on crack extension in brittle materials, and the role of microfractures in controlling the fracture strength of glass. Griffiths’ crack theory is basic to models of fracture in brittle materials, whether at the scale of rock fracture in laboratory specimens or fault propagation in the earth’s crust (see e.g. [42,82,109,110]).

Griffiths’ ideas were later developed by Weibull [136] into a model explaining the variations in strength of otherwise similar specimens of rock and many other substances. Weibull supposed that the underlying cause was the random distribution of microfracture lengths in the specimen, and used an argument based on the distribution of the length of the largest such microfracture to deduce a form

for the distribution of strengths. Indeed it is from these studies that the ‘Weibull distribution’ takes its name.

The branching process, percolation, and cellular automata interpretations of the earthquake process start from similar general premises. The underlying idea is that, instead of progressing smoothly, as might a fault or fracture in a homogeneous elastic medium, the progress of the fault in a medium containing many weaknesses is controlled by the essentially random locations of these weaknesses. The various models which have been proposed differ mainly in the assumptions governing these random locations.

In Otsuka’s original ‘go-game’ model [93], points were laid down on a lattice in much the same way as in the game of ‘Go’, but at random, using a simulation technique, with the interpretation that the enclosed pieces determined a rupture area. In [108] this was idealized into a model linking weaknesses located on a Bethe lattice, where every node has one input link and the same fixed number of outward links, and each node may or may not be a point of weakness.

Otsuka’s model has both branching process and percolation model interpretations, with a considerable literature surrounding extensions of both. There also links to other, apparently more deterministic, approaches to the generation of the empirical laws, for example through block slider models or in the general class of complex systems. Although the models differ in approach and detail, the size distributions and the like often turn out to be very similar to those derived from the branching models. As in statistical mechanics, the properties have their origin in the mass interactions of many small components, and are relatively insensitive to the details at the microscopic level. The simpler statistical models, such as the branching models, allow these distributions to be explored directly by analytical means. In complex system theory the aim is rather to show how similar results arise from approximating the deterministic equations governing large families of interacting bodies.

### Branching Models

The conceptual framework here is that the crack initiates from an initial weakness (dislocation or microfracture) and spreads to one or more others, or terminates, the ‘others’ being interpreted as ‘offspring’ and the initial weakness as the ‘ancestor’. Each ‘other’ then acts as an ancestor in its own right, and the process continues until either all branches have died out (subcritical and critical cases) or the process explodes (supercritical state). The behavior is controlled by a ‘criticality parameter’  $\rho$ , effectively the

mean number of offspring per ancestor. The subcritical, critical, and supercritical cases correspond respectively to  $\rho < 1$ ,  $\rho = 1$  and  $\rho > 1$ .

This model was developed in general form in [123], following [109] and the earlier work on the ‘go-game’ model in Japan. Related ideas occur in many places papers by Kagan and Kagan and Knopoff, see especially the extended branching model described in [54].

The distribution of the size or energy release of the rupture is then obtained by counting the total number of offspring before the process dies out (critical or subcritical cases). The remarkable feature here is that even when the individual offspring distributions are very regular, the total size distribution approaches a power-law (Pareto) form whose basic parameters are independent of the details of the offspring distribution. In the limiting critical case, the power-law distribution for sizes has  $\Pr\{N > n\} \sim Cn^{-1/2}$ , corresponding roughly, assuming equal energies/event on average, to a G-R law with  $b \approx 0.75$ .

When the process is just subcritical the Pareto distribution becomes a tapered Pareto distribution, with power-law behavior for moderate to large events, and an exponential tail-off at high magnitudes which cuts in at a point determined by the distance from criticality,  $\delta = 1 - \rho$ . Again the behavior is otherwise largely independent of the details of the offspring distribution.

Many further developments and ramifications of this underlying model have been proposed. One of the deepest is the simulation model for earthquakes developed in [54], starting from the scale of dislocations or other defects in the rock fabric, and incorporating temporal, directional, and distance factors into the model evolution to develop an impressive array of properties akin to those of real earthquakes. The model still awaits a full analytical treatment.

It is also remarkable that a branching model underlies one of the most successful models for earthquake occurrence at the regional level, namely the ETAS model described in Sect. “[The ETAS Model](#)”. The fact that the same mechanism seems implicated at both levels lends plausibility to Kagan’s conjecture that the physical process is one and the same at all scales, and that our attempts to decompose it into elements at the fracture formation and inter-fracture stages are more a result of our perceptions and measuring instruments than they are of the underlying physical processes.

### Percolation Models

The classical percolation model starts from a two- or three-dimensional lattice, the sites (or alternatively the

bonds between lattice points) being randomly and independently labeled ‘open’ or ‘closed’ with a fixed probability  $p$  and its complement  $1 - p$ . A crack initiated at an open site links up all contiguous open sites until it can spread no further. In both cases, a critical regime, characterized by a critical value of the probability  $p$ , marks the transition between subcritical (small finite events only) and supercritical (infinite or explosive events) regimes. As with the branching models, it is assumed that the crust is generally in or just below the critical state.

An underlying difficulty is that the available observational data are insufficient to provide any easy control over the best interpretation. As with the branching process model again, the percolation models lead to forms of the G-R law, and with additional features can often be extended to cover aftershock phenomena. [15], and [65,66,67] are among the many papers which discuss and develop these ideas. [5] highlights some of the difficulties of interpretation.

Percolation processes are extensively used in statistical physics to model phase transitions, and their appearance here invites an interpretation of fracture as a phenomenon analogous in some ways to a phase transition. Ideas from the phase transition context that have been transferred to both earthquakes and fracture mechanics include especially features characteristic of the approach to the critical conditions required for the occurrence of a phase transition: the development of long-range correlations, the appearance of power-law or fractal distributions, and approximate self-similarity. Many authors have sought to develop these analogies, often using analogue or simulation models, and attempted to use the appearance of different interaction ranges to identify the approach to near-critical stress conditions in the crust. See [9,10,26,121], as well as the papers cited above, for further references and discussion of such ideas.

### Cellular Automata and Self-Organizing Criticality

A third type of model with a similar general role is the cellular automaton, with the distinction that the application here is not to a single faulting or fracture episode, but to a whole network of interacting faults. The simple basic form, first applied to the earthquake context by Bak and Tang [3], again relates to a two-dimensional lattice model. With each point of the lattice is associated a certain integer stress or force, say  $Z_{ij}$  for points on a 2-dimensional lattice  $\{i, j\}$ . The external force (the ‘immigrants’ in this context) is manifested through the addition of unit force to a site chosen at random through the lattice or on its boundary. When the force exceeds a certain critical value

$Z_c$  on a given site, a ‘microfracture’ occurs, and single units of force are transferred to each of the four directly adjacent sites, while four units are subtracted from the force at the original site. Such transfers may overload one or more of the adjacent sites, which then in turn transfer units of stress to their neighbors (including possibly the initial site), and so on until the system is at rest. Then another unit is added in and a further redistribution of stress takes place. The whole episode is interpreted as an earthquake, and the total number of steps in the episode is taken as proportional to the energy of the earthquake.

The process as a whole is said to exhibit ‘self-organizing (or ‘self-organized’) criticality’. Even if the process is started from a situation where the forces are set to zero at all sites, they will gradually build up, first to the stage where small individual episodes take place, then, as more and more sites approach the the critical value of stress, the episodes become larger, until a stochastically stationary state is reached where the input of stress units is just balanced by the loss of stress units from points on the boundary of the region. So long as the region under consideration is sufficiently large, a process reaching the critical regime exhibits many of the features already indicated as characteristic of the approach to a phase-change: a G-R relation, long-range correlation effects, and (with some elaborations) an Omori-type phenomenon for aftershock sequences.

## Models for Paleoseismological and Historical Earthquakes

### Background and Data

We move now to models designed for use with data on earthquake occurrences. These generally belong to the second class of stochastic models referred to in the introduction. They should be able to be fitted to real catalog data; simulations from them should mimic real catalogs; and they should be useful in real applications, capable in particular of generating probability forecasts.

We have grouped the models into two main types, those developed to model large earthquakes on historical or even geological time scales, and those developed for use with modern instrumental catalogs, where smaller events are included. The main difference between the two types of model is their treatment of clustering; this is largely ignored in models of the first type, but plays a central role for models of the second type. Models of the first type are considered in the present section, models of the second type in Sect. “[Point Process Models for Regional Catalogs](#)”.

The distinction between the groups is associated with one of the longest and still unresolved debates over earth-

quake mechanism, namely the validity of the *characteristic earthquake hypothesis*. Crudely stated, this asserts that, for any given fault or fault segment, there exists an earthquake of approximately fixed magnitude, which is determined by the physical attributes of the fault, and repeats itself after time intervals of approximately fixed length. Since faults occur over a very wide range of sizes (themselves having a power law or Pareto distribution), this does not contradict, but rather suggests a different origin for, the GR distribution.

The empirical evidence for such a hypothesis is equivocal. Its main support comes from the paleoseismological studies on repeated events along a single fault (e. g. [113]), but the data from such studies is usually so limited that it is hard to accept the evidence as conclusive. Other supporting evidence, again observed sometimes but not always, is the occurrence of a hump, corresponding to repeating earthquakes with similar magnitudes, in the frequency-magnitude distribution for selected regions. For large regions (scale of major faults), this may occur around magnitudes 6–7, suggesting that the larger events from that region occur more regularly (with higher relative frequency) than would be expected from the GR model. One difficulty with the hypothesis is that several of the best-known sequences, such as the Parkfield earthquake sequence, ultimately deviate from the prescribed regularity. Studies of microearthquakes suggest that in some circumstances similar-sized small events may repeat themselves several times in almost identical locations.

Such results suggest that no simple, universal mode of behavior is likely to be found in earthquakes from particular fault structures. Indeed, recent studies by Ben-Zion and colleagues (see [7,8,9,10] and further references therein) have emphasized the possible role played by evolving heterogeneities and damage rheology in the occurrence patterns on a fault system, and have suggested that, according to their ages and past histories, some faults may exhibit characteristic earthquake behavior while others exhibit GR behavior and others again may alternate between the two.

For paleoseismic studies, each data point is extracted with effort from trenching along fault traces or similar exercises. Moreover, only the largest events leave traces identifiable over thousands of years or longer, while estimating magnitudes and other characteristics is at best informed guess-work. Even the dates, usually determined from some form of radio-carbon or other isotope-based method, can be subject to substantial errors.

Similarly in historical studies, such as Ambraseys' history of Persian earthquakes [1], only the largest events affecting a given region or territory are likely to appear sufficiently prominently in the historical records to allow the

size and epicenter of the earthquake to be estimated even roughly. Periods of civil war, famine, foreign invasion and bureaucratic neglect leave gaps and further uncertainties which are difficult if not impossible to resolve.

Despite such difficulties, these data provide the only records we have of seismic activity over periods stretching backwards in time beyond the last hundred years or so, and are worthy of the most serious attempts to collect and interpret.

We consider a sequence of three models, starting from the simple renewal model, then considering variants more closely linked to the elastic rebound model.

### Renewal Models

With paleological data in particular, attention is generally focused on major events along a single fault, where magnitudes are poorly constrained, and the stochastic elements are introduced primarily to describe, and if possible predict, the time intervals between events.

For a renewal process, magnitudes are neglected, and it is assumed that the successive intervals are independent, both of each other and of other processes, with a common distribution. The independence assumptions are questionable, but with no further information available, this is at least a reasonable starting model.

Let  $f(x)$  denote the density and  $F(x)$  the distribution function of the common interval distribution. If the observation record, over  $(0, T)$ , say, comprises an interval of length  $\ell_0$  to the first recorded event, then  $n$  complete intervals  $\ell_1, \ell_2, \dots, \ell_n$ , and finally an unfinished interval  $\ell_{n+1}$ , the likelihood is given by

$$L(\ell_0, \ell_1, \dots, \ell_n, \ell_{n+1}) = a(\ell_0) \left[ \prod_{i=1}^n f(\ell_i) \right] b(\ell_{n+1}), \quad (5)$$

where  $\ell_0$  and  $\ell_{n+1}$  are the incomplete intervals from the commencement of study to the first event, and from the last event to the end of the study, respectively,  $a(x) = [1 - F(x)]/\mu$ ,  $b(x) = 1 - F(x)$ ,  $\mu = \int_0^\infty uf(u)du$ .

The term  $a(x)$  at the beginning of the sequence is the appropriate form to use if the process can be supposed stationary, but nothing is known about events before the commencement of the observation period. The term  $b(x) = 1 - F(x)$  at the end of the sequence merely acknowledges the fact that the final interval has begun but not yet concluded.

The main uses of the model are in estimating long-term average or static hazards, in which case the mean of the interevent times plays the crucial role and the form of the distribution is largely irrelevant (so that a Poisson



approximation would generally be adequate). The other application is to estimating the residual time to the next event, which is governed by the extended hazard function

$$h(y|x) = f(x+y)/[1-F(x)] \quad (y \geq 0; x > 0), \quad (6)$$

giving the density of the distribution of the time  $y$  from the present to the next event, given that time  $x$  has elapsed since the last event.

Distributions commonly used in these situations include the Weibull, gamma, log-normal and Brownian first passage time (inverse normal). Recent work has tended to favor the last of these: see [79]. Two further recent studies which look carefully at the statistical issues, including those relating to errors in the occurrence times, are in [64] and [121]. In long period studies, care needs to be taken that consistent procedures have been used over the whole period, particularly in the determination of magnitude thresholds, and rules for the exclusion of aftershocks (which must be removed here since otherwise they would contradict the assumption of i.i.d. intervals).

### Time- and Slip-Predictable Models

The time-predictable model, introduced by Shimazaki and Nakata in [116], is a widely used alternative for major events along a given fault, when magnitudes as well as inter-occurrence times are available.

As in the elastic rebound model, it is supposed that stress along a particular fault builds up linearly until a critical value is reached, representing in some sense the strength of the fault. The size (magnitude) of the resulting event is not known beforehand, but is supposed to be selected randomly either from the standard G-R form or a variant suggested by the characteristic earthquake model.

Once it has occurred, the stress along the fault is instantaneously reduced by an amount determined by the magnitude of the event. Then stress build-up continues until the critical stress level is reached again. The time needed for this to occur is determined by the stress released by the previous event, and so is known, whence the 'time-predictable' title.

The major unknown in the model is the rate of stress build-up between events. If only observation times and magnitudes are available, this can be estimated, albeit crudely, by regressing the observed time intervals onto the magnitudes of the preceding events. If the magnitudes are determined up to a normal error term, with variance independent of the magnitude, this will result in a log-normal distribution for the length of the time interval following an event of given magnitude, and indeed this is commonly

used. Another approach would be to use geological data to provide an initial ('prior') distribution for the slip rate, and put the further analysis into a Bayesian framework.

A simple model used for predictive purposes in some of the papers by the Working Group on Californian Earthquakes (see [137] for example), can be represented as

$$\log T_i = A + M_i + \epsilon_i \quad (7)$$

where the  $\epsilon_i$  are independent, normally distributed errors with zero mean and constant variance, the  $M_i$  are the observed magnitudes of the events, and  $A = -\log V$  is the logarithm of the slip rate, estimated from geological and GPS studies. Given the time  $x$  since the last major event on the fault, the distribution of the remaining time  $y$  until the next event is governed by the extended hazard function of the log-normal distribution, as in the discussion (6) of the renewal model, but with the mean adjusted to take into account the extra information provided by the magnitude of the previous event.

An underlying but subtle logical difficulty with the model is that if applied on a long time basis, the assumption of i.i.d. lognormal errors leads to unbounded fluctuations in the accumulated sums

$$V \sum_1^n T_i - \sum_1^N S_i = V \sum_1^N S_i (\epsilon_i - 1),$$

where  $S_i$  is an estimate of the slip from an event with magnitude  $M_i$ . The cumulative sum on the right hand side can oscillate without bound, implying the unphysical possibility of indefinitely large fluctuations in the accumulated stress.

Shimazaki and Nakata also suggested a dual version, the *slip-predictable* model, characterized by a return after each event to a constant resting stress. The time at which the next event will occur is unknown, but given the time since the last event, the minimum expected size is determined by the stress accumulated since the previous event. [57] develops a more detailed version for use in earthquake engineering applications.

### The Stress-Release Model

The stress-release model is an attempt to address similar issues from within a stochastic point process framework (see for example Chap. 7 of [20]), incorporating both occurrence times and magnitudes. As in the previous case, it is assumed that the rate of stress build-up is constant (say  $\rho$ ), and that sizes of successive events are i.i.d. and independent of the stress level at the time of occurrence.

Most commonly they are assumed to follow the exponential form associated with the GR law, but this is not inherent in the model.

The crucial difference with the time-predictable model is that, instead of assuming that the strength of the crust is fixed, it is assumed to be variable with distribution function say  $\Phi(s)$  with density  $\phi(s)$ . The probability that the next earthquake occurs when the stress passes through  $s, s + ds$ , but not before, is then given by the hazard function  $\Psi(s) = \phi(s)/[1 - \Phi(s)]$ . This hazard function  $\Psi(s)$  determines the pattern of occurrence probabilities. Most commonly, it is taken to have an exponential form  $\Psi(s) = Ae^{\lambda s}$ , corresponding to the double exponential distribution function  $\Phi(S) = 1 - e^{-A[e^{\lambda s} - 1]}$  for the breaking strength itself. This has a well-marked mode at  $(-\log A)/\lambda$  if  $A$  is rather small.

In stochastic point process terms, the quantity  $\lambda^*(t) = \Psi[X(t)]$  can be interpreted as the *conditional intensity* of the model, meaning approximately the instantaneous occurrence rate, given the history of the process up to time  $t$ :

$$\lambda^*(t)dt \approx E[dN(t) | \mathcal{H}(t)] \approx \Pr\{dN(t) > 0 | \mathcal{H}(t)\}. \quad (8)$$

Roughly speaking, the process behaves locally like a Poisson process with instantaneous rate  $\lambda^*(t)$ , which in the stress-release model can be written more explicitly as

$$\lambda^*(t) = \Psi[X(t)] = \Psi \left[ X(0) + \rho t - \sum_1^{N(t)} S_n \right]. \quad (9)$$

This model has several useful features. First, the fact that a simple explicit form exists for the conditional intensity means that it can be readily incorporated into standard procedures for maximum likelihood estimation, simulation, and prediction (see again Chap. 7 of [20]). In particular, the likelihood ratio for a set of observed events  $(t_i, M_i)$  over the interval  $[0, T]$  can be written in the form

$$\log L/L_0 = \left[ \sum \log[\lambda^*(t_i)/\lambda] - \int_0^T [\lambda^*(u) - \lambda] du \right] + \sum_1^{N(T)} \log[g(M_i)/g_0(M_i)] \quad (10)$$

where  $\lambda$  is the rate of the background (null) model, assumed constant rate Poisson,  $g(x)$  and  $g_0(x)$  are the densities of the proposed and background magnitude distributions, and the magnitudes are assumed independent.

Second, as in earlier, related work by Knopoff [58], the current stress level, say  $X(t)$ , is Markovian, for the current value of  $X(t)$  determines the probability of the next jump

occurring, while the remaining components (size of jump, rate of build-up between jumps) are independent of the past history of the process. Hence the extensive knowledge of Markov processes can be brought to bear on the properties of  $X(t)$  (e. g. [11]).

A third point is that as the stress level increases, the rate of occurrence of new events will remain relatively high until a large enough event occurs to reduce the stress level to substantially lower values. The model therefore embodies a modest form of accelerated moment release [43].

The model assumes only a simple scalar concept for regional stress, much as in the early chapters of [83], and does not allow for stress interactions between regions. To address the latter point, the *coupled stress release model* was introduced by Shi Yaolin and students [68], to allow stress transfers between regions as well as simple stress drops. Further discussions and examples are in [4] and [71].

## Point Process Models for Regional Catalogues

### Data Consistency and Declustering

Regional catalogs, based on instrumental data from the last century or so, present a very different picture, but one with its own problems also. Of these, the two most important are the maintenance of consistency and the problem of clustering (or declustering).

It is characteristic of such catalogs that the networks supplying the data undergo many changes with the passing of the years. Although it is just these changes that have made possible the more serious statistical studies of recent years, they create their own problems in terms of lack of data consistency. Using such data for any form of long-term study requires continual vigilance over questions such as improvements and other changes in the individual network stations and their instruments, shifts in magnitude definitions or thresholds, changes in the routines used in determining epicenter locations, policy decisions over the events to be listed in the catalog, etc. Unless such factors are carefully listed and properly allowed for, they can easily lead to misinterpretation of statistical features observed in the data. As just one illustration, [18] gives some vivid examples of features of apparent physical interest which in fact have their origins in catalog artefacts induced by changes in magnitude registration.

An even more vexed question is whether, and if so how, to remove major clusters from (i. e. 'decluster') the catalog. Large aftershock sequences look simple to identify and remove, but the process is considerably more difficult than it might appear.

The possible justifications for doing so are two-fold. If it is believed that the large events are different in kind from the smaller events, then declustering is simply a procedure to isolate the events of primary importance. This assumption was once standard, and in any case the large events appear to be responsible for the major part of the large-scale tectonic motion. But with data on small events becoming ever more plentiful and increasingly reliable, their role is undergoing reassessment.

The second justification for removing aftershocks and other clusters is that they are a nuisance. They negate the assumptions of independence which lie at the basis of most standard statistical tests (e.g. for trends or periodic effects), they greatly complicate analysis and interpretation, and they require elaborate and difficult techniques to deal with explicitly.

Nevertheless, most statisticians, myself included, would tend to look askance at throwing away a substantial portion of the data on the basis of what are inevitably somewhat ad-hoc rules. Many procedures for removing aftershocks have been proposed, and the fact that none has gained general acceptance is evidence of this underlying problem. Moreover, while declustering removes the most obvious earthquake clusters, it rarely removes the clustering completely. The interpretation of results based on the remaining data remains equivocal, partly physical and partly man-induced.

For such reasons we do not discuss declustering in detail in the present article, but concentrate rather on procedures for modeling the data without removing the aftershocks.

A general caution in handling clustered data is not to presume that standard statistical procedures, especially tests, can be applied without modification. In general, the presence of clustering severely affects significance levels. For example, attention is drawn in [81] to the dangers of assessing the significance of precursory effects without properly allowing for clusters. A similar point occurs in assessing the significance of periodic effects, as was pointed out by Schuster in [112], as well as more recently in [129] and no doubt in other places.

To allow for the effects of clustering, and to examine the structuring features themselves, some form of explicit modeling is generally desirable. For example, one possible approach to highly clustered data is to remove as much of the gross clustering as possible with a basic cluster model, and then examine the residuals from fitting the model. Ogata, Zhuang and colleagues have recently developed various techniques, described in [86] and [143] for example, for examining the residuals from catalog data initially fitted by the ETAS model. Alternatively the cluster

model can be fitted locally (i.e. with parameters allowed to vary in time or in space and time), and the parameter variations examined to shed more light on the features of interest: [89] contains a compelling example of such an analysis.

We proceed to describe three types of cluster model, starting from the ETAS model itself. All three models are defined through the form of the conditional intensity function, as outlined in the discussion of the stress-release model. In all three models again, magnitudes are allocated independently and randomly, either according to the GR law, or some variant such as the tapered Pareto distribution for seismic moments. The final feature in common is that in all three models the main component in the conditional intensity is a linear combination of contributions from past events.

### The ETAS Model

The ETAS model (the initials standing for Epidemic Type Aftershock Sequence) first appeared in Ogata's paper [86], but was preceded by a series of studies by Ogata and colleagues in Tokyo on processes which, like the ETAS model itself, have conditional intensities of the linear, Hawkes type, following [34,35]. Earlier cluster models included the Neyman–Scott process, reincarnated as a 'trigger model' in [133] and [126].

In its basic time-magnitude form, the ETAS model has conditional intensity

$$\lambda^*(t, M) = \beta e^{-\beta(M-M_0)} \cdot \left\{ \mu + A \sum_{i:t_i < t} e^{\alpha(M_i-M_0)} f(t-t_i) \right\}, \quad (11)$$

where the first term on the RHS is the GR density for magnitudes,  $\mu$  is an arrival rate for background (ancestor) events, the constant  $A$  is related to the criticality of the process, the 'productivity function'  $e^{\alpha(M-M_0)}$  describes how the number of first-generating offspring increases with magnitude of the parent event, and  $f(u) = p c^p / (c+u)^{1+p}$ ,  $c > 0$ ,  $p > 0$  is the density (here a Pareto form) for the distribution of the temporal lag between the arrival or birth of the parent and that of its offspring.

In the full space-time-magnitude version

$$\lambda^\dagger(t, x, M) = \beta e^{-\beta(M-M_0)} \cdot \left\{ \mu h(x) + A \sum_{i:t_i < t} e^{\alpha(M_i-M_0)} f(t-t_i) g(x-x_i) \right\}. \quad (12)$$

The new terms are the density  $h$  of new arrivals over the spatial region, and the density  $g$  in space for the location of an ‘offspring’ event about its parent. We suppose that  $f, g$  and  $h$  are all normalized to form probability densities.

One of the main attractions of the Hawkes’ processes, including the ETAS model, is that they have a branching process interpretation, first pointed out in [36] and implicit already in the description of the conditional intensity. For example, the criticality parameter (mean number of offspring per ancestor, averaged over the magnitude distribution for the ancestor) is given by  $A/(1 - \alpha/\beta)$  for both the above forms. Thus a stable version of the process can exist only if  $\alpha < \beta$ , and then only if  $A$  is small enough. Of course, branching process ideas appear in many earthquake occurrence models, notably in Kagan’s work (e. g. [36]).

While the branching process interpretation gives much insight into the structure of the ETAS model, statistical analysis depends crucially on the representations in (11) and (12), since they lead to the relatively tractable form (10) for the likelihood.

For computational purposes, the likelihood of a general marked point process, of which the ETAS models are examples, is often written most conveniently in terms of the conditional intensity  $\lambda_g^*(t)$  for the *ground process*, the overall occurrence of points, irrespective of location or mark, and the conditional mark (in our case space and magnitude coordinates) distribution  $f^*(x, M|t)$ , so that  $\lambda^*(t, x, M) = \lambda_g^*(t) f^*(x, M|t)$ . The star indicates that the quantities so labeled are in general conditional on (and hence functions of) the histories up to time  $t$ . Provided  $f^*$  is normalized to a probability density for any given past history, we can write the likelihood ratio in the form

$$\log L_1/L_0 = \left[ \sum_{i=1}^{N(T)} \log[\lambda_g^*(t_i)/\lambda] - \int_0^T [\lambda_g(t) - \lambda] dt \right] + \sum_1^{N_g(T)} \log[f^*(x_i, M_i | t_i)/f(x_i, M_i)], \quad (13)$$

where the terms  $\lambda$  and  $f(x, M)$  relate to the rate and mark distribution for the background process (null model), here taken to be a constant rate Poisson process with independent (and usually GR) magnitudes.

This form represents the likelihood ratio as the sum of two terms, the first involving the time points only, and the second involving the marks (spatial locations) given the time points. In many models, the parameters appearing in the two terms have no common variables, in which case optimization can be carried out for the two terms separately.

Because the ETAS model fits well to catalogue data over a wide range of scales and contexts, in recent years its properties have been examined in detail, with the aim of verifying its ability, or otherwise, to reproduce specific features of the real process, such as Bath’s law or the occurrence of foreshocks; see, for example, [37,38].

Moreover, the procedures developed by Ogata and colleagues for fitting versions of the model in which the parameters can vary in both location and time, and for detecting local departures from a good fit of the model, have made the ETAS model a powerful diagnostic tool. In this way it has been used to estimate local variations in the stress field (e. g. [89,91]), or changes in seismicity due to the intrusion of ground water [30,58].

For long it was believed that an immigration component, coupled to a subcritical branching structure for the offspring, was the only way to produce a stable process with branching structure. However, it was shown recently in [12] that when the temporal lag distribution  $f(\cdot)$  of (11) is very long-tailed, a critical Hawkes process can sustain itself indefinitely as a ‘process without ancestors’. Another somewhat unexpected extension, described in [132], is to a self-similar version over an infinite range of magnitudes.

Perhaps the one serious limitation of the ETAS model is its rather poor performance as an intermediate-term predictor. The reason for this is that its predictive power is basically dependent on its ability to fit aftershock sequences. Hence it does not show significant gains, even against the Poisson model, until the time intervals between forecasts are of similar order of magnitude to the time intervals between the larger events in an aftershock sequence.

### The Kagan–Jackson Models

Kagan and Jackson have proposed a number of forms of which we refer to two, the long-term and short-term versions of [41]. The long-term version has its origins in [51], while the short-term version has its origins in [55].

In the long-term model, the current value of the conditional intensity, within a spatial region  $A$  and based on observations within  $(0, t)$ , has the form

$$\lambda^*(t, x, M) = f_t(x)g(M)h(t) \quad (14)$$

where  $h(t)$  is the overall current risk (ground process intensity),  $g(M)$  is a (fixed) magnitude distribution, commonly that corresponding to the tapered Pareto form for seismic moments, and

$$f_t(x) = \sum_{0 < t_i < t} k(x - x_i) / \sum_{0 < t_i < t} \int_A k(x' - x_i) dx'$$

$k$  being a spatial kernel function. Thus  $f_t$  is a normalized sum of contributions from previous events within the observation period and spatial region. It is time-independent except insofar as the advent of additional events requires additional renormalization.

Although the model is well-defined by its conditional intensity and an initial condition at  $t = 0$ , and can be fitted by likelihood methods much as for the ETAS model, the renormalization introduces a non-linear component into the model which makes its properties more difficult to analyze than those of the ETAS model. Moreover, like other forms of moving average model, it is non-ergodic: there is no unique stationary form to which it will converge from different initial conditions. Nevertheless, it serves the principal purpose of providing a baseline comparison for other putative prediction models for the same region.

The short-term form is very close to the spatial ETAS model. As in the ETAS model, each past event is associated with both a spatial and a temporal decay function, the temporal decay following the Omori law. Again, however, it involves a renormalization rather than the introduction of an explicit immigration term into the model, although some versions of the model allow a small quota of ‘surprises’ in parts of the region with no previous earthquakes.

An important role for both models has been to focus attention on the need for systematic, long-term evaluation and comparison of forecasting models, and to provide a more relevant null model than the constant-rate Poisson process.

### The EEPAS Model

This model grew out of several decades of experimentation with precursory swarm models by Evison and Rhoades; see [24] and [104] for more of the history and underlying concepts. The precursory swarm models identify groups of moderate-sized earthquakes as precursory swarms and use these as possible precursors of large earthquakes. Since both precursory phenomena and forecast phenomena belong ultimately to the same process of earthquake formation, a more satisfactory approach is to try and develop a joint model for both phenomena. This, in effect, is what the EEPAS model achieves. It has its own rationale, based on a theory of growth and development of crustal fractures outlined in the papers cited above, and has been successfully applied in several major seismic regions (e. g. [100] and [105]).

Much as in the ETAS model, the conditional intensity has the general form

$$\lambda^*(t, m, x) = \mu\lambda_0(t, m, x) + \sum_{t_i < t} w_i \eta(m_i) r(M | M_i) f(t - t_i | M_i) g(x - x_i | M_i), \quad (15)$$

but the details are significantly different.

First, the conditional intensity in (15) is not taken to apply to the whole catalogue from which the events on the right side are derived, but only for events above a higher threshold. Thus, the model might be used for modeling (and predicting) events over magnitude 5.8, but would take explanatory data from the catalogue of events with  $M \geq 4$ .

Second, the functions  $f$  and  $g$  are not based on Omori-type decay formula, but on logarithmic regressions for the time and space delays between an initiating event and the event it may precede, and the magnitudes of the two events. Thus for example

$$f(u | M_i) = \frac{1}{u\sigma_T\sqrt{2\pi}} \exp\left[-\frac{(\log u - a_T - b_TM_i)^2}{2\sigma_T^2}\right],$$

with an analogous expression for  $g(w)$ , while  $r(m)$  takes the form

$$r(m | M_i) = \frac{1}{\sigma_M\sqrt{2\pi}} \exp\left[-\frac{(m - a_M - b_MM_i)^2}{2\sigma_M^2}\right].$$

They differ considerably from the functional forms used in the ETAS model, but are similar to relations used in the precursory swarm models.

The weight factors  $w_i$  are commonly set to unity, but in more refined analyzes may be down-weighted when the triggering event has been identified as an aftershock. One way of finding suitable weights is to carry out an initial ETAS stochastic declustering, as in [143], and base the weights on the probability that a given event is independent.

The further normalizing factor  $\eta(m_i)$  in (15) is introduced, much as in the Kagan–Jackson models, to offset the absence of any immigration term, and to compensate for the input from earthquakes below the magnitude threshold. It is adjusted for each magnitude class  $m_i$  so that the overall rates follow the G-R law.

In practice the contribution from the baseline rate density is often so small as to be negligible.

Our impression is that the EEPAS model is currently the best-performing of the general seismicity models in the sense of producing the highest average probability gains or entropy scores (see Sect. “Assessing Probability Forecasts”) for predicting moderate to large events on intermediate time scales.

## Stochastic Models with Precursors

### General Considerations

The search for reliable earthquake precursors has a long and troubled history. High hopes in the 1970s met many disappointments, some at least arising from an inadequate appreciation of the many statistical pitfalls. These difficulties are now much better appreciated, but even so there are relatively few studies based on a satisfactory statistical model, incorporating a proper assessment of the uncertainties, and showing a significant precursory effect.

From a modeling point of view it is important to distinguish between *complete* and *partial* models. In a complete model, both the earthquakes and the precursors are included as components of an overarching joint process, the earthquakes forming one marginal process and the precursors another. In principle such a complete model should be the aim, but in most situations either the background physics is insufficiently understood, or observations on the precursors are inadequate, or the statistical analysis is too difficult, to allow such a joint analysis to proceed. In many situations also, the modeling and analysis lie outside the realms of conventional statistical models (in dealing with self-similarity, for example), raising further procedural problems.

In a partial model, no attempt is made to model the precursors as a stochastic process. They are treated as given, and their data used in regression-like procedures to modify the probabilities of earthquake occurrence. They can be used retrospectively to examine the performance of a suggested predictive relation, but their use in developing probability forecasts is limited to the short term because there is no model to forecast the future behavior of the precursors.

In the later parts of this section we illustrate some of the modeling approaches that can be used with precursors, using examples drawn mainly from our own experience.

One issue that commonly arises is that the precursory signal data are derived from observations taken at fixed sampling intervals. To match such data to the point process data for the earthquakes, it is generally easiest to switch the whole analysis to a discrete-time study. In this case the continuous time point process models of the previous sections need to be replaced by approximating discrete time models. The two most common of these are the logistic, or binary data, models, and the discrete Poisson process models, illustrated in the first two examples below. In these two classes of models, the probabilities  $p_n$  that an event occurs in the  $n$ th interval (respectively, the means  $\mu_n$  of the Poisson distribution for that interval) play the

role of the conditional intensity function  $\lambda(t)$  of the continuous time models of the previous section.

### Example 1: Logistic Regression Analysis of M8 Series

Logistic regressions are used to directly assess the effect of precursor observations on the event probabilities  $p_n$ . Suppose that at the  $n$ th interval, observations  $(U_1^n, U_2^n, \dots, U_k^n)$  are available on  $k$  precursors. Dropping the  $n$  for brevity, the logistic regression takes the general form

$$\log \left[ \frac{p}{1-p} \right] = \alpha_0 + \sum_{k=1}^K \alpha_k U_k \quad (16)$$

where the left side is the logit (log-odds) transform of the event probability, and the right side is the regression term. This representation corresponds to the canonical form for a binomial distribution as a member of the exponential family. Standard routines exist for estimating the parameters  $\alpha_k$  by maximum likelihood or closely related methods, and form part of the generalized linear model procedures.

As an example we consider the model used in [32,33] on the output from the M8 algorithm on New Zealand data.

The M8 algorithm itself is not a stochastic process model, but a decision procedure for calling an earthquake alert based on the analysis of 7 contributing series from earthquake data within a specified 'region (circle) of investigation'. It is the best known of a number of pattern-recognition algorithms developed by the Russian group headed by Keilis-Borok during the 1970s and subsequently. The basic form of the algorithm is described in [56] and [59], with recent reviews in [60,61].

The heart of the algorithm consists of a set of decision rules, based on the joint behavior of the 7 time series, for calling an alert, or more specifically the announcement of a 'TIP' (time of increased probability of an event above a given magnitude threshold) over the region of investigation. The time series are updated every six months, and a TIP extends for three years in the first instance.

A key feature of the analysis in [32,33] is that, in each six-month interval, the values of the seven series are combined by a non-linear formula (linear methods seem less effective) into the value  $U_1^n$  of a 'critical series' which is then used as the single precursor in a logistic regression model. The non-linear formula mimics the structure of the decision rules used in declaring a TIP.

The logistic regression analysis then provides the probability for the occurrence of an event over the specified magnitude threshold within the region of investigation for

the current 6-monthly period. Much of the further discussion in [32,33] is concerned with combining the outputs from overlapping regions of investigation.

Note that the analysis is typical of that for a partial model; a complete model for M8 would model the joint distribution of the M8 series and the target events.

### Example 2: Discrete-Time Poisson-Type Model for ULF Electric Signals

In a discrete-time Poisson-type model, the number of events  $Z_n$  in the  $n$ th time interval is modeled as a Poisson variable with mean  $\mu_n^*$  that is treated as a function of the past history in much the same way as the conditional intensity in the continuous-time model. The likelihood ratio against a constant mean Poisson process takes the form

$$\log L/L_0 = \sum_1^N Z_n \log(\mu_n^*/\mu) - \sum_1^N (\mu_n^* - \mu). \quad (17)$$

The analogy with the point process form (10) is very obvious, particularly as the time intervals become small so that with high probability the  $Z_n$  are either 0 or 1.

The way in which  $\mu_n^*$  depends on the past can be very general, subject only to the constraint  $\mu_n^* \geq 0$ . In particular,  $\mu_n^*$  can depend on prior observations both of the process itself and of auxiliary (precursory) variables. If a linear conditioning of the type (16) is required, the multiplicative form

$$\log \frac{\mu_n^*}{\mu} = \sum_{k=1}^K \alpha_k^{(n)} U_k^{(n)} \quad (18)$$

can be used, ensuring that  $\mu_n$  is positive, and slotting into the canonical form for the Poisson distribution as a further member of the exponential family, so that the generalized linear model procedures become available. However, the form of the likelihood (17) is usually simple enough to be maximized directly even in more general cases.

An example is the analysis of ULF (ultra low frequency) electric field data in [144].

The ULF signal referred to here is made up of small fluctuations in electric potential measured some meters below the ground surface by sensitive and well shielded electrodes. Its role as a precursor, and the physical explanation of the phenomenon, if it exists, are still unclear. The data analyzed in the paper cited come from some thirty years of recordings from stations around Beijing in China.

Here the base-line (reference) model is a self-exciting (Hawkes type) process in discrete time for the daily earthquake numbers in a wider region around Beijing. Such a model is needed as a reference model because it

takes into the inherent clustering effects of the earthquakes themselves. Otherwise there is a temptation to interpret (wrongly) all improvements over the constant rate Poisson process as due to the signals and not to the inherent clustering of the earthquakes. The regressands were the daily readings of ULF anomalies at a set of some half-dozen recording stations around Beijing, simplified to form 0-1 series of observations above a threshold. In fact the daily numbers of earthquakes were small enough for the corresponding continuous and discrete time models to be essentially identical.

Two analyzes were carried out, first with a linear Hawkes-type representation incorporating the effects of past earthquakes alone (self-exciting model), none of the ULF data being used, and second with a double (mutually exciting) linear Hawkes-type representation for the effects of both past earthquakes and ULF signals on the current rate. Likelihood ratios were taken first with respect to a constant rate Poisson process, optimizing parameters in both cases, then (as a ratio of ratios) for the first and second models against each other, to allow the improvement due to adding in the information from the ULF signals to be assessed.

In this study the model was also tested in reverse mode, to see whether the earthquakes improved the likelihood performance of a Hawkes-type model for the ULF signals alone. The results were positive in the direct mode and negative in the reverse mode. Either way, the models were partial, not complete, as no attempt was made to provide a joint model for the earthquakes and the ULF anomalies together.

Discrete-time Poisson-type models are also used as the basis for model testing within the RELM testing center in Southern California [111]. The modeler supplies the Poisson rates  $\mu(n, r, m)$  not only for each time interval ( $n$ ) but also for each spatial bin ( $r$ ) and magnitude bin ( $m$ ). The (approximative) assumption is then made that all the Poisson variables relating to a given time interval are conditionally independent given the current Poisson rates, so that a likelihood ratio of the form (17) can still be used, and made the basis of comparing different proposed models.

### Example 3: Point Process Regression Models

As already hinted at, continuous time (point process) procedures can be developed along similar lines to those in the previous example, if the past history includes information on auxiliary (precursory) variables as well as the history of the point process itself.

The form of such dependences can be very general, but if a linear form is wanted it can be incorporated through

expressions of the form

$$\log[\lambda^*(t)/\lambda_0^*(t)] = \sum_{k=1}^K \alpha_k U_k(t) \quad (19)$$

for the ratio of intensities, in a similar way to the multiplicative form for the discrete Poisson model. In the special case that  $\lambda_0(t)$  corresponds to a renewal process, this is the well-known *Cox regression model* [18]. A comprehensive treatment of models of this kind, with mainly medical and social science applications, is in [2].

Few examples of this type are known to us in the seismological literature, and this may be one area with scope for further development. For example, it is possible that the time-predictable model could be reformulated as a Cox regression model, by building in the dependence on the size of the previous event as a regressand. The closest to a model of this kind that already considered in this article is perhaps the EEPAS model, where the events used in developing the conditional intensities are mainly smaller than the events being modeled, and so could be described as precursors. Thus the conditional intensity for the larger events being modeled is regressed onto the magnitudes, times and locations of the smaller events in the catalogue. If the renormalization and immigration terms were omitted, the conditional intensity (15) of the EEPAS model would have a similar basic form to (19) above.

#### Example 4: Foreshocks

Foreshocks, with their potential for earthquake prediction, have been of interest to seismologists since the days of Omori and other earthquake pioneers. The hope that increased quantity and quality of catalogue data would lead to a more definitive picture of how, when and why foreshocks occur has not yet been realized, however. Features which would discriminate foreshocks from other earthquakes or earthquake clusters have proved hard to identify, although careful studies by Ogata and colleagues (e. g. [90]) do suggest that limited opportunities for discrimination may exist.

As precursors, foreshocks retain a specific but limited role. Some degree of forecasting power is available simply from the fact that any newly-observed event, outside those in clearly defined aftershock sequences, has the potentiality of being a foreshock. Recent studies of foreshock occurrence from this point of view are given (among others) in [46] for Southern California, [96] for large global events, [80] for New Zealand earthquakes.

The studies suggest that, leaving aside events in an obvious aftershock sequence, between 5 and 10% of earth-

quakes with magnitudes 4 and over are likely to be followed by a larger event within time and space windows of the order of 4–5 days and 20–30 km radius respectively. No very sophisticated stochastic model is required to describe such a feature: within the defined window, the probability that a larger event will occur is simply increased from its background value to about 5%, and multiplied by a standard GR factor to take into account magnitude variation; a further separate factor can be used to take into account sub-regional variations of foreshock probabilities.

In the last few decades foreshocks have been studied from a more general point of view, as evidence or otherwise of the self-similarity of earthquake occurrence. In a branching model such as the ETAS model, no foreshock feature is explicitly built into the model, but from among a given parent's offspring, one will occasionally appear with a magnitude larger than that of its parent – with frequencies again in the vicinity of 5%, approximately irrespective of the magnitude of the parent.

That just such an interpretation may apply also to real earthquakes is suggested in [25]. Foreshocks then are not a specific physical phenomenon, but just parent events which happen to have offspring larger than themselves. An interesting connection between the probability that an initial event is a foreshock (followed by a larger event as offspring), and the distribution of the Bath's law magnitude gap, is put forward in [25] and [38].

#### Further Topics

There are many further topics where stochastic modeling ideas are relevant, even if they do not necessarily involve the development and fitting of a full model. For example, the last few decades have seen considerable work on the development of procedures for producing and assessing probability forecasts, and for quantitatively describing spatial or space-time point patterns. In this section we list what seem to us to be some of the more important topics of this kind, space precluding more than very cursory accounts.

#### Generating Probability Forecasts

There is no longer any very hard and fast line between predictions and probability forecasts. It has long been recognized that any prediction of the time and place of a forthcoming event must be accompanied by some statement of the uncertainties in the prediction. But this, as discussed in the introduction, is precisely the motivation for introducing stochastic models. Using such a model, the uncertainties can be rephrased in terms of the probabilities of occurrence within specified time intervals and spatial regions,



i. e. by probability forecasts. In our view such probability forecasts represent the most useful way of summarizing information about the uncertainties regarding future events. In statistical jargon, they represent the ‘predictive distributions’ for future events, and form the basis not only for the probability forecasts themselves, but also for any associated cost-benefit analyses.

Even within a probabilistic framework, the idea that precise, medium-term or even short-term forecasts may be possible has looked increasingly like a pipe dream over the last few decades. Nevertheless, it is not yet entirely ruled out. While the emphasis in the last few decades has been on increased surveillance, improved knowledge and understanding of long-term hazards, and the reduction of risks from earthquake hazards, medium term (months to years) probability forecasts are winning a new role as refinements of more traditional long-term, static hazards for both building and insurance purposes. In addition very short term forecasts play a useful role in connection with the progression of aftershock sequences, and in developing real-time warnings for trains, gas supplies, and other facilities at high risk from a serious earthquake.

All of the models outlined in the previous sections can be used to develop probability forecasts. Typically, such forecasts are derived from simulations rather than from analytical studies. For example, simulation schemes for use with conditional intensity models are described, with further references, in Chap. 7 of [20].

For complete models, forecasting schemes making use of such procedures involve first fitting a model on all the data up to the current time, then simulating the model as far as is desired into the future, and using these simulations to estimate any required probabilities or expected values as in a Monte Carlo study.

Forecasts for partial models can proceed along similar general lines, although in the first instance the dependence on auxiliary variables restricts the probability forecasts to just the next forecasting period. To obtain forecasts beyond this first step, and in the absence of any updated precursor data, it is necessary to develop a further set of 2-step forecasts, using only the current precursor data, and so on successively, the forecasts gradually reducing in power.

### Assessing Probability Forecasts

Two main approaches to assessing probability forecasts are through likelihoods and probability gains, and through their performance in decision schemes based on the forecast probability exceeding some threshold. Since most procedures are reduced in practice to forecasts for finite

forecasting periods, we outline assessment procedures for this case only.

The *probability gain* for each forecast (i. e. for each time, time-space or other interval for which a probability is forecast) is the ratio of the forecast probability for the observed occurrence in that interval to the corresponding probability from a standard reference model, such as a simple or compound Poisson process, for that interval.

The sample averages of the log-probability-gains, or *entropy scores* in the terminology of [34], over all or certain classes of intervals or of observed outcomes, provide useful insights into the performance of the forecasts. From them one can quickly perceive the outcomes which the scheme is forecasting well, and those which it is forecasting badly.

The overall average, or more properly the expected value, of these entropy scores is called in [34] the *information gain* relative to the reference model. In a complete model, the information gain is a numerical characteristic of the model, giving an upper bound for the improvement in performance that can be expected for the proposed model, relative to the reference model, when the proposed model is the true model. In many cases it reduces to the expected value of the mean likelihood ratio. For example, if successive forecasts, derived from the use of a particular model, give probabilities  $p_1^*, p_2^*, \dots, p_N^*$  for the successive observed outcomes, and  $\bar{p}$  is a constant probability used as reference, then the sum

$$(1/N) \sum_1^N \log[p_n^*/\bar{p}]$$

is the empirical entropy score for that model as well as the mean log-likelihood ratio. It approximates the corresponding expected value (the information gain) if the model used is the true model. If the true model is unknown, as will usually be the case in practice, the above average approximates the difference in the Kullback–Leibler distances between the given model and the true model, and the reference model and the true model. For a set of models, that giving the largest entropy score should be that closest to the true model in the sense of Kullback–Leibler distance. In a partial model, expectations cannot be taken over values of the auxiliary variable, since its distribution is not included in the model specification, but at least empirical averages based on past observations can be developed. [34] gives further background and examples; the basic idea of using log-probabilities (i. e. loglikelihoods) as an indicator of forecasting performance goes back at least to [52].

We noted in Subsect. “**Example 2: Discrete-Time Poisson-Type Model for ULF Electric Signals**” that such com-

parisons of likelihood ratios form the basis of the assessment procedures used in the RELM testing center for probability forecasting schemes; see further [111].

The more traditional procedures for assessing probability forecasts suppose that the schemes are first turned into prediction schemes by predicting that an event will occur whenever the forecast probability of that event exceeds a certain threshold value. The results can then be put into a 2 x 2 table (occurrences or non-occurrences, versus predictions or non-predictions). If the entries in the table are labeled as

- (a) number of successful forecasts of occurrence,
- (b) number of failures to predict,
- (c) number of successful forecasts of non-occurrence,
- (d) number of false alarms,

the commonly used *R-score*, or *Hanssen–Kuiper skill score*, can be defined as

$$R = \frac{ac - bd}{(a + b)(c + d)} = \frac{a}{a + b} - \frac{d}{c + d} = \frac{c}{c + d} - \frac{b}{a + b}. \tag{20}$$

The *R-score* varies between  $-1$  and  $+1$ , with  $0$  denoting forecasts independent of outcomes, and the two extreme values perfect non-prediction and perfect prediction, respectively. [91] illustrates the use of the *R-score* to evaluate Chinese yearly forecasts, even though these are based on expert opinion rather than any probability model.

The usefulness of this approach was greatly extended in [84,85], by allowing the threshold probability  $p_c$  to vary and calculating for each such value the ratios

$$v(p_c) = \frac{c}{(a + c)}$$

and

$$\tau(p_c) = \frac{(a + b)}{(a + b + c + d)}$$

The resulting  $v - \tau$  diagram, obtained by plotting these quantities against each other, provides a comprehensive summary of the behavior of the probability forecasting scheme. The diagram typically consists of a convex (downwards) curve reducing to the diagonal joining the points  $(0, 1)$  and  $(1, 0)$  in the case of purely random forecasts, and to the two  $(0, 1)$  segments of the axes in the case of perfect prediction. It is essentially a Q-Q (quantile-quantile) plot of the distribution of the proportion of time on trial, and the the proportion of failures to predict. See the papers quoted or [34] for further details.

It should be emphasized that the procedures described are concerned with the scientific issues of assessing the quality of proposed models rather than the practical issues of issuing and using probability forecasts. The latter raise just as many difficulties, if not more, than the former. Among these are the need to take into account errors in the model as well as the uncertainties described by the model itself, the need to develop decision-making frameworks that can take advantage of the information within probability forecasts, and the need to address the social, political and economic consequences of issuing forecasts.

### Change-Point Models

Change-point models are used retrospectively to identify time-points at which a change occurs in quantities such as a mean value or rate. They are not themselves precursors, but are used rather to indicate the onset of a period which is anomalous in some sense, and may therefore have some precursory significance. In practice, one of the biggest difficulties in using such methods with seismicity data lies not in identifying the change point, but determining the region from which the data for the change-point analysis should taken.

The procedure consists essentially of dividing an observation time period, say  $(0, t)$ , into two segment  $(0, t_1)$  and  $(t_1, t)$ , and finding the value of  $t_1$  which maximizes the discrepancy in the values of the quantity being studied. The value of this maximized discrepancy can then be tested for significance using the null hypothesis that the values in both periods are equal.

Such a change-point technique is developed in [86] to detect the onset of precursory quiescence within a selected observation region. The data is first fitted to the ETAS model. The time axis is then transformed by the random time transformation

$$\tau = \int_0^t \lambda^*(u) du$$

which has the effect of transforming the original point process, with conditional intensity  $\lambda^*(t)$ , into a unit rate Poisson process (see, for example, Sect. 7.4 in [20]). If, however, there is a change in the parameters of the original ETAS model, this will show up as a change in the rate or other perturbation in the unit rate Poisson process, for which many tests are available.

This technique can be applied both to background events in a specified region (e. g. [140]), or to events within the course of an aftershock sequence, anticipating the occurrence of a large aftershock (e. g. [77]).

Ogata, Toda and colleagues make use of similar but more sophisticated procedures for identifying the regions and onset times of stress-shadowing through careful monitoring of activity of small events; see in particular [89,91].

A different type of change-point analysis has been developed by W. Smith for detecting changes in  $b$ -value. Although one of the earliest possibilities suggested, the precursory role of  $b$ -value changes as a precursor has never been unambiguously identified. Early papers such as [82] or [109] suggested a link to physical features such as heterogeneity of the crustal material, or changing levels of stress. The latter interpretation is at least partially supported by the branching and similar models for earthquake mechanism outlined in Sect. “**Branching Models**”.

Smith’s approach is based on modifications of the CUSUM procedure widely used in quality control contexts for detecting departures from normal behavior in a production process. They are essentially methods for detecting a change in slope of cumulative occurrences or other similar sums. For further details and reviews of earlier work on  $b$ -value changes, see [117,118].

Finally, there is the possibility of a Bayesian approach to change-point problems in seismology; some discussion and an example are given in [94].

Change point models are closely related to *hidden Markov models* in which the rates and other characteristics of the observed process change with the state of a Markov process which is hidden or at best only partly observable. Such models have been widely used in speech recognition and modeling of IT traffic (see, e. g. [72]). Their use in seismology is relatively new (see [23] for a recent example) but their potential seems worthy of further exploration.

### Moment Measures and Correlation Functions

Moment structure plays an important role in most stochastic processes, and the same is true for point processes. The main attention is on second order or correlation properties. Indeed, where distributions are more or less Gaussian, the distinction between models and second order properties is largely nominal, since a Gaussian model is fully described by its means and covariances. The same is not true for count data, for which the second order properties form an important but not in general a definitive aspect of the overall model structure, since a range of different models can be developed to fit particular second order characteristics.

Early papers in the seismological context include [16, 53,128]. There is also a considerable literature in astrophysics, relating especially to the distribution of galaxies

and the role of 2-point, 3-point and higher order correlation functions (e. g. [76]).

Second-order properties for point processes mean properties of point-pairs. Their behavior is described through the second moment measure, whose density (when it exists) is given for  $x_1 \neq x_2$  by

$$m_2(x_1, x_2)dx_1dx_2 = E[N(dx_1)N(dx_2)] .$$

Apart from a renormalization, this is also the 2-point correlation function of [53]. It is also the basis of the covariance measure, with density

$$c(x_1, x_2) = m_2(x_1, x_2) - m_1(x_1)m_1(x_2) \quad (x \neq y) .$$

This also exists in various renormalized forms, for example the radial correlation function defined, when the process is isotropic, by

$$\rho(r) = dK(r)/dA(r)$$

where  $A(r) = \pi r^2$  is the area of a circle centered on an arbitrary point of the process, and  $K(r)$  (often referred to as ‘Ripley’s  $K$ -function’) is the expected number of additional points (i. e. apart from the point at the center) in the same circle (see e. g. [22,104]).

In a general treatment, the coordinates  $x_1, x_2$  may combine both space and time components. The second moment densities then give information about the expected density of occurrence of additional points at given time or space intervals about a given point taken as origin. In this way they can display distance variations in the strength of the clustering tendency.

The second order techniques are particularly valuable when the process is stationary in time, so that spectral methods can be used. In this situation the covariance measure becomes a function of time through the difference  $t_1 - t_2$  in the time coordinates of the two points being considered:

$$c_2(x_1, x_2) = c_2(t_1 - t_2; y_1, y_2) ,$$

where  $y_1, y_2$  may represent locations or magnitudes. Taking Fourier transforms with respect to time leads to a spectral density, multivariate if the dependence on space or magnitude is retained, which can be used for the analysis of periodic effects in point processes. [133] is an early example of the spectral analysis of earthquake occurrence data. More general discussions can be found in [13] or Chap. 8 of [20].

### Principal Component Analysis

Principal components are the names given to the eigenvectors in the orthogonal decomposition of a model-derived

or empirical covariance matrix. Being symmetric and positive definite, the diagonal representation of such a matrix has non-negative eigenvalues, which measure the proportion of variation associated with the given eigenvector. Thus the principal components associated with the largest eigenvalues define those linear combinations of the observation vector components which explain the greatest amount of variability.

In geophysical applications, notably in meteorology, the observations typically arise as time sequences of observations from a set of recording stations. Each time point gives a vector observation. The principal components associated with the largest eigenvalues then often describe recognizable weather occurrence patterns, and are used as a means for identifying and classifying such patterns.

This idea admits many specializations and extensions. For stationary processes in time, for example, the covariance matrix has a special structure ( $c_{i,j} = c_{i-j}$ ) which for infinite series, or a finite series with periodic structure, causes the eigenvectors to reduce to complex exponentials  $e^{i\omega r^n}$ . These periodic terms form the principal components of the sequence, while the corresponding eigenvalues are amplitudes associated with the given frequencies. Thus standard spectral analysis can be interpreted as a special form of principal component analysis. The so-called Karhunen–Loeve theory (see e. g. [69]) is an extension to very general covariance operators admitting a diagonal decomposition.

Adaptations of these ideas for earthquake data have recently been considered by Rundle, Tiampo and colleagues (see e. g. [107,120]). Here the data is typically of point process form (0-1 functions) on a sequence of observation regions, centered either on a lattice of space or space-time points, or on the centers of a family of small fault segments over an interacting fault system. Their papers combine the Karhunen–Loeve theory with ideas originating in dynamical systems theory, and seek to extend the analysis to provide short or medium term probability forecasts.

### Fractals and Fractal Dimensions

Many of the models so far considered show long-range dependence, power-law distributions, and some form of self-similarity (similar appearances on different scales). The term ‘fractal’ is commonly used in this context, with a variety of interrelated interpretations. A wide-ranging review of such properties for earthquakes is given in [121].

In the basic text [74] of Mandelbrot, fractals are introduced as a class of geometric objects (sets) with features which repeat on a family of reducing scales. The Cantor set in one dimension and the snowflake curve in 2 dimen-

sions are well-known examples. Their characteristic feature is that the Hausdorff dimension of the fractal set may differ from the dimension of the space in which it is embedded. Thus the Cantor set has dimension  $\log_3(2) < 1$  despite being embedded in a 1-dimensional space. As in this example, these aberrant dimensions often have a non-integer value, whence the name.

Various empirical procedures have been devised for determining these fractional dimensions. For example, for a set in 2 dimensions, we may count the proportion of cells from a 2-dimensional lattice which have a non-empty intersection with the set, and consider the behavior as the cell size in the lattice approaches zero. Typically this results in a log-log relation between number of affected cells and the cell dimension, corresponding in fact to power-law behavior. Such methods can be applied also to physical sets such as the traces of faults on the earth’s surface (see [121] and [45] for examples). If a suitably linear portion of the log-log plot can be found, the slope is used as an estimate of their fractal dimension.

Because power-laws arise inevitably in the study of fractal sets, the term *fractal behavior* has come to be used loosely to describe the behavior of any probability distribution for which

$$1 - F(x) = \Pr(X > x) \sim cx^{-\alpha} ,$$

or even when power-law forms arise in features such as time or space correlation functions.

A different fractal concept derives from the work of Renyi [99]. Renyi’s dimension estimates (also called multifractal dimensions) reflect the irregularities of a measure (distribution) rather than a set. Thus, if a probability distribution on a square is used as an example, and the square is divided into a lattice of small squares  $\Delta_i$  each of side  $\Delta$ , we may consider the limit as  $\Delta \rightarrow 0$  of the ratios

$$d_q(\Delta) = \left[ \log \sum_i p(\Delta_i)^q \right] / \log \Delta \quad (q \neq 1) \quad (21)$$

with

$$d_1(\Delta) = \left[ \sum_i p(\Delta_i) \log p(\Delta_i) \right] / \log \Delta . \quad (22)$$

For each  $q$ ,  $d_q$  defines a *multifractal dimension* associated with the distribution, and for varying  $q$  the family of such dimensions defines a type of transform of the underlying probability distribution.

The term multifractal is used because the distribution may exhibit different forms of singularity, associated with different power-law behavior, at different points of

the space. In applications, for example in turbulence and meteorology as well as seismology, the process is visualized as a descending sequence of random eddies ('random cascades') with dimensions shrinking to zero. It is in this sense that Kagan [49] coined the phrase 'frozen turbulence' to describe earthquake processes in the crust.

The computations in the equations above can be applied empirically to quantities such as the counts or the energy release from earthquakes and used to estimate the multifractal dimensions of some underlying spatial distribution. For general discussions of multifractals, see [95] and [31]; the latter gives an extended discussion of applications of the theory to earthquake data.

With count data, there is a link between the multifractal dimension  $d_q$  for integer  $q$  and the  $q$ -point correlation function. In particular  $d_2$  describes the growth rate of the two-point correlation function for vanishingly small separations; see [53,131].

### Self-Similarity

Self-similarity is used in a broad sense to describe processes in space or time where, much as in a fractal set, key structural features are preserved on a descending sequence of scales. Plots of fault traces or epicenters do indeed suggest such scale invariance properties.

In stochastic process theory, self-similarity (auto-modeling) is used to denote a precisely defined property of random processes or random measures. In particular, a random measure  $\xi(A)$  defined on sets in the plane (or other Euclidean space) is said to be *self-similar*, with similarity index  $H$ , if a change of scale in the space can be compensated for by a change of scale in the quantity being measured:

$$\xi(rA) \stackrel{D}{=} r^{-H} \xi(A), \quad r > 0 \quad (23)$$

where  $\stackrel{D}{=}$  means the two sides have equal probability distributions.

Self-similar random measures of this form arise in turbulence theory, seismology, finance, and elsewhere. In one dimension, the concept can be reinterpreted as implying that a stochastic process has *self-similar increments*. Brownian and fractional Brownian motions both have increments of this type. However, the Brownian motions take both positive and negative values, whereas any random measures associated with earthquakes (for example, by way of energy release) should be non-negative.

Thus the self-similar processes arising in seismology have a rather special character, since they must be non-negative as well as self-similar. The only known examples

are purely atomic random measures:  $\xi(A)$  is obtained by summing the contributions (energies, for example) from point-events in  $A$ .

One such example is well-known: the *stable* random measures, which are characterized by a strong independence property, and a power law distribution for event sizes equivalent to a G-R relation for magnitudes. Superficially, this example matches the earthquake data rather well, but it fails to incorporate one key feature of earthquake occurrence: the lack of independence inherent in earthquake clustering. Recently, however, it was shown in [132] that it is possible to define a variant on the ETAS model which combines self-similarity with a non-trivial dependence (cluster) structure. It is not known at present how wide this class of models may be.

### Block-Slider and Related Mechanical Models

The mechanical models attempt to illuminate the processes of earthquake occurrence by devising complex mechanical systems that will reproduce many observed features of earthquake catalogs. The reason for including a brief section on such models in the present article is that one of their underlying purposes is to demonstrate that deterministic models of sufficient complexity can exhibit just that random-appearing behavior that characterizes the appearance of earthquake data. Two important issues then arise. The first is whether the mechanical models can be matched closely enough to any real seismic system to produce forecasts similar to or better than those produced by the stochastic models. The second is whether those stochastic models currently fitted to earthquake catalogs also provide good descriptions of the catalogs produced by the mechanical models. Such studies may then suggest ways in which the stochastic models themselves can be improved.

In the pioneering article [14], the mechanical system comprises a series of blocks, linearly connected by springs, which are pulled over a rough plane. The resultant movement is not smooth, but jerky, as the tension in one of the springs builds up to the point where it overcomes the frictional forces opposing motion and a slip occurs. Often movement of one block will initiate movement in a whole set of blocks.

The population of such movement sequences is then compared to a population of earthquakes. Although the motion of the system is entirely deterministic, analytic solutions are too complex to obtain in explicit form, and the behavior is irregular, with features of (pseudo)randomness. In particular a crude form of GR law generally results.

Many further mechanical models have been invented and studied, both in the laboratory and via numerical solution of the appropriate dynamical equations. Often the masses of the blocks or the spring parameters are selected at random from some plausible overall population. It is not this randomness which is the cause of the random-looking behavior, however, but rather the complexity of the movements (trajectory) of any particular system of blocks and springs or other similar elements.

One of the difficulties in matching such models to particular fault systems is to match the initial conditions, which are not easy to determine for real processes, even if they can be reproduced in the artificial system. It is precisely the relative simplicity of the stochastic model, which comes at the expense of describing many important physical details, which allows it to be fitted to catalogue data and then used to produce rough probabilistic forecasts.

In [70], one of the relatively few studies to fit point process models to the output from such mechanical systems, outputs were taken from the four different models described in [6] (see also [10]) and fitted by the simple stress release model. The four models illustrated different patterns of behavior, varying from highly random behavior with typical GR distributions, to regular, characteristic earthquake behavior. After suitable adjustment of the frequency-magnitude law in the stress-release model, it was found that the stress release model could be fitted to all four versions, and used to describe the overall energy state of the mechanical system; as such it operated as a crude predictor of the next major event in the system.

### Future Directions

There are many possible directions in which the applications of stochastic models in seismology could be extended and deepened. The fundamental limitations in the past have most commonly been limitations in data, and consequent limitations in the physical understanding of the processes. It has been the great improvement in data which has allowed the development of better and more insightful seismicity models in the last two decades; this tendency is continuing strongly and may well lead to unexpected new developments.

Two directions in particular appear to me to hold out scope for further development of stochastic models for earthquake occurrence. The first is related to the collection and integration of data on earth deformation. The extensive data now becoming available from GPS measurements have already led to new discoveries, for example 'slow earthquakes' which relieve strain without being

registered on conventional seismometers. The underlying problem is how to link the data on strain to data on seismicity. This is not easy, and is likely to require new ideas on both the physical and statistical sides.

The second relates to the systematic collection and evaluation of data on potential precursor events. Many of the ideas initiated in the 1970's have been abandoned, with the result that for most potential phenomena there exists no substantial body of data by which their effectiveness can be adequately tested, or the underlying physical processes modeled. Many controversial and so far inadequately explained phenomena fall into this category. They have the potential to generate projects of interest and importance from the physical as well as the hazard estimation viewpoints. Unfortunately collecting and archiving such data is a long-term process with uncertain future outcomes, and therefore difficult to fund under current funding criteria. The balance of scientific, public and even government opinion may change, however, and further work in these fields may be anticipated.

Fundamental work on earthquake mechanism – still a largely unsolved problem – is also likely to attract attention during the next decades, and to require a combination of physical and statistical modeling.

If these are rather long-term developments, there are many smaller scale, more immediate problems that require further investigation. On the theoretical side, one important issue is the development of improved models and procedures for analyzing data with self-similar characteristics. Another area where further research is needed, particularly in subduction regions, is in the development of better physical and statistical models for deep earthquakes, including their possible links to different forms of activity (seismic, volcanic etc) closer to the surface.

The current interest in developing testing centers for probability forecasting is likely to promote more strenuous efforts to develop improved forecasting models, whether based just on catalogue data, or allied with deformation data, or with data on other precursory phenomena. Even the existing models are capable of producing time-varying forecasts which in principle could lead to significant reductions in earthquake risk. The problem here is how to realize these reductions in practice, for example through improved insurance or disaster mitigation activities. Many questions arise, of an operations research as much as a statistical character, which warrant further study and effort. The probability forecasts themselves, if they are to be useful in such contexts, need to incorporate the uncertainties in the underlying models and hence to take on a more explicitly Bayesian character, and to be presented in such a way that they can be related to the many additional fac-

tors that have to be borne in mind when making real-life decisions.

### Acknowledgments

I am very grateful to friends and colleagues, especially David Harte, Mark Bebbington, David Rhoades and Yehuda Ben-Zion, for helpful discussions, correcting errors and plugging gaps.

### Bibliography

- Ambraseys NN, Melville CP (1982) *A History of Persian Earthquakes*. Cambridge University Press, Cambridge
- Andersen PK, Borgan Ø, Gill RD, Keiding N (1993) *Statistical Models Based on Counting Processes*. Springer, New York
- Bak P, Tang C (1989) Earthquakes as a self-organized critical phenomenon. *J Geophys Res* 94:15635–15637
- Bebbington M, Harte DS (2003) The linked stress release model for spatio-temporal seismicity: formulations, procedures and applications. *Geophys J Int* 154:925–946
- Bebbington M, Vere-Jones D, Zheng X (1990) Percolation theory: a model for earthquake faulting? *Geophys J Int* 100:215–220
- Ben-Zion Y (1996) Stress, Slip and earthquakes in models of complex single-fault systems incorporating brittle and creep deformations. *J Geophys Res* 101:5677–5706
- Ben-Zion Y, Dahmen K, Lyakhovskiy V, Ertas D, Agnon A (1999) Self-driven mode-switching of earthquake activity on a fault system. *Earth Planet. Sci Lett* 172:11–21
- Ben-Zion Y, Eneva M, Liu Y (2003) Large earthquake cycles and intermittent criticality on heterogeneous faults due to evolving stress and seismicity. *J Geophys Res* 108:2307V. doi:10.1029/2002JB002121
- Ben-Zion Y, Lyakhovskiy V (2002) Accelerated seismic release and related aspects of seismicity patterns on earthquake faults. *Pure Appl Geophys* 159:2385–2412
- Ben-Zion Y, Rice J (1995) Slip patterns and earthquake populations along different classes of faults on elastic solids. *J Geophys Res* 100:12959–12983
- Borovkov K, Vere-Jones D (2000) Explicit formulae for stationary distributions of stress release processes. *J Appl Prob* 37:315–321
- Brémaud P, Massoulié L (2001) Hawkes branching processes without ancestors. *J Appl Prob* 38:122–135
- Brillinger DR (1981) *Time Series: Data Analysis and Theory*, 2nd edn. Holden Day, San Francisco
- Burridge R, Knopoff L (1967) Model and theoretical seismicity. *Bull Seismol Soc Am* 57:341–371
- Chelidze TL, Kolesnikov YM (1983) Modelling and forecasting the failure process in the framework of percolation theory. *Izvestiya Earth Phys* 19:347–354
- Chong FS (1983) Time-space-magnitude interdependence of upper crustal earthquakes in the main seismic region of New Zealand. *J Geol Geophys* 26:7–24, New Zealand
- Console R, Lombardi AM, Murru M, Rhoades DA (2003) Båth's Law and the self-similarity of earthquakes. *J Geophys Res* 108(B2):2128V. doi:10.1029/2001JB001651
- Cox DR (1972) Regression models and life tables (with discussion). *Roy J Stat Soc Ser B* 34:187–220
- Dahmen K, Ertas D, Ben-Zion Y (1998) Gutenberg-Richter and characteristic earthquake behavior in simple mean-field models of heterogeneous faults. *Phys Rev E* 58:1494–1501
- Daley DJ, Vere-Jones D (2003) *An Introduction to the Theory of Point Processes*, 2nd edn, vol I. Springer, New York
- Davison C (1938) *Studies on the Periodicity of Earthquakes*. Murthy, London
- Diggle PJ (2003) *Statistical Analysis of Spatial Point Patterns*. 2nd edn. University Press, Oxford
- Ebel JB, Chambers DW, Kafka AL, Baglivo JA (2007) Non-Poissonian earthquake clustering and the hidden Markov model as bases for earthquake forecasting in California. *Seismol Res Lett* 78:57–65
- Evison F, Rhoades D (2001) Model of long-term seismogenesis. *Annali Geofisica* 44:81–93
- Felzer KR, Abercrombie RE, Ekström G (2004) A common origin for aftershocks, foreshocks and multiplets. *Bull Amer Seismol Soc* 94:88–98
- Fisher RL, Dahmen K, Ramanathan S, Ben-Zion Y (1997) Statistics of earthquakes in simple models of heterogeneous faults. *Phys Rev Lett* 97:4885–4888
- Griffiths AA (1924) Theory of rupture. In: *Proceedings 1st Int Congress in Applied Mech, Delft*, pp 55–63
- Gutenberg B, Richter C (1949) *Seismicity of the Earth and Associated Phenomena*, 2nd edn. University Press, Princeton
- Habermann RE (1987) Man-made changes of seismicity rates. *Bull Seismol Soc Am* 77(1):141–159
- Hainzl S, Ogata Y (2005) Detecting fluid signals in seismicity data through statistical earthquake modelling. *J Geophys Res* 110. doi:10.1029/2004JB003247
- Harte D (2001) *Multifractals: Theory and Applications*. Chapman and Hall/CRC, Boca Raton
- Harte D, Li DF, Vreede M, Vere-Jones D (2003) Quantifying the M8 prediction algorithm: reduction to a single critical variable and stability results. *NZ J Geol Geophys* 46:141–152
- Harte D, Li D-F, Vere-Jones D, Vreede M, Wang Q (2007) Quantifying the M8 prediction algorithm II: model, forecast and evaluation. *NZ J Geol Geophys* 50:117–130
- Harte D, Vere-Jones D (2005) The entropy score and its uses in earthquake forecasting. *Pure Appl Geophys* 162:1229–1253
- Hawkes AG (1971) Spectra of some self-exciting and mutually exciting point processes. *Biometrika* 58:83–90
- Hawkes AG, Oakes D (1974) A cluster representation of a self-exciting process. *J Appl Prob* 11:493–503
- Helmstetter A, Sornette D (2002) Subcritical and supercritical regimes in epidemic models of earthquake aftershocks. *J Geophys Res* 107:2237. doi:10.1029/2001JB001580
- Helmstetter A, Sornette D (2003) Båth's law derived from the Gutenberg-Richter law and from aftershock properties. *Geophys Res Lett* 103(20):2069. doi:10.1029/2003GL018186
- Ishimoto M, Iida K (1939) *Bull Earthq Res Inst Univ Tokyo* 17:443–478
- Iwata T, Young RP (2005) Tidal stress/strain and the *b*-values of acoustic emissions at the Underground Research Laboratory. Canada. *Pure Appl Geophys* 162:(6–7):1291–1308. doi:10.1007/s00024-005-2670-2 (P\*1357)
- Jackson DD, Kagan YY (1999) Testable earthquake forecasts for 1999. *Seismol Res Lett* 70:393–403

42. Jaeger JC, Cook NGW (1969) *Fundamentals of Rock Mechanics*. Methuen, London
43. Jaume SC, Bebbington MS (2004) Accelerating seismic moment release from a self-correcting stochastic model. *J Geophys Res* 109:B12301. doi:10.1029/2003JB002867
44. Jeffreys H (1938) Aftershocks and periodicity in earthquakes. *Beitr Geophys* 53:111–139
45. Jeffreys H (1939) *Theory of Probability*, 1st edn (1939), 3rd edn (1961). University Press, Cambridge
46. Jones LM, Molnar P (1979) Some characteristics of foreshocks and their possible relationship to earthquake prediction and premonitory slip on a fault. *J Geophys Res* 84:3596–3608
47. Kagan Y (1973) Statistical methods in the study of the seismic process. *Bull Int Stat Inst* 45(3):437–453
48. Kagan Y (1991) Seismic moment distribution. *Geophys J Int* 106:121–134
49. Kagan Y (1991) Fractal dimension of brittle fracture. *J Non-linear Sci* 1:1–16
50. Kagan Y (1994) Statistics of characteristic earthquakes. *Bull Seismol Soc Am* 83:7–24
51. Kagan Y, Jackson DD (1994) Probabilistic forecasting of earthquakes. *Geophys J Int* 143:438–453
52. Kagan Y, Knopoff L (1977) Earthquake risk prediction as a stochastic process. *Phys Earth Planet Inter* 14:97–108
53. Kagan Y, Knopoff L (1980) Spatial distribution of earthquakes: the two-point correlation function. *Geophys J Roy Astronom Soc* 62:303–320
54. Kagan Y, Knopoff L (1981) Stochastic synthesis of earthquake catalogues. *J Geophys Res* 86:2853–2862
55. Kagan Y, Knopoff L (1987) Statistical short-term earthquake prediction. *Sci* 236:1563–1567
56. Keilis-Borok VI, Kossobokov VG (1990) Premonitory activation of the earthquake flow: algorithm M8. *Phys Earth Planet Inter* 61:73–83
57. Kiremidjian AS, Anagnos T (1984) Stochastic slip predictable models for earthquake occurrences. *Bull Seismol Soc Am* 74:739–755
58. Knopoff L (1971) A stochastic model for the occurrence of main sequence earthquakes. *Rev Geophys Space Phys* 9:175–188
59. Kossobokov VG (1997) User manual for M8. In: *Algorithms for Earthquake Statistics and Prediction*. IASPEI Softw Ser 6:167–221
60. Kossobokov VG (2005) Earthquake prediction: principles, implementation, perspectives. Part I of *Computational Seismology* 36, "Earthquake Prediction and Geodynamic Processes." (In Russian)
61. Kossobokov VG (2006) Testing earthquake prediction methods: The West Pacific short-term forecast of earthquakes with magnitude  $M_wHRV \geq 5.8$ . *Tectonophysics* 413:25–31
62. Libicki E, Ben-Zion Y (2005) Stochastic branching models of fault surfaces and estimated fractal dimensions. *Pure Appl Geophys* 162:1077–1111
63. Lombardi A (2002) Probabilistic interpretation of Båth's law. *Ann Geophys* 45:455–472
64. Lomnitz CA (1974) *Plate Tectonics and Earthquake Risk*. Elsevier, Amsterdam
65. Lomnitz-Adler J (1985) Asperity models and characteristic earthquakes. *J Roy Astron Soc* 83:435–450
66. Lomnitz-Adler J (1985) Automaton models of seismic fracture: constraints imposed by the frequency-magnitude relation. *J Geophys Res* 95:491–501
67. Lomnitz-Adler J (1988) The theoretical seismicity of asperity models; an application to the coast of Oaxaca. *Geophys J* 95:491–501
68. Liu J, Chen Y, Shi Y, Vere-Jones D (1999) Coupled stress release model for time dependent earthquakes. *Pure Appl Geophys* 155:649–667
69. Loève M (1977) *Probability Theory I*, 4th edn. Springer, New York
70. Lu C, Vere-Jones D (2001) Statistical analysis of synthetic earthquake catalogs generated by models with various levels of fault zone disorder. *J Geophys Res* 106:11115–11125
71. Lu C, Harte D, Bebbington M (1999) A linked stress release model for Japanese historical earthquakes: coupling among major seismic regions. *Earth Planet. Science* 51:907–916
72. Macdonald II, Zucchini W (1997) *Hidden Markov and Other Models for Discrete-Valued Time Series*. Chapman and Hall, London
73. Main IG, Burton PW (1984) Information theory and the earthquake frequency-magnitude distribution. *Bull Seismol Soc Am* 74:1409–1426
74. Mandelbrot BB (1977) *Fractals: Form, Chance and Dimension*. Freeman, San Francisco
75. Mandelbrot BB (1989) Multifractal measures, especially for the geophysicist. *Pure Appl Geophys* 131:5–42
76. Martínez VJ, Saar E (2002) *Statistics of the Galaxy Distribution*. Chapman & Hall/CRC, Boca Raton
77. Matsu'ura RS (1986) Precursory quiescence and recovery of aftershock activities before some large aftershocks. *Bull Earthq Res Inst Tokyo* 61:1–65
78. Matsu'ura RS, Karakama I (2005) A point process analysis of the Matushiro earthquake swarm sequence: the effect of water on earthquake occurrence. *Pure Appl Geophys* 162 1319–1345. doi:10.1007/s00024-005-2762-0
79. Matthews MV, Ellsworth WL, Reasenber PA (2002) A Brownian model for recurrent earthquakes. *Bull Seism Soc Amer* 92:2232–2250
80. Merrifield A, Savage MK, Vere-Jones D (2004) Geographical distributions of prospective foreshock probabilities in New Zealand. *J Geol Geophys* 47:327–339, New Zealand
81. Michael A (1997) Test prediction methods: earthquake clustering versus the Poisson model. *Geophys Res Lett* 24:1891–1894
82. Mogi K (1962) Study of elastic shocks caused by the fracture of heterogeneous materials and its relation to earthquake phenomena. *Bull Earthq Res Inst Tokyo Univ* 40:125–173
83. Mogi K (1985) *Earthquake Prediction*. Academic Press, Tokyo
84. Molchan GM (1990) Strategies in strong earthquake prediction. *Phys Earth Plan Int* 61:84–98
85. Molchan GM, Kagan YY (1992) Earthquake prediction and its optimization. *J Geophys Res* 106:4823–4838
86. Ogata Y (1988) Statistical models for earthquake occurrence and residual analysis for point processes. *J Amer Stat Soc* 83:9–27
87. Ogata Y (1998) Space-time point process models for earthquake occurrences. *Annals Inst Stat Math* 50:379–402
88. Ogata Y (1999) Estimating the hazard of rupture using uncertain occurrence times of paleoearthquakes. *J Geophys Res* 104:17995–18014



89. Ogata Y (2005) Detection of anomalous seismicity as a stress change sensor. *J Geophys Res* 110(B5):B05S06. doi:10.1029/2004JB003245
90. Ogata Y, Utsu T, Katsura K (1996) Statistical discrimination of foreshocks from other earthquake clusters. *Geophys J Int* 127:17–30
91. Ogata Y, Jones L, Toda S (2003) When and where the aftershock activity was depressed: Contrasting decay patterns of the proximate large earthquakes in southern California. *J Geophys Res* 108(B6):2318. doi:10.1029/2002JB002009
92. Omori F (1894) On aftershocks of earthquakes. *J Coll Sci Imp Acad Tokyo* 7:111–200
93. Otsuka M (1972) A chain reaction type source model as a tool to interpret the magnitude-frequency relation of earthquakes. *J Phys Earth* 20:35–45
94. Pietavolo A, Rotondi R (2000) Analyzing the interevent time distribution to identify seismicity patterns: a Bayesian non-parametric approach to the multiple change-point problem. *Appl Stat* 49:543–562
95. Pisarenko DV, Pisarenko VF (1995) Statistical estimation of the correlation dimension. *Phys Lett A* 197:31–39
96. Reasenber PA (1999) Foreshock occurrence before large earthquakes. *J Geophys Res* 104:4755–4768
97. Reasenber PA, Jones LM (1989) Earthquake hazard after a mainshock in California. *Sci* 243:1173–1176
98. Reid HF (1911) The elastic-rebound theory of earthquakes. *Bull Dept Geol Univ Calif* 6:413–444
99. Renyi A (1959) On the dimension and entropy of probability distributions. *Acta Math* 10:193–215
100. Rhoades DA (2007) Application of the EEPAS model to forecasting earthquakes of moderate magnitude in Southern California. *Seismol Res Lett* 78:110–115
101. Rhoades DA, Evison FF (2004) Long-range earthquake forecasting with every event a precursor according to scale. *Pure Appl Geophys* 161:147–171
102. Rhoades DA, Evison FF (2005) Test of the EEPAS forecasting model on the Japan earthquake catalogue. *Pure Appl Geophys* 162:1271–1290
103. Rhoades DA, Van Dissen RJ (2003) Estimation of the time-varying hazard of rupture of the Alpine Fault of New Zealand, allowing for uncertainties. *NZ J Geol Geophys* 40:479–488
104. Ripley BD (1988) *Statistical Inference for Spatial Processes*. University Press, Cambridge
105. Robinson R (2000) A test of the precursory accelerating moment release model on some recent New Zealand earthquakes. *Geophys J Int* 140:568–576. doi:10.1046/j.1365-246X.2000.00054.x
106. Robinson R, Benites R (1995) Synthetic seismicity models for the Wellington region of New Zealand: implications for the temporal distribution of large events. *J Geophys Res* 100:18229–18238. doi:10.1029/95JB01569
107. Rundle JB, Klein W, Tiampo K, Gross S (2000) Dynamics of seismicity patterns in systems of earthquake faults. In: *Geocomplexity and the Physics of Earthquakes*. Geophysical Monograph 120, American Geophysical Union
108. Saito M, Kikuchi M, Kudo M (1973) An analytical solution of Go-game model of earthquakes. *Zishin* 26:19–25
109. Scholz CH (1968) The frequency-magnitude relation of microfaulting in rock and its relation to earthquakes. *Bull Seism Soc Am* 58:399–415
110. Scholz CH (1990) *The Mechanics of Earthquakes and Faulting*. Cambridge University Press, New York
111. Schorlemmer D, Gerstenberger MC, Wiemer S, Jackson DD, Rhoades DA (2007) Earthquake likelihood model testing. *Seismol Res Lett* 78:17–29
112. Schuster A (1897) On lunar and solar periodicities of earthquakes. *Proc Roy Soc London* 61:455–465
113. Schwartz DP, Coppersmith K (1984) Fault behavior and characteristic earthquakes: examples from the Wasatch and San Andreas Faults. *J Geophys Res* 89:5681–5698
114. Shi YL, Liu J, Chen Y, Vere-Jones D (1999) Coupled stress release models for time-dependent seismicity. *J Pure Appl Geophys* 155:649–667
115. Shi Y, Liu J, Zhang G (2001) An evaluation of Chinese annual earthquake predictions, 1990–1998. *J Appl Prob* 38A:222–231
116. Shimazaki K, Nakata T (1980) Time-predictable recurrence model for large earthquakes. *Geophys Res Lett* 7:179–282
117. Smith WD (1986) Evidence for precursory changes in the frequency-magnitude *b*-value. *Geophys J Roy Astron Soc* 86:815–838
118. Smith WD (1998) Resolution and significance assessment of precursory changes in mean earthquake magnitude. *Geophys J Int* 135:515–522
119. Stoyan D, Stoyan H (1994) *Fractals, Random Shapes and Point Fields*. Wiley, Chichester
120. Tiampo KF, Rundle JB, Klein W, Ben-Zion Y, McGinnis SA (2004) Using eigenpattern analysis to constrain seasonal signals in Southern California. *Pure Appl Geophys* 16:19–10, 1991 V2003. doi:10.1007/s00024-004-2545-y
121. Turcotte DL (1992) *Fractals and Chaos in Geology and Geophysics*. Cambridge University Press, Cambridge
122. Utsu T (1961) A statistical study on the properties of aftershocks. *Geophys Mag* 30:521–605
123. Utsu T, Ogata Y (1997) *IASPEI Softw Libr* 6:13–94
124. Utsu T, Ogata Y, Matu'ura RS (1995) The centenary of the Omori formula for a decay law of aftershock activity. *J Phys Earth* 43:1–33
125. Vere-Jones D (1969) A note on the statistical interpretation of Båth's law. *Bull Seismol Soc Amer* 59:1535–1541
126. Vere-Jones D (1970) Stochastic models for earthquake occurrence. *J Roy Stat Soc B* 32:1–62
127. Vere-Jones D (1977) Statistical theories for crack propagation. *Pure Appl Geophys* 114:711–726
128. Vere-Jones D (1978) Space-time correlations of microearthquakes – a pilot study. *Adv App Prob* 10:73–87, supplement
129. Vere-Jones D (1978) Earthquake prediction: a statistician's view. *J Phys Earth* 26:129–146
130. Vere-Jones D (1995) Forecasting earthquakes and earthquake risk. *Int J Forecast* 11:503–538
131. Vere-Jones D (1999) On the fractal dimension of point patterns. *Adv Appl Prob* 31:643–663
132. Vere-Jones D (2003) A class of self-similar random measures. *Adv Appl Prob* 37:908–914
133. Vere-Jones D, Davies RB (1966) A statistical analysis of earthquakes in the main seismic region of New Zealand. *J Geol Geophys* 9:251–284
134. Vere-Jones D, Ozaki T (1982) Some examples of statistical inference applied to earthquake data. *Ann Inst Stat Math* 34:189–207

135. Vere-Jones D, Robinson R, Yang W (2001) Remarks on the accelerated moment release model for earthquake forecasting: problems of simulation and estimation. *Geophys J Int* 144:515–531
136. von Bortkiewicz L (1898) *Das Gesetz der kleinen Zahlen*. Teubner, Leipzig
137. Weibull W (1939) A statistical theory of the strength of materials. *Ingvetensk Akad Handl* no 151
138. Working Group on Californian Earthquake Probabilities (1990) Probabilities of earthquakes in the San Francisco Bay region of California. US Geological Survey Circular 153
139. Yin X, Yin C (1994) The precursor of instability for non-linear systems and its application to the case of earthquake prediction – the load-unload response ratio theory. In: Newman WI, Gabrielov AM (eds) *Nonlinear dynamics and Predictability of Natural Phenomena*. AGU Geophysical Monograph 85:55–66
140. Zheng X, Vere-Jones D (1994) Further applications of the stress release model to historical earthquake data. *Tectonophysics* 229:101–121
141. Zhuang J (2000) Statistical modelling of seismicity patterns before and after the 1990 Oct 5 Cape Palliser earthquake, New Zealand. *NZ J Geol Geophys* 43:447–460
142. Zhuang J, Yin X (2000) The random distribution of the loading and unloading response ratio under the assumptions of the Poisson model. *Earthq Res China* 14:38–48
143. Zhuang J, Ogata Y, Vere-Jones D (2004) Analyzing earthquake clustering features by using stochastic reconstruction. *J Geophys Res* 109(B5):B05301. doi:10.1029/2003JB002879
144. Zhuang J, Vere-Jones D, Guan H, Ogata Y, Ma L (2005) Preliminary analysis of precursory information in the observations on the ultra low frequency electric field in the Beijing region. *Pure Appl Geophys* 162:1367–1396. doi:10.1007/s00024-004-2674-3

## Earthquake Scaling Laws

RAUL MADARIAGA

Ecole Normale Supérieure, Laboratoire de Géologie,  
Paris, France

### Article Outline

Glossary

Definition of the Subject

Introduction

Earthquakes and Seismic Radiation

Earthquake Fault Models:

The Scaling of Geometry and Stress

Earthquake Dynamics and the Scaling of Energy

Kinematics and Statistical Models for Fault Slip

Future Directions

Acknowledgments

Bibliography

### Glossary

**Seismic moment** The most fundamental measure of the size of an earthquake. In the simplest situation it represents the moment of one of the couples of forces that make up a dipolar source. In more general cases it is a 3 by 3 symmetric tensor of elementary force couples.

**Seismic radiation** The seismic waves emitted by a seismic source. For point sources these are spherical P and S waves emitted by the point tensor source.

**Seismic spectrum** The absolute value of the Fourier transform of the displacement field radiated by an earthquake in the far field. For almost all earthquakes it has a common shape: flat at low frequencies and decays like the inverse squared power at high very high values of frequency.

**Corner frequency** The low and high frequency asymptotes of the earthquake spectrum intersect at a characteristic frequency, called the corner frequency. The corner frequency scales with the size of the earthquake measured by the seismic moment.

**Radiated or seismic energy** Total energy of the seismic waves radiated by a seismic source. It can be computed from the energy flow relatively far from the source of the earthquake.

**Apparent stress** Originally defined as the product of seismic efficiency times the average stress during earthquake slip. In practice, it is computed from the ratio of radiated energy to moment release of the earthquake multiplied by the shear modulus.

**Energy release rate** Amount of energy per unit surface used to make a rupture advance by a unit distance.

**Static stress drop** The static change in shear traction between the sides of the fault occurs during an earthquake. In principle, it could be determined by measuring stress before and after the earthquake. In practice stress drop is computed using very specific source models, like a circular crack.

**Dynamic stress drop** The stress change in shear traction as a function of time while the rupture is still growing. It can only be estimated from seismic records obtained in the near field by elaborate inversion schemes. The relation between static and dynamic stress drop can only be estimated once the friction law between the sides of the fault has been defined.

### Definition of the Subject

Earthquake scaling laws provide some of the most basic knowledge about seismic sources. Since the end of the 70s, a very successful model for earthquakes was developed by seismologists. In this model earthquakes are due to rapid slip on pre-existing faults driven by steady load due to plate motion and resisted by friction between the fault walls. This model may be used to predict many of the general properties of seismic radiation that can be derived from a simple spectral shape of type omega-squared. In this article I derive general expressions for energy, moment and stress in terms of measured spectral parameters. The available data shows that earthquakes can be reduced to a single family in terms of three parameters: moment, corner frequency and radiated energy. Using specific models of rupture these three parameters can be reinterpreted in terms of moment, size and stress drop. Although details differ between the models proposed by seismologists, both seismic spectra and the wave-number spectra of slip distributions can be explained with a simple circular crack model. This does not mean that a circular crack is the best earthquake model; it means that the ensemble average of seismic sources has properties that are similar to those of simple circular shear cracks. A direct result of scaling laws is that total fracture energy must scale like radiated and strain release energy, so that fracture energy should scale with fault size as observed for many earthquakes and in certain laboratory experiments.

### Introduction

It has been 40 years since Aki [4] published his seminal paper on the scaling law of earthquakes that established that to first order seismic moment scales like the third power of the fault size. This paper came just a year after Aki [3]

made the very first measurement of seismic moment, the torque of one of the couples that make up a basic source mechanism. Not much later, in 1970, Brune introduced a very simple source model and established a generalization of Aki's [4] spectral model of earthquakes that is now known as the omega-squared model of earthquakes. Almost simultaneously, Kostrov [43], Savage [66], Sato and Hirasawa [65] and Madariaga [46] developed models of a circular faults. Radiation from a circular crack explained well the omega-squared model. Digital data was not available at the time when these models were introduced. It is nowadays an almost standard observatory practice to study the scaling law, but as many recent studies have shown, proper estimation of spectral parameters like moment and corner frequency and high frequency decay for a large range of earthquake sizes is not trivial and it is often difficult to obtain from a single instrument. In the 1990s good quality digital data, both broad band seismograms and accelerograms, became available not just from surface instruments, but also from boreholes opening the way to a new appraisal of the scaling law.

We will briefly review some of these observations and we will try to establish general properties of earthquake radiation based on the recent work by McGarr [53], Abercrombie and Rice [2], Ide et al. [35], Prieto et al. [59], etc. Our purpose is not to review the very extensive literature on the determination of seismic source parameters; most of those papers assume specific scaling laws like circular cracks, Brune's relation between corner frequency and earthquake size, etc. Our purpose here is to derive the scaling law from same basic physical concepts and to test the validity of some common assumptions in seismology. Because different authors are interested in certain specific aspects of earthquake sources, often the data from different authors is hard to combine. Some authors use the Brune's [17] empirical model as a basis for the scaling law; others use the radiation from quasidynamic circular cracks as a model leading to substantial variations. Even more serious differences come from data processing, some authors using specific attenuation corrections in order to correct for Earth's Q, others use small events as empirical Green functions, etc. These corrections are very important, but they are beyond our goal which is to try to extract information about scaling and its inferences for earthquake physics. Two aspects will be particularly discussed: how to make model independent estimates of source parameters and how to establish model independent scaling laws. Some recent evidence indicates, for instance, that stress change during earthquakes as measured by apparent stress varies independently of Moment and size, so that earthquakes are probably quantified by three independent

parameters that need to be carefully chosen. The other will be an aspect that is often overlooked in the literature: Brune's [17] model, as well as Madariaga's [46] circular crack model make very specific statements about the partition of seismic energy at the source. In particular, the scaling law implies that fracture energy is not a constant but that it scales with earthquake size in a manner that was predicted on the basis of fracture dynamics [46]; friction experiments by Ohnaka and Shen [56] and Ohnaka [55]; arguments about scaling by McGarr and Fletcher [54], and dynamic seismic source inversions [34,57].

### Earthquakes and Seismic Radiation

It is now well established that earthquakes are due to faulting and that the simplest way to measure them is to use the seismic moment  $M_0$ , introduced by Aki [3], and given by

$$M_0 = \mu \bar{D} S \quad (1)$$

where  $\mu$  is the shear or rigidity modulus of the material surrounding the fault,  $\bar{D}$  is the mean value of the final slip on the fault and  $S$  is the area of the fault rupture.  $M_0$  is the moment of one of the couples that constitute a double couple, the simplest possible model of a point-like earthquake. Radiation from a point double couple source has been completely solved by seismologists and is thus the natural starting point for the development of earthquake scaling laws.

The far-field displacement  $\mathbf{u}_c$  radiated by a point double couple source can be written in the following form:

$$\mathbf{u}_c(r, t) = \frac{1}{4\pi\rho c^3} \frac{1}{R} \mathbf{e}_c^T \cdot \mathbf{M}_0 \cdot \mathbf{e}_R(t - R/c), \quad (2)$$

(see, e. g. [5]) where the subscript  $c$  stands for P or S waves, i. e.  $c$  is either  $\alpha$ , the P wave speed, or  $\beta$ , the S wave speed;  $\rho$  is the density;  $R$  is the distance of the observation point from the source.  $\mathbf{e}_R$  is the radial unit vector in the direction of radiation. The unit vector  $\mathbf{e}_c$  is the polarization of the wave, that is  $\mathbf{e}_c = \mathbf{e}_R$  for P waves, or  $\mathbf{e}_T$  the appropriate transverse unit vector for SH or SV waves.  $\mathbf{M}_0(t)$  is the moment tensor of the source, a symmetric tensor of order three that describes the geometry and amplitude of the seismic source (see [5] for details).

Very often in seismology it is assumed that the geometry of the source can be separated from its time variation, so that the moment tensor can be written in the simpler form:

$$\mathbf{M}_0(t) = \mathbf{I}_0 M_0 s(t) \quad (3)$$

where  $\mathbf{I}_0$  is a time-invariant tensor that describes the orientation of the source,  $M_0$  is the scalar moment tensor of

the source, and  $s(t)$  is the time variation of the moment, the source time function determined by seismologists. In the following we assume that  $s(t)$  is causal, i. e. it is zero up to  $t = 0$ , and normalized so that

$$\int_0^{\infty} s(t) dt = 1 .$$

Using (3) we can now write a simpler form of (2):

$$u_c(r, t) = \frac{1}{4\pi\rho c^3} \frac{\mathcal{R}_c}{R} \Omega(t - R/c) . \quad (4)$$

For P waves,  $u_c$  is the radial component; for S waves, it is the appropriate transverse component for SH or SV waves. In (4) we have introduced the standard notation  $\Omega(t) = M_0 ds(t)/dt$  for the source time function, the signal emitted by the source as seen in the far field. The term  $\mathcal{R}_c(\theta, \phi)$  is the radiation pattern, a function of the takeoff direction of the ray from the source. In a spherical coordinate system  $(R, \theta, \phi)$  centered at the source, the radiation patterns are given by Aki and Richards [5]. We simply quote here the case of a so-called strike-slip earthquake where the fault is vertical, normal to the  $y$  axis, and slip is parallel to the  $x$  axis, so that only the  $M_{xy} = M_{yx}$  components of the moment tensor are different from zero. In this case

$$\begin{aligned} \mathcal{R}_P &= \sin^2 \theta \sin 2\phi , \\ \mathcal{R}_{SV} &= \frac{1}{2} \sin 2\theta \sin 2\phi , \\ \mathcal{R}_{SH} &= \sin \theta \cos 2\phi . \end{aligned}$$

On the  $z = 0$  plane ( $\theta = \pi/2$ ), there are no SV waves. On the other hand, on this plane, the radiation patterns of P and SH waves have typical quadrupole distributions proportional to  $\sin 2\phi$  and  $\cos 2\phi$  respectively.

### Spectral Domain Approach

At high frequencies, the signals radiated by earthquakes may become quite complex because of multipathing, scattering, etc., so that the actually observed seismogram  $u(t)$  resembles the source time function  $\Omega(t)$  only at long periods. It is usually verified that complexities in the wave propagation affect much more the phase of seismic waves than the spectral amplitudes in the Fourier transformed domain. The spectral domain approach was introduced by Aki [4], Wyss and Brune [73] and Brune [17]. Radiation from a simple point moment-tensor source can be obtained from (4) by Fourier transformation. Displacement pulses radiated from a point moment tensor in the Fourier transformed domain is then

$$u_c(r, \omega) = \frac{1}{4\pi\rho c^3} \frac{\mathcal{R}_c}{R} \tilde{\Omega}(\omega) e^{-i\omega R/c} . \quad (5)$$

Where  $\tilde{\Omega}(\omega)$  is the Fourier transform of the source time function  $\Omega(t)$ . A well-known property of the Fourier transform is that

$$\lim_{\omega \rightarrow 0} \tilde{\Omega}(\omega) = M_0 * \int_0^{\infty} \dot{s}(t) dt = M_0 \quad (6)$$

so that the in the low-frequency limit of the source time function spectrum approaches the scalar moment.

From the observation of many earthquake spectra, and from the computation of magnitudes in different frequency bands, Aki [4] and Brune [17] concluded that the seismic spectra had a universal shape with a flat low frequency asymptote given by (6), a certain characteristic frequency called **corner frequency** by seismologists, and a decay at high frequencies that tends asymptotically to  $\omega^{-2}$ . I will not repeat here the arguments that led to this model: Aki [4] proposed this spectral shape in order to explain the differences in magnitude determined from seismic waves of different frequencies. Brune's [17] argument was based on simple concepts about the high frequency radiation and the amount of energy radiated by a seismic event. The omega-squared model has now been confirmed by numerous observations and detailed studies of seismic radiation. As an example, Fig. 1 shows the displacement record of a M5 earthquake that occurred inside the Nazca plate, 99 km under Santiago de Chile on 7 January 2003. The spectrum shown at the bottom presents the typical low frequency asymptote proportional to the seismic moment, and high frequency decay proportional to  $\omega^{-2}$ . The corner frequency is approximately 1.2 Hz for this event.

In its simpler form the  $\omega^{-2}$  spectrum is [17]:

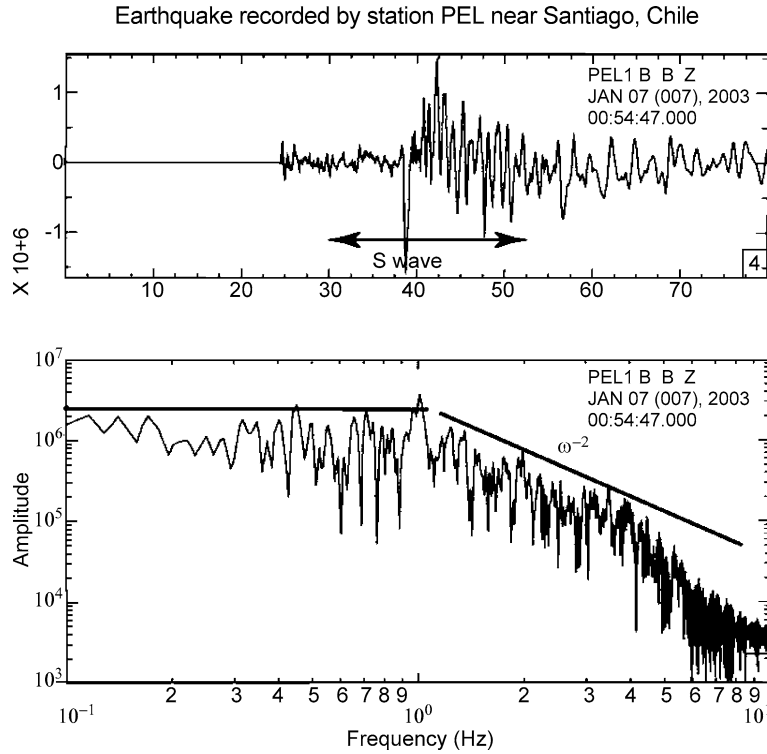
$$\tilde{\Omega}(\omega) = \frac{M_0}{1 + \omega^2/\omega_0^2} \quad (7)$$

where  $\omega_0$  is the corner frequency. Based on considerations about the spectrum of random signals, Boatwright [12] proposed the alternative model:

$$\tilde{\Omega}(\omega) = \frac{M_0}{[1 + \omega^4/\omega_0^4]^{1/2}} . \quad (8)$$

In these simple omega-squared models, seismic sources are characterized by only two independent scalar parameters: the seismic moment  $M_0$  and the corner frequency  $\omega_0$ .

Brune's model (7) explains well the spectrum of S waves, it describes also that of P waves, but the corner frequencies are different for P and S. For a long time observations were not able to distinguish between the P and S waves spectra. The first clear observations of the corner frequency ratio were made by Hanks [26]. Recent work by Abercrombie and Rice [2], Prieto et al. [59], Ide et al. [35]



Earthquake Scaling Laws, Figure 1

Example of a seismic signal and its spectrum. Recording at the PEL broad-band station of the GEOSCOPE network of a  $M = 5$  intermediate depth earthquake. The event occurred inside the subducted Nazca plate under Santiago, Chile at 99 km depth. The top panel shows the displacement, integrated from the broad band velocity record. The shear wave section used for computing the spectrum is shown with arrows. The bottom panel shows the amplitude Fourier spectrum. The low frequency and high frequency trends are indicated by the thick straight lines. The corner frequency is located at the intersection of the two asymptotes

has confirmed that P spectra have a higher corner frequency than S waves and that very roughly

$$\omega_0^P \cong 1.6\omega_0^S. \quad (9)$$

This is very close to  $\sqrt{3}$ , the ratio of P to S wave speed. This ratio is similar to that predicted for the quasi-dynamic circular crack model [46].

As mentioned earlier, not all earthquakes have displacement spectra as simple as (7), but the omega-squared model is a simple starting point for understanding seismic radiation.

From (7), it is possible to compute the spectra predicted for ground velocity:

$$\tilde{\dot{\Omega}}(\omega) = \frac{i\omega M_0}{1 + \omega^2/\omega_0^2}. \quad (10)$$

Ground velocity spectra are characterized by a peak situated roughly at the corner frequency  $\omega_0$ . In actual earthquake ground velocity spectra, this peak is usually

broadened and contains oscillations and secondary peaks, but (10) is a good approximation to the spectra of ground velocity for frequencies lower than a certain cut-off frequency called  $f_{\max}$  by Hanks [27], and is close to 6–7 Hz in many areas. At frequencies higher than  $f_{\max}$ , attenuation, propagation and scattering modify the velocity spectrum.

### Seismic Energy Radiated by Point Moment-Tensor Sources

In order to establish the most basic scaling relationship for seismic sources we have to compute the energy radiated by an omega-squared source like (8). This was actually the way Brune originally calibrated his source time function and Aki established the variation of corner frequency with moment. Assuming that the source is embedded in a homogeneous medium, and that the observation point is sufficiently far from the source, the energy flow per unit solid angle,  $e_r$  is proportional to the square of the particle velocity  $v_c$  (see [11,12,17,18,31]), so that the total flow per unit

solid angle is:

$$e_r^c = \rho c R^2 \int_0^\infty v_c^2(t) dt \quad (11)$$

where  $\rho c$  is the P or S wave impedance,  $v_c(t)$  is the ground velocity and  $R$  is again the distance of the observation point from the source. We can compute the radiated energy density replacing  $v_c(t)$  by the time derivative of the far field displacement (4), and get:

$$e_r^c = \frac{1}{16\pi^2 \rho c^5} \mathcal{R}_c^2 \int_0^\infty \dot{\Omega}^2(t) dt. \quad (12)$$

As expected the energy flow per unit solid angle does not depend on the distance from the source  $R$ . We can now apply Parseval's theorem to express the energy in terms of the source spectral amplitude

$$\int_0^\infty \dot{\Omega}^2(t) dt = \frac{1}{\pi} \int_0^\infty \omega^2 |\tilde{\Omega}(\omega)|^2 d\omega$$

and compute the total radiated energy,  $E_r$ , for each type of wave. Integrating (12) over the angles  $\theta$  and  $\phi$  we get

$$E_r^c = \frac{1}{4\pi^2 \rho c^5} \langle \mathcal{R}_c^2 \rangle \int_0^\infty \omega^2 |\tilde{\Omega}(\omega)|^2 d\omega \quad (13)$$

where

$$\langle \mathcal{R}_c^2 \rangle = \frac{1}{4\pi} \iint_{\Omega} \mathcal{R}_c^2(\theta, \phi) \sin \theta d\theta d\phi$$

is the mean-squared radiation pattern. This expression for the total radiated energy is very interesting because it does not depend on any assumption about earthquake dynamics, just on the shape of the spectra.

For the  $\omega$ -squared model (7) the integral over circular frequency in (13) can be evaluated exactly to  $(\pi/4M_0^2\omega_0^3)$ , so that the radiated energy is simply

$$E_r^c = \frac{1}{16\pi} \langle \mathcal{R}_c^2 \rangle \frac{M_0^2 \omega_0^3}{\rho c^5} \quad (14)$$

where we grouped in the last term all the dimensional variables. Let us remark that the numerical factor  $1/16\pi$  depends on the particular model assumed for the spectrum near the corner frequency. Thus, for the Boatwright model (8), the coefficient is slightly larger ( $\sqrt{2}/16\pi$ ).

Since radiated energy and moment have the same dimensional units it is customary to rewrite this expression in the non-dimensional form:

$$\frac{E_r^c}{M_0} = \frac{\langle \mathcal{R}_c \rangle^2}{16\pi} \frac{M_0}{\rho} \frac{\omega_0^3}{c^5}. \quad (15)$$

Very often this expression is written in terms of frequency  $f_0$  instead of the circular frequency ( $\omega_0 = 2\pi f_0$ ), so that

$$\frac{E_r^S}{M_0} = \frac{\pi^2 \langle \mathcal{R}_c \rangle^2}{2} \frac{M_0}{\rho} \frac{f_0^3}{c^5}. \quad (16)$$

The average radiation patterns are well known,  $\langle R_p \rangle^2 = 4/15$  and  $\langle R_s \rangle^2 = 6/15$ , see, e.g. Haskell [31], so that for S waves most authors use the following expression to quantify the ratio between the S wave radiated energy and the seismic moment

$$\frac{E_r^S}{M_0} = 1.9739 \frac{M_0}{\mu} \frac{f_0^3}{\beta^3}. \quad (17)$$

Where we used the definition of S wave speed ( $\mu = \rho\beta^2$ ). This non-dimensional relation makes no assumptions about the rupture process at the source except that the spectrum decays like  $\omega^{-2}$  at high frequencies. Note that the numerical coefficient is smaller by a factor of four from that computed by Singh and Ordaz [69]. The factor of four seems to be a misprint.

### Apparent Stress

A very important parameter of seismic sources that can be computed independently of any particular source geometry is the apparent stress. Originally, apparent stress was defined as the product of the seismic efficiency  $\eta$  by the average stress  $\bar{\sigma}$  that acts across the fault during the earthquake

$$\sigma_a = \eta \bar{\sigma},$$

efficiency  $\eta$  was in turn defined as the ratio between the radiated energy  $E_r$  and the total released energy  $W$ . Unfortunately, neither  $W$  nor the average stress can be directly inverted from seismic observations because seismic waves have no information about the average stress that acts on the fault. Wyss and Brune [73] proved that for uniform average stress,  $W$  could be written as

$$W = \frac{\bar{\sigma}}{\mu} M_0$$

so that apparent stress can be defined as

$$\sigma_a = \frac{\mu E_r^S}{M_0}.$$

This expression, originally derived for uniform average stress has become one of the most useful measures of stress

on seismic sources (see, e. g. [53]). Using the expression for radiated energy (17) we get:

$$\sigma_a = \frac{\mu E_r^S}{M_0} = 1.9739 M_0 \frac{f_0^3}{\beta^3} \tag{18}$$

an expression that depends only on three measurable quantities: total S wave radiated energy, seismic moment and corner frequency. Because the energy flow can usually be computed only for those directions where stations are available, (18) can never be evaluated very accurately. This problem still persists at present; in spite of the deployment of increasingly denser instrumental networks, there will always be large areas of the focal sphere that remain outside the domain of seismic observations because the waves in those directions are refracted away from the station networks, energy is dissipated due to attenuation, etc.

Equation (18) shows that energy moment ratio is a non-dimensional number that depends only on observable quantities ( $E_r$ ,  $M_0$  and  $f_0$ ) and wave speed. No assumption is made in Eq. (18) about particular models to convert corner frequencies into source dimensions. Let us finally remark that if  $\sigma_a$  is computed from the estimated radiated energy  $E_r$ , then the apparent stress may be considered as an independent parameter that may be used to test the relation (18).

### Time Domain Approach

In the previous section we approached seismic radiation from the spectral point of view. An alternative way to understand radiation is to approach it from the time domain. The obvious question is what is the time domain signal associated with  $\omega^{-2}$  spectrum (17)? Brune [17] proposed one of them:

$$\Omega(t) = M_0 \omega_0^2 t e^{-\omega_0 t} \quad \text{for } t > 0 \tag{19}$$

which is a causal function (i. e., it is zero for  $t < 0$ ). It is also normalized so that

$$\int_0^\infty \Omega(t) dt = M_0.$$

The high frequency content for this function is controlled by the slope discontinuity at the origin. This is however only one of many functions sharing the same spectral amplitude described by (7) and (8), leaving sufficient freedom for variations in the source time function shape.

Kanamori and Rivera [40] defined a function of finite duration  $T$ , finite moment and minimum radiated energy. They found such signal by solving a variational problem

for fixed signal duration  $T$ . The time signal is the parabola defined by

$$\Omega(t) = \frac{6M_0}{T^3} t(T-t) \tag{20}$$

for  $0 < t < T$  and 0 elsewhere. The radiated energy for this signal can be computed using the time domain expression (12). Using the integral

$$I_v = \int_0^\infty \dot{\Omega}^2(t) dt = \frac{12}{T^3} M_0^2$$

we can compute the total radiated energy as a function of  $T$ . In order to write the energy in the same form as the frequency domain expressions we determine the corner frequency from the Fourier transform of (20):

$$\tilde{\Omega}(\omega) = \frac{6M_0}{\omega^3 T^3} \left[ (\omega T - 2i) + (\omega T + 2i)e^{i\omega T} \right]. \tag{21}$$

At low frequencies this expression tends to  $M_0$ , as expected, while at high frequencies it behaves like

$$\lim_{\omega \rightarrow \infty} \tilde{\Omega}(\omega) = \frac{12M_0}{\omega^2 T^2} \frac{[1 + e^{i\omega T}]}{2}.$$

So that its envelope decreases asymptotically as  $12M_0(\omega T)^{-2}$  that is, it has the same inverse omega-squared behavior as Brune's model (7). The corner frequency for this signal computed from the intersection of the asymptotes is  $\omega_0 = \sqrt{12}/T$ . Inserting into (17) we get the following energy moment ratio:

$$\frac{E_r^c}{M_0} = \frac{\langle \mathcal{R}_c \rangle^2}{4\sqrt{12}\pi} \frac{M_0}{\rho} \frac{\omega_0^3}{c^5}. \tag{22}$$

which has a numerical coefficient that is slightly larger than that of (17). This is apparent contradiction to the method used by Kanamori and Rivera, who derived (20) from the condition that  $E_r$  be a minimum for a given moment. The reason these two results are not contradictory is that Brune's signal has infinite duration and therefore the variational principle posed by Kanamori and Rivera does not apply to it. Thus the two expressions give very similar answers, but clearly the signal (20) is not the only one that minimizes radiated energy. It is interesting to observe that all these signals decay like  $\omega^{-2}$  at high frequencies. The reason is that high frequency radiation in these models is controlled by the slope discontinuities in displacement. In Brune's signal (19) the slope discontinuity is at the origin, while in (20) there are two slope discontinuities at the origin and at  $t = T$ .



In conclusion, the radiation models proposed by seismologists share the following properties: (1) the amplitude is controlled by the seismic moment, (2) the displacement spectrum decreases like  $\omega^{-2}$  at high frequencies, (3) the spectral shape has a corner frequency  $f_0$  and (4) the radiated energy to moment ratio satisfies the relation

$$\frac{E_r^c}{M_0} = C_r \frac{M_0 f_0^3}{\rho c^5}, \tag{23}$$

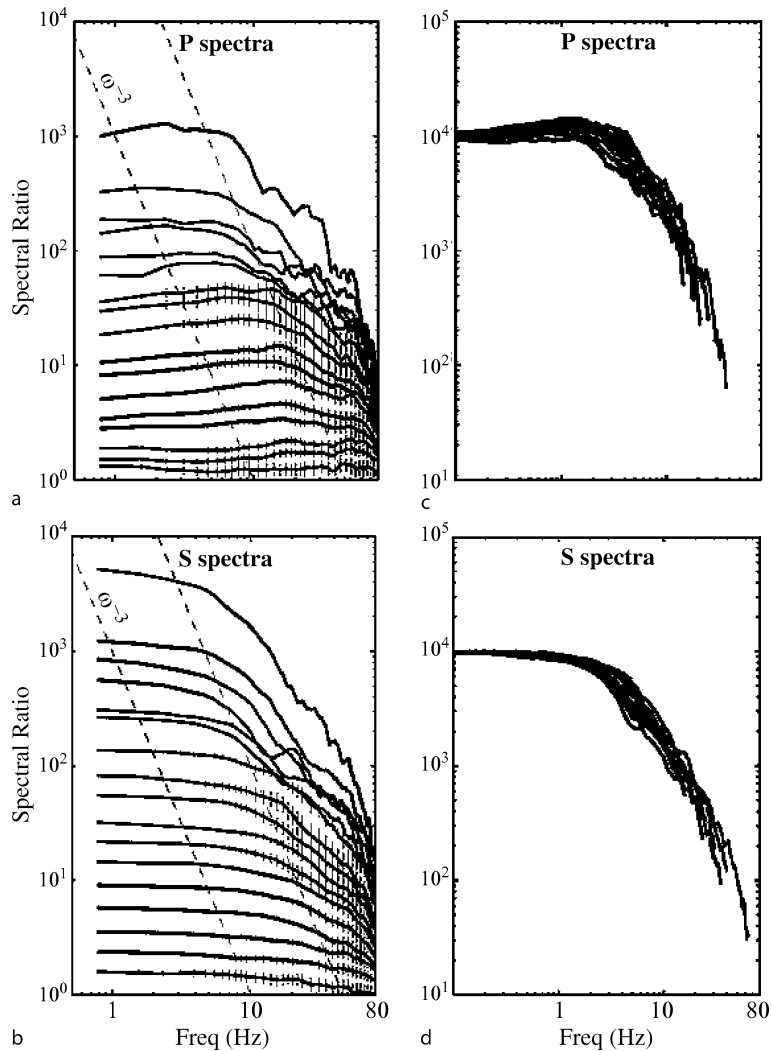
where  $C_r$  is a numerical constant on the order of two. Its consequences for energy balance are quite interesting as we shall promptly discuss.

**Aki's Scaling Law**

From observation of seismic data most authors (for recent data see, e. g., Abercrombie [1], Ide and Beroza [33] Ide et al. [35], McGarr [53]). Abercrombie and Rice [2], Prieto et al. [68] have concluded that apparent stress  $\sigma_a$  is almost independent of moment for most earthquakes. If that is correct, its immediate consequence is that moment scales like the inverse third power of the corner frequency:

$$M_0 \propto f_0^{-3}. \tag{24}$$

*If apparent stress is constant, seismic moment is inversely proportional to the cube of the corner frequency. This result*



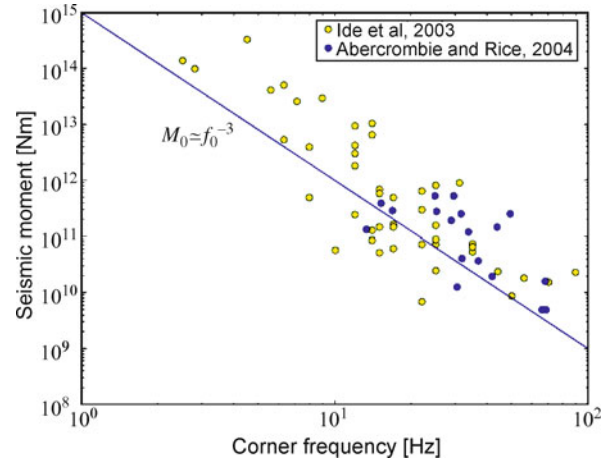
Earthquake Scaling Laws, Figure 2  
 Scaling of the displacement spectrum in the far field as illustrated by several spectra for different size earthquakes reported by Prieto et al. [59]. Body wave spectra have the typical Brune [17] spectrum and can be collapsed into a single scaling figure by gliding the corner frequencies along an  $\omega^{-3}$  line

is independent of the particular source model used and can be tested directly from seismic observations. Figure 2 from Prieto et al. [59] illustrates this scaling law. On the left-hand side a series of spectra for different size earthquakes are shown. The corner frequencies align along a line with slope  $\omega^{-3}$ . Letting all the spectra glide along this line they computed the spectra stack shown at the right-hand side. The same properties are shared by P and S wave spectra. The P wave corner frequencies are higher by a factor of 1.6 than, those of shear waves (9).

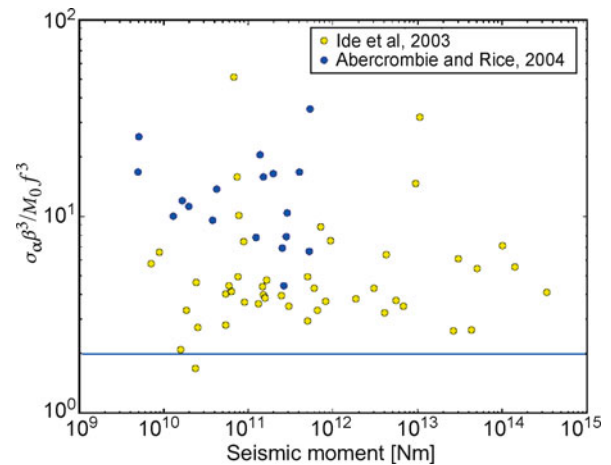
Originally, the scaling law (24) was proposed by Aki [4] following a study of spectral data from several collocated earthquakes of different magnitude. Comparing their spectra, he concluded that corner frequencies scaled with seismic moment like (24). This is the so-called **scaling law of seismic spectrum**. This law has been tested by numerous authors with increasingly reliable digital data. Figure 3 shows an example derived from data published by Ide et al. [35] and Abercrombie and Rice [2]. In this Figure I plotted the corner frequency of S waves as measured by the authors as a function of seismic moment. The moment vs. corner frequency plot clearly follows the trend of Eq. (24). The fact that in Fig. 3 moment and corner frequency scale like (24) is often taken as a proof that earthquakes scale with a single parameter: the seismic moment. This is however not sufficient to prove scaling because seismic sources require at least three independent parameters for their quantification. It has been traditional to add an additional model dependent relation in order to derive length, time and stresses from (23). The most common assumption is that the corner frequency is related to the radius of an equivalent circular crack using the frequency radius relation proposed by Brune [17]. Some authors use other similar relations derived from quasidynamic models (e. g. [46,66]). With that assumption, corner frequency can be converted into fault size and stress drop can be derived from moment and source radius.

In order to test whether earthquakes scale with a single parameter, it is necessary to obtain an additional objective measure of seismic sources. The best current candidate is the radiated energy of S waves defined by (13). Originally proposed by Boatwright [11,12,13], estimates of seismic energy have become quite common but they are still difficult to obtain as discussed by many authors including Boatwright and Fletcher [15], Abercrombie [1], McGarr [53], Ide et al. [32] Singh and Ordaz [69], etc. In the following I will test the scaling law of earthquakes by computing the non dimensional ratio  $C_r = \mu E_r \beta^3 / (M_0 f_0^3)$  from published data.

In Fig. 4 I test expression (18) for the data studied by Ide et al. [35] and Abercrombie and Rice [2]. The fig-

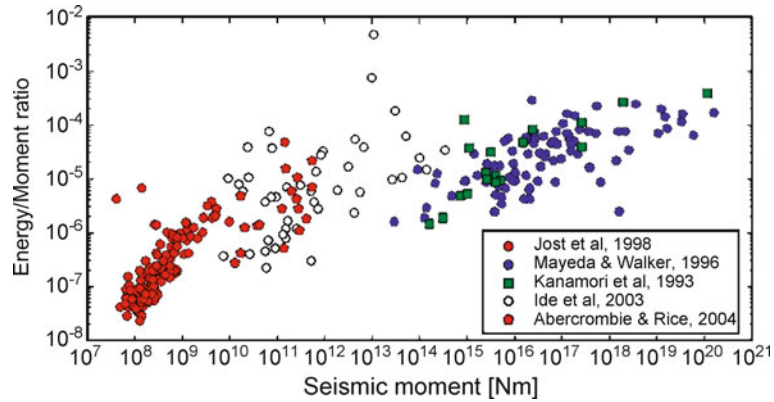


Earthquake Scaling Laws, Figure 3  
 Self-similarity of earthquake spectra. The figure includes data from two different studies by Ide et al. [32] and by Abercrombie and Rice [2]. These authors measured all the quantities needed for testing the energy/moment scaling relation (23). The line labeled  $M_0 \cong f_0^{-3}$  indicates the trend of variation of moment with corner frequency predicted by Aki's scaling law



Earthquake Scaling Laws, Figure 4  
 The nondimensional coefficient  $C_r$  defined by expression (23) plotted as a function of seismic moment. Data from Abercrombie [2] and Ide et al. [32]. The non dimensional coefficient predicted by Brune's spectral model should be equal to 1.9739 as indicated by the horizontal line. Actually it varies over almost two orders of magnitude in this data set. This may be due to errors in the measurement of radiated energy, moment and corner frequencies or it may reflect an important departure from single parameter scaling of seismic sources

ure plots the non dimensional ratio  $C_r = \mu E_r \beta^3 / M_0 f_0^3$  as a function of seismic moment. According to (23), for strict scaling of seismic sources with a single parameter, this non-dimensional ratio should be a constant, indepen-



Earthquake Scaling Laws, Figure 5

Energy moment ratio as a function of moment for 14 orders of magnitude of moment. This figure is inspired on similar plots by McGarr [53] and Ide and Beroza [33]. We used data from Jost et al. [36], Mayeda and Walker [52], Kanamori et al. [41], Ide et al. [35] and Abercrombie and Rice [2]. The data of Jost et al. [36] was not corrected for attenuation as suggested by Ide and Beroza [33]. At this scale it is obvious that moment is the most broadly variable parameter, but there is also a large spread of apparent stress, over at least three orders of magnitude

dent of the seismic moment. Figure 4 shows that there is substantial scattering of the values of the non-dimensional ratio. There are many reasons for this variation; the most obvious one is experimental error due to uncorrected path or site effects, like attenuation, scattering and site amplification. I believe that even if those errors were corrected there would remain some variation due to source complexity that is not fully explained by the assumption that all earthquakes scale with a single parameter. In the literature authors generally assume that the scaling law applies and proceed to compute model-dependent quantities like static or dynamic stress drop using very specific models of rupture (see also p. 36 in [10]).

Clearly moment, energy/moment ratio and apparent stress are broadly distributed. This is illustrated in Fig. 5, inspired by previous figures of the same kind by McGarr [53] and Ide and Beroza [33]. The main difference between the results reported by these two authors is that Ide and Beroza [33] introduced a correction for attenuation at high frequencies. The figure shows the energy/moment ratio as a function of moment over 15 orders of magnitude of seismic moment. The sources of data [2,35,36,41,52] are not the same as those used by McGarr [53] and Ide and Beroza [33], because not all the data they used is published. I did not include small mine earthquakes from South Africa because it is not clear whether those events are due to frictional slip of pre-existing surfaces (see, e.g. [63]). For such a broad range of moments, the variation in energy-moment ratio is bounded, but it ranges over close to three orders of magnitude. It is clear that for the group of earthquakes reported in Fig. 5, apparent stress changes

were important and deserve further work. This is not completely unexpected: a multitude of observations point out that stresses in the seismic zones are highly variable as well as the geometry of faulting.

### Earthquake Fault Models: The Scaling of Geometry and Stress

The previous discussion focused on the properties of seismic radiation from moment tensor sources and the time and frequency dependence of the moment rate function. Actually, the Brune spectrum and Aki's scaling law can be retrieved from seismic waves without any reference to a particular fault model. In order to understand how the moment is related to source dimensions and the origin of omega-squared radiation, we have to introduce a specific fault model. We will proceed in two steps: first we will study a simple source model that explains most of the observations and, in a second step, we will discuss how this model can be generalized.

A fault is defined as a rupture in the earth crust with a relative displacement of its two sides. The relative displacement between the two sides of the fault (or fault slip) will be denoted by  $\mathbf{D}(\mathbf{x}, t)$ , a vector function of position on the fault ( $\mathbf{x}$ ) and time ( $t$ ). Thus, in general,  $\mathbf{D}$  may vary in amplitude and direction over the fault plane and at each point is a function of time. The scalar seismic moment of an earthquake defined in (1) depends on source area and slip on the fault. It is essentially a static measurement of earthquake size. Corner frequency, on the other hand, is a measure of the duration of the earthquake signal which

is controlled by the time it takes for the rupture front to propagate across the fault. Thus, corner frequencies depend clearly on fault size, but one can expect this relation to be complex and very dependent on the details of the rupture process. This is often ignored in practical work, and simple source models are adopted in order to express the scaling law (23) in terms of source dimension. The argument in favor of this approach is that in order to build scaling laws that extend over several orders of magnitude of moment we may ignore the details of slip and geometry. Let us start by studying a simple circular crack, probably the simplest realistic fault model one can consider.

In a simplified model of fracture, the relative slip  $D$  of the two sides of a fault is produced by the relaxation of the shear stress transmitted across the fault. Shear stress changes with time due to the slow motion of plates, orogeny and a number of other processes that transfer stresses in the Earth's crust. When shear stress exceeds the strength of the material or the friction that maintains the fault locked, slip on the fault starts and, simultaneously, shear stress relaxes to a lower value until all motion on the fault ceases. This process is obviously very complex; detailed studies of even the simplest models of faulting show that stress relaxation at any point on the fault is complex function of not just local stress release on the fault but of other slipping points and of wave propagation on the fault. Solving such a complex problem is only worth the effort for very special, very well recorded events. For most other earthquakes we want to make global estimates of stress relaxation, and related them to the simple spectral model we studied in the previous section.

Let the shear stress acting on the fault plane before and after the occurrence of an earthquake be  $T_0$  and  $T_f$ , respectively. We define stress drop  $\Delta\sigma$ , as the difference

$$\Delta\sigma = T_0 - T_f. \quad (25)$$

The stress drop represents the part of the acting stress which is used to produce the slip of the fault so that  $\Delta\sigma$  is related to slip  $D$ . The relation between stress drop and slip will be in general very complex: it will depend on the geometry of the fault, but also on certain fundamental assumptions about the stress field in the earth. In general, stress will be much more heterogeneous than slip because, at least in the static case, stress is a generalized derivative of stress. We will discuss these properties briefly in Sect. "Earthquake Dynamics and the Scaling of Energy". We will assume here that the stress distribution has to be such that a finite amount of energy is released during faulting. This assumption is the basis of fracture mechanics, leading to the condition that stress may have at most inverse square root singularities on the fault surface

(see, e.g. [22,45,60]). In the early work on faults, many authors assumed that earthquakes were due to dislocations, slip distributions that present slip discontinuities at their borders. The best known example of such model is the Haskell [30] rectangular dislocation model. This model produces non-integrable stress concentrations around the edges of the fault that store an infinite amount of strain energy [48]. Thus, even if this model radiates a finite amount of energy it can not be used to estimate energy balance during seismic rupture. This paradox was well known in mechanics, where dislocations and cracks are treated as very different phenomena.

### The Static Circular Crack

A simple model that may be used to explain many of the scaling laws observed in earthquake seismology is that of a static circular ("penny shaped") crack of radius  $a$  lying on the  $x, y$  plane. We assume that the fault is loaded by a uniform initial shear stress  $T_0$ , and that  $T_f$ , the final stress, is also uniform inside the fault. The slip on the fault produced by a constant static stress drop  $\Delta\sigma$  inside a circular crack was computed by Eshelby [20] and Keilis-Borok [42]

$$D(r) = \frac{24}{7\pi} \frac{\Delta\sigma}{\mu} \sqrt{a^2 - r^2}, \quad (26)$$

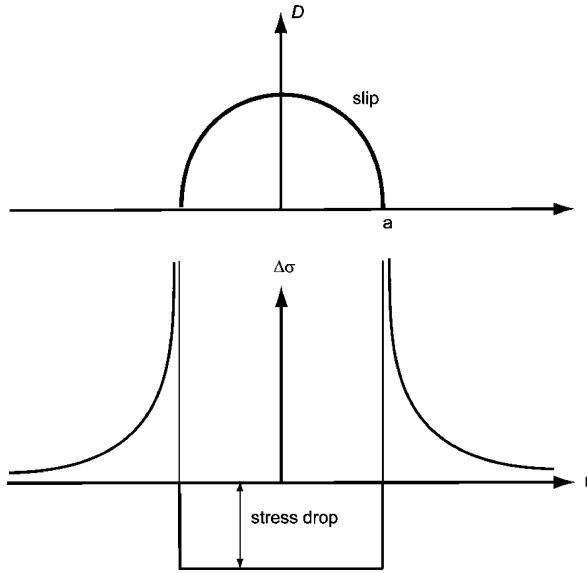
where  $r$  is the radial distance from the center of the crack on the  $(x, y)$  plane,  $a$  is the radius of the crack, and  $\mu$  is the elastic rigidity of the medium surrounding the crack. Slip has the typical elliptical shape associated with cracks. The distribution of slip and the stress change for a circular crack are schematically shown in Fig. 6. Using the definition of the seismic moment (1) we can determine the scalar seismic moment for this circular fault:

$$M_0 = \frac{16}{7} \Delta\sigma a^3 \quad (27)$$

so that the moment is the product of the stress drop times the cube of the fault size. This simple relation will be used to explain the seismic scaling law in terms of fault radius and stress drop. Other fault geometries produce somewhat different scaling laws, including products of fault length and fault width. Unfortunately, as far as I know, no other geometry can be solved in such a simple closed form as the circular crack.

We can also compute the static strain energy change in the elastic medium surrounding the circular fault. This is defined as

$$\Delta W = \frac{1}{2} \int_S \Delta\sigma D dS. \quad (28)$$



Earthquake Scaling Laws, Figure 6  
 The simple static circular crack. The upper panel shows the slip distribution as a function of radius. The bottom panel shows the stress change produced by the slip at the top

From simple thermodynamic considerations,  $\Delta W$  should be negative so that stress drop and slip should have opposite signs. For the circular crack this can be easily computed replacing the slip distribution (26) in this integral. Integrating, we find

$$\Delta W = \frac{8}{7} \frac{\Delta\sigma^2}{\mu} a^3 \quad (29)$$

where we have omitted the negative sign, so that  $\Delta W$  should be interpreted as the reduction of strain energy from the elastic body caused by the earthquake. Dividing (27) into (29) we find that the strain energy to moment ratio for the circular crack is just.

$$\frac{\Delta W}{M_0} = \frac{1}{2} \frac{\Delta\sigma}{\mu} \quad (30)$$

In the very early studies of seismic rupture it was sometimes assumed that radiated energy was equal to strain energy change, i.e.  $\Delta W = E_r$ . In that case apparent stress  $\sigma_a = 1/2 \Delta\sigma$  his assumption is sometimes referred to as the Orowan [58] model. This model is very unlikely to hold for real earthquakes: if all the available energy were radiated, there would be no energy left for producing rupture propagation and consequently rupture should propagate exactly at the P wave or the S wave speeds or should stop immediately in front of any obstacle (see [21,44]).

The circular crack model has been used to quantify numerous earthquakes for which the moment was estimated from the amplitude of seismic waves, and the source radius was estimated from corner frequencies, surface deformation, etc. The result is that for shallow earthquakes in the seismogenic zones like the San Andreas Fault, or the North Anatolian Fault in Turkey, average stress drops are of the order of 1–10 MPa. For deeper events in subduction zones, stress drops can reach several tens of MPa. Thus, average stresses do not vary much, compared to the variation of moment over more than 15 orders of magnitude.

### Brune's [17] Model of Seismic Radiation

Brune developed a model of seismic radiation based on the observation that seismic spectra had the omega-squared spectral shape (7). Brune [17] proposed a model only for shear waves; although, as we already mentioned, P waves have a similar spectral shape with a corner frequency that is different from that of S waves. In his 1970 paper proposed also a relation between the corner frequency  $f_0$  and the source radius  $a$  of a circular fault:

$$f_0 = 0.3724 \frac{\beta}{a} \quad (31)$$

That is, the corner frequency is inversely proportional to the size of the fault. The origin of the coefficient 0.3724 that appears in (31) is very important because it has major consequences for the partitioning of elastic energy. Although the steps followed by Brune were different, we can obtain his results in the following way.

Replacing the expression for moment of a circular crack (27) into the expression for energy moment ratio (17) we get

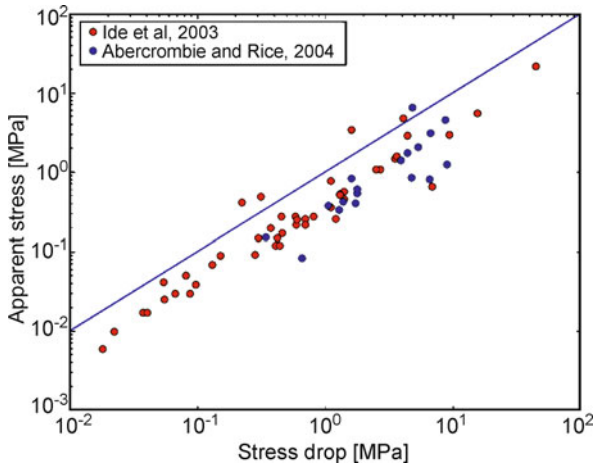
$$\frac{E_r^S}{M_0} = 4.5118 \frac{\Delta\sigma}{\mu} \frac{a^3 f_0^3}{\beta^3} \quad (32)$$

inserting Brune's expression for the corner frequency (31) into (32) we get the apparent stress

$$\sigma_a = \frac{\mu E_r^S}{M_0} = 0.2331 \Delta\sigma \quad (33)$$

So that apparent stress drop in Brune's model is proportional to the static stress drop  $\Delta\sigma$ . This result, derived by Singh and Ordaz [69], has a very interesting consequence for the energy balance in earthquakes. Indeed, inserting the relation (30) between moment and strain energy change during an earthquake into (34), we get

$$E_r^S = 0.466 \Delta W \quad (34)$$



Earthquake Scaling Laws, Figure 7

Apparent stress versus stress drop for the data reported by Ide et al. [32] and Abercrombie and Rice [2]. This figure suggests that apparent stresses computed directly from radiated energy and moment are not independent of stress drop computed for a static circular crack model

That is Brune's model makes the implicit assumption that the energy radiated in the form of S waves is 46.6% of the available strain energy. This was computed by Brune [17]; he found that S waves carried 44% of the available energy in his Equation (40). The number cited here, 46.6% is due to Brune's [18] correction of (31). I think that this is a very important consequence of the corner frequency source radius relationship (31), that is not often cited in the literature.

We can now test whether stress drops computed using the static circular crack model produces are or not independent of the apparent stresses computed from the energy moment ratio  $E_r/M_0$ .

In Fig. 7 we plot the ratio of apparent stress to stress drop for the data of Ide et al. [35] and Abercrombie and Rice [2]. The relation is roughly linear but the ratio is not well defined, although it is clearly less than 1 for most of the events.

### Earthquake Dynamics and the Scaling of Energy

Earthquakes are dynamic processes in which rupture propagates under the control of friction that acts between the two sides of the fault as they slip. The study of the friction law that actually operates on seismic faults is a major problem in seismology and fracture mechanics. Laboratory experiments, seismic observations and field studies are needed to solve this complex problem. In this section we will attempt to establish some general

properties of seismic sources without getting involved with fine details about friction and rupture propagation (see the contribution by Ampuero in the present volume for a fuller discussion). The main question in this context is: can we establish some general properties of seismic ruptures that are independent of the details of friction? Do the observations of seismic spectra and scaling laws constrain in any way the overall properties of seismic sources? This approach has been taken in recent years by many authors, some have tried to convert slip models inverted from near field seismic observations to determine energy balance [16,34]; others have tried to do the same by remarking that dynamic ruptures only propagate at reasonable rupture speeds for a very limited range of seismic parameters [57]; or, very recently, have tried to derive general properties of the friction law from the scaling of seismic spectra [2]. We will follow the latter approach because I believe that it is very promising.

### The Dynamic Circular Crack Model

Perhaps, the simplest fault model that can be imagined is a circular crack that grows from a point at a constant or variable rupture speed and then stops at the rim of the fault, arrested by the presence of unbreakable barriers. This model is the natural dynamic equivalent to the static circular crack discussed in the previous section. The circular crack problem is posed in terms of stresses not of slip, but the rupture process is fixed in advance so that rupture does not develop spontaneously. This is the only unrealistic feature of this model, hence it is considered as quasidynamic, that is, rupture is kinematically defined, but slip is computed solving the elastodynamic equations. This model was carefully studied by a number of authors in the 1970s [43,46,47,62,65].

Let us consider a rupture that starts from a point and then spreads self-similarly at constant rupture speed  $v_r$  without ever stopping. Slip on this model is driven by stress drop inside the fault. The solution of this problem is somewhat difficult to obtain because it requires very advanced use of self-similar solutions to the wave equation and its complete solution for displacements and stresses must be computed using the Cagniard de Hoop method [62]. Fortunately, the solution for slip across the fault found by Kostrov [43] is surprisingly simple. Slip in the circular fault is parallel to the direction of stress drop on the fault and it has the typical elliptical shape:

$$D(r, t) = C(v_r) \frac{\Delta\sigma}{\mu} \sqrt{v_r^2 t^2 - r^2} \quad (35)$$

where  $r$  is the radius in a cylindrical coordinate system centered on the point of rupture initiation.  $v_r t$  is the instantaneous radius of the rupture at time  $t$ .  $\Delta\sigma$  is the dynamic stress drop assumed to be constant inside the rupture zone,  $\mu$  is the elastic rigidity, and  $C(v_r)$  is a slowly varying function of the rupture velocity  $v_r$ . For most practical purposes  $C \sim 1$ . This simple solution constitutes a key result containing one of the most important properties of circular cracks. Slip inside the fault scales with the ratio of stress drop over rigidity times the instantaneous radius of the fault. As rupture develops, slip increases with the size of the rupture zone.

### Energy Release Rate for a Dynamic Circular Crack

We can determine the energy release rate for Kostrov’s model (35) from the behavior of stresses near the edge of the crack. At time  $t$ , the fault radius is  $r = v_r t$ , the slip velocity field derived from (35) has the form

$$V(r, t) = C(v_r) \frac{\Delta\sigma}{\mu} \frac{v_r^2 t}{\sqrt{v_r^2 t^2 - r^2}} \tag{36}$$

so that, near the rupture front, the velocity field presents the well known inverse-squared root singularity predicted by dynamic crack theory [22,45]. We can then approximate the singularity in slip rate in the general form

$$V(r, t) = \frac{V_d}{\sqrt{2\pi}} \frac{1}{\sqrt{v_r t - r}} \tag{37}$$

where  $V_d$  is the velocity intensity factor, a measure of the amplitude of the square root singularity in slip velocity that moves with the rupture front. This velocity singularity is associated with a dynamic stress concentration ahead of the rupture front

$$\Delta\sigma(r, t) = \frac{K_d}{\sqrt{2\pi}} \frac{1}{\sqrt{r - v_r t}} \tag{38}$$

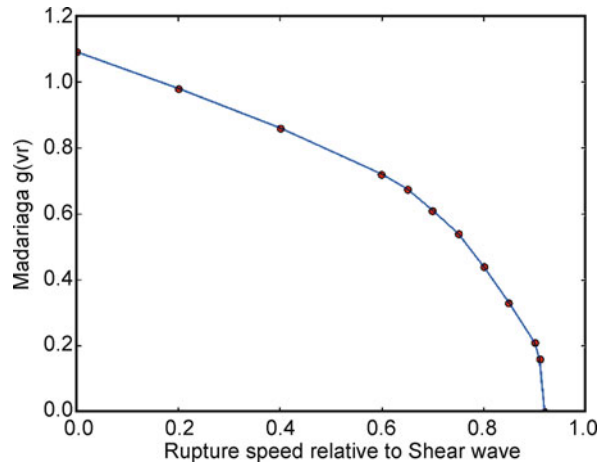
which is also of the inverse square root type.  $K_d$  is the dynamic stress intensity factor. The amplitudes  $V_d$  and  $K_d$  are linearly related to each other with a coefficient that is different for fracture modes II and III. Avoiding details that are discussed by Freund [22], we can write

$$V_d = a(v_r) \frac{K_d}{\mu} v_r \tag{39}$$

where  $a(v_r)$  is a coefficient that depends on the instantaneous rupture velocity  $v_r$ .

We can compute the energy release rate near the border of the fault directly from these expressions, using several results from Kostrov [43] and Madariaga [47]:

$$G_c(v_r) = \frac{K_d V_d}{v_r} = a(v_r) \frac{K_d^2}{2\mu} \tag{40}$$



Earthquake Scaling Laws, Figure 8  
**The  $g(v_r)$  function of rupture velocity. This function controls the fraction of strain energy that is used as fracture energy in a quasi-dynamic circular shear crack that ruptures at a constant rupture speed**

that is, the energy release rate  $G_c$  near the crack front is proportional to the square of the dynamic stress intensity factor. Since the dynamic stress intensity factor  $K_d$  tends to zero at high rupture speeds, the energy release rate  $G_c$  also decreases at high speed rates. Thus, the faster the rupture, the less energy is spent in making the rupture advance.

For the circular crack, the energy release rate is not uniform around the perimeter of the circular fault because it is different in mode II (in-plane) and mode III (anti-plane). From results by Madariaga [46] we get for a circular crack

$$G_c(v_r, r) = \frac{g(v_r)}{3} \frac{\Delta\sigma^2}{\mu} r. \tag{41}$$

Where  $g(v_r)$  is a monotonically decreasing function of rupture speed shown in Fig. 8. An exact expression for  $g(v_r)$  was proposed by Madariaga [46], but his Eq. (30) contains a misprint that was corrected by Ide [32].

### The Scaling of Energy Release Rate with Earthquake Size

We can now estimate the energy used for the propagation of the seismic rupture using the previous estimate for Brune’s [17] model (34). We can establish the following global energy balance for an earthquake that is well modeled by Brune calibration of the shear wave spectra.

As rupture propagates, strain energy released by faulting  $\Delta W$  is used in part to produce seismic waves, and in part to make fracture advance (measured by the energy re-

lease rate  $G_c$ ). Assuming that  $G_c$  is constant over the fault surface we find the following earthquake energy balance

$$\Delta W = E_r + G_c S \quad (42)$$

where  $S$  is the area of the fault. In order to estimate  $G_c$ , we need to estimate the total radiated energy  $E_r$ . Brune's model is only suitable to compute S wave radiated energy which as shown by Eq. (34) is 46.6% of the available strain energy. Energy transported by P waves can be computed using either theoretical arguments or observations of P to S energy ratios. This question was examined by Boatwright and Fletcher [15] who concluded that, theoretically at least, a crack-like source should produce about 15 times more energy carried by S waves than P wave energy. This number is confirmed by observations reported by Abercrombie and Rice [2], Prieto et al. [59]. We can thus estimate that radiated energy in the Brune model is roughly 50% of the strain energy released by the earthquake; the other 50% goes into rupture work, i. e.

$$G_c S \approx \frac{1}{2} \Delta W. \quad (43)$$

Using the expression (29) for  $\Delta W$  we get

$$G_c \approx \frac{1}{2} \frac{\Delta W}{\pi a^2} = \frac{4}{7\pi} \frac{\Delta \sigma^2}{\mu} a. \quad (44)$$

That is, if we adopt Brune's model, *energy release rate scales with fault size*.

In the previous calculation we assumed that  $G_c$  was uniform inside the fault. A better assumption would be that, as the rupture propagates,  $G_c$  grows with the radius as in the simple quasi-dynamic model of Madariaga [46],

$$G_c(r) = \frac{\Delta \sigma^2}{3\mu} r g(v_r) \quad (45)$$

where  $g(v_r)$  will be determined for Brune's model. The total energy release during rupture is then

$$\int_S G_c(r) dS = \frac{2\pi}{9} \frac{\Delta \sigma^2}{\mu} g(v_r) a^3.$$

Using (43) again we get

$$g(v_r) = \frac{18}{7\pi} = 0.818. \quad (46)$$

Thus, Brune's model is equivalent to a circular quasidynamic shear crack propagating such that the energy release rate grows with fault radius like

$$G_c = \frac{6}{7\pi} \frac{\Delta \sigma^2}{\mu} r.$$

Thus, whether we use a constant energy release rate on the fault (44), or a more realistic model where energy release grows with the radius of the fault, we find that energy release rate grows with fault radius in Brune's model. This result confirms that energy release rate scales with the fault radius and that it adjusts as the fault grows [55,56].

The relation between energy release rate and fault size was studied by Abercrombie and Rice [2] using their own data and data from a number of previous studies. They reached the conclusion that  $G_c$  grows with radius roughly like  $r^{0.4}$ , not like the radius as in (41) and (47). Abercrombie and Rice estimated  $G_c$  from the expression

$$G_c = \frac{1}{2} (\Delta \sigma - \sigma_a) D$$

where  $D$  is slip. This expression is entirely compatible with ours, so that the reason  $G_c$  scales with the power 0.4 of the radius is that stress drops scale with earthquake size, specially if  $\Delta \sigma$  is not linearly related to  $\sigma_a$  as shown in Fig. 7. After checking Fig. 8 of Abercrombie and Rice [2], I conclude that they obtained scaling with a power of 0.4 using the full data set, including much larger earthquakes. For the Cajon pass earthquakes studied by Abercrombie and Rice the power of radius in scaling is much closer to one. Seismic data have now reached the quality necessary to test different hypothesis about the scaling of rupture energy in order to see whether larger earthquakes are really different from smaller ones.

### Scaling of Energy, Magnitude and Moment

Kanamori [38] introduced the so-called moment magnitude,  $M_w$ , assuming that all available strain energy was converted into seismic waves, i. e. that  $E_r \approx \Delta W$ . This assumption means that no energy is used to propagate fracture so that  $G_c = 0$ . Using the definition of strain energy change, (28), we get Kanks and Kanamori [28]:

$$E_r = \frac{1}{2} \Delta \sigma \bar{D} S, \quad (47)$$

where  $\bar{D}$  is the average slip of the earthquake and  $S$  its source area.

This expression shows that from the radiated energy we only have information about the stress drop, not about the absolute stress level acting on the fault during faulting. Rewriting expression (27)

$$M_0 = \frac{16}{7\pi^{3/2}} \Delta \sigma S^{3/2} \quad (48)$$



and taking logarithms

$$\log M_0 = \frac{3}{2} \log S + \log \left( \frac{16\Delta\sigma}{7\pi^{3/2}} \right). \quad (49)$$

From this equation it follows that, for constant stress drop scaling,  $\log S$  should be proportional to  $2/3 \log M_0$ . This hypothesis had been shown empirically to be valid for a large range of values of  $M_0$  [39]. We noticed however that more recent data (Fig. 6) shows that stress drop varies over 3 orders of magnitudes, at least in the data reported by Ide and Beroza [33].

Assuming constant stress drop, Kanamori [38] defined the moment magnitude  $M_w$  based on the empirical relation of Gutenberg and Richter [23,24] between surface wave magnitude and seismic energy (in Joules):  $\log E_r = 1.5M_S + 4.8$ . Hanks and Kanamori [28] proposed the moment magnitude scale

$$M_W = \frac{2}{3} \log M_0 - 6.07 \quad (50)$$

where  $M_0$  is measured in Nm. The moment magnitude has become the standard way to measure the size of earthquakes. Both  $M_w$  and seismic moment  $M_0$  (Nm) can be related to other magnitude measurements by a number of empirical relations. (see, e. g. [39]).

Radiated seismic energy can not be computed directly from (47) because it needs to be corrected for the part of the strain energy that is used to propagate the fracture (see, (42)).

### More General Scaling Relations Derived from the Scaling Law of Earthquake Spectra

In the two previous sections several scaling relations have been established which relate the parameters involved in the fracture process of earthquakes. It has been shown that slips and slip velocities scale linearly with stress drop, which is the most fundamental scaling parameter (pp. 202–211 in [67]). If the average stress drop measured over the whole fault plane is roughly constant for all earthquakes, the slip on the fault should scale with the dimensions of the fault ( $L$ ) which for small earthquakes represents the length ( $L$ ) or radius of the fault and for large earthquakes the width ( $W$ ). An unsolved issue, due mostly to lack of data for very long strike slip earthquakes, is that of a possible difference in the scaling of the seismic moment with fault length, between large and small earthquakes. Strike slip earthquakes with seismic moment less than  $10^{21}$  Nm ( $M_w < 8$ ) should scale with  $L^3$ , while larger ones with  $L^2$ .

Many other relations can be established starting from the basic scaling laws discussed earlier in this chapter. For instance, maximum and average slip for earthquakes scale like the cubic root of moment for most earthquakes [53,54], but these scaling relations can be derived from the basic relations discussed earlier.

### More Realistic Radiation Model

In reality earthquakes occur in a complex medium that is usually heterogeneous and dissipative. Seismic waves become diffracted, reflected, and in general suffer from multipathing in those structures. Accurate seismic modeling would require perfect knowledge of Earth's structure. It is well known and understood that structural complexities dominate signals at certain frequency bands. For this reason the simple model presented here can be used to understand the main features of earthquakes at long wavelengths, while the more sophisticated approaches that attempt to model every detail of the wave form are reserved only for more advanced studies. Here, like in many other areas of geophysics, a balance between simplicity and concepts must be kept against numerical complexity that may not always be warranted by lack of knowledge of the details of Earth's structure. If the simple approach were not possible, then many standard methods to study earthquakes would be impossible to use. A good balance between simple, but robust concepts and the sophisticated reproduction of the complex details of real wave propagation is a permanent challenge for seismologists.

### Why Does the Spectrum Decay Like $\omega$ Squared?

We have seen that seismic data is in very broad agreement with the general features of Brune's spectral model (7). We have explained the scaling of low frequencies in terms of simple static source models, the corner frequency and the high frequency decay are explained by the energy balance Eq. (32). Seismic energy must be finite and a well defined fraction of the available strain energy. These conditions require that the spectrum of seismic energy is integrable in expression (13). General properties of Fourier transform can be invoked to demonstrate that in the time domain displacement signals are continuous functions of time with discontinuous derivatives. That is, the velocity field emitted by a seismic source in the far field contains jumps in particle velocity as is the case with the seismic signals proposed by Brune (19) or by Kanamori and Rivera (20). For the dynamic circular crack studied in this section, the velocity jumps are emitted when rupture stops abruptly at the rim of the circle of radius  $a$ . The

nature of these stopping phases has been carefully studied by a number of authors, including Madariaga [46,47], Boatwright [12], Spudich and Frazer [71] and Bernard and Madariaga [8]. Their study is very complex and beyond the purpose of the present article, we will use a simpler approach based on a scaling argument.

Let us now consider how the stopping phases scale with earthquake size and stress drop. In the omega squared model, the high frequency decay of the far-field displacement produced by shear waves can be written in the very general form

$$u_s(r, \omega) = C(\theta, \varphi) \frac{M_0 \omega_0^2}{\mu \beta} \frac{1}{R} \omega^{-2},$$

where we lumped the numerical coefficients and the radiation pattern in the single non-dimensional coefficient  $C(\theta, \varphi)$ .  $R$ ,  $\theta$  and  $\varphi$  are spherical coordinates at a reference point on the fault. Using the expressions for  $M_0$  and corner frequency in terms of fault radius and stress drop we obtain

$$u(r, \omega) = C(\theta, \varphi) \frac{\Delta \sigma a \beta}{\mu} \frac{1}{R} \omega^{-2} \quad (51)$$

with a slightly different dimension-less coefficient. The far field waves scale at high frequencies like the product of stress drop times the radius of the fault. That is exactly what was predicted by Madariaga [47] for a circular crack. It is interesting to remark that the factor  $\Delta \sigma a$  actually comes from the product of the stress intensity factor  $K$  and the square root of the radius  $a$ . Although the scaling of high frequencies was derived here for a very special circular crack model, it can be easily generalized to ruptures of any shape, splitting the factor  $\Delta \sigma a$  into a stress intensity factor and the square root of the local radius of curvature of the wave front of the stopping phase.

The previous model for the radiation of omega squared high frequency waves can be extended to more complex source models, in particular to source models that contain a number of subfaults (see, e. g., [14]). The high frequency seismic waves emitted by such a model are due to stopping phases emitted all along their propagation process. Each such stopping phase contributes to enriching the high frequency contents of the seismic waves. The incoherent sum of those phases produces a total spectrum that scales with fault radius as in (51). In this sense, omega squared decay is the signature of the presence of cracks on the fault. Recent work has shown that omega squared waves are emitted every time the rupture front changes rupture velocity, or that the rupture is deviated from a plane by the presence of fault kinks or discontinuities.

## Kinematics and Statistical Models for Fault Slip

So far we have discussed a dynamic crack approach to understanding earthquake scaling. Another method to describe seismic sources radiation was introduced by Haskell [30]. He assumed that earthquakes could be described by simple propagating dislocations leaving a constant slip in their trail. The most common such model is that of a flat rectangular fault with constant slip in it. Such model is mechanically impossible, because it needs infinite amount of energy to be created. Curiously, though, the Haskell model produces a finite amount of radiated energy and the radiated field can be computed exactly both in the far [31] and near field [48]. Because the strain energy change produced by Haskell's model is infinite, an energy balance equation like (42) can not be established. In Haskell model all of the energy released from the medium is absorbed by the dislocation motion and seismic radiation is just a secondary feature of the source processes.

Dislocation models have appeared in a different form, derived from statistical considerations about the distribution of slip on faults. In his seminal paper, Andrews [6] established some basic statistical properties of slip distributions that are actually based on Haskell's [31] original study of the power spectra of slip and the correlation functions of slip. His analysis is based on some general features of fractal surfaces and distributions; here we will look at these scaling relationships from the point of view of the circular crack model that we discussed in previous sections.

The slip function of a circular crack was defined in (26). This is a function of radius only so that its Fourier transform is very easy to compute using the Hankel transform:

$$\tilde{D}(k) = 2\pi \int_0^\infty D(r) J_0(kr) r dr \quad (52)$$

where  $k$  is the radial wave-number,  $J_0$  is the Bessel function of degree zero. Inserting the expression (26) in (52) and integrating we get

$$\tilde{D}(k) = \frac{48}{7} \frac{\Delta \sigma}{\mu} a^3 \frac{\sin(ak) - ak \cos(ak)}{k^3 a^3}. \quad (53)$$

At low wave-number, when  $k$  tends to 0, the spectrum (53) tends to the value

$$\tilde{D}(0) = \frac{16}{7} \frac{\Delta \sigma}{\mu} a^3 = \frac{M_0}{\mu}$$

that is, the low wave number limit of the slip spectrum is the seismic moment, just as the low frequency limit of Brune's spectrum is the seismic moment. This is of course

not a coincidence but a consequence of the fact that the seismic moment is the source of low frequency waves. At high wave numbers, when  $ka \gg 1$ , the spectrum  $D(k)$  behaves like

$$\lim_{k \rightarrow \infty} \tilde{D}(k) = -\frac{48}{7} \frac{\Delta\sigma}{\mu} a^3 \frac{\cos(ak)}{a^2 k^2}. \quad (54)$$

The slip spectrum decays like  $k^{-2}$  at high wave numbers, a property that seems to be as universal as the high frequency decay of seismic spectra with omega-squared.

The important issue is why is it  $k^{-2}$ ? The origin of the high frequency behavior of the slip spectrum can be determined with some simple properties of the Hankel transform (52). Take a circular fault of finite radius  $a$ . Then for different types of slip discontinuity we get the following asymptotic behavior:

	$\lim_{r \rightarrow a} D(r)$	$\lim_{k \rightarrow \infty} D(k)$
Constant	1	$k^{-3/2}$
Crack-like	$(a-r)^{1/2}$	$k^{-2}$
Conical	$(a-r)$	$k^{-5/2}$
Smooth	$(a-r)^{3/2}$	$k^{-3}$

Thus the high wave number behavior is a reflection of the discontinuity of slip at the border of the fault. Andrews [6] studied a slip distribution that behaves like  $(a-r)^{3/2}$  near the edge of the fault. In his case the high wave number decay is  $k^{-3}$ . Thus the high wave number behavior of slip distributions is controlled by the discontinuities of slip, a crack like discontinuity producing a  $k$ -squared distribution. Note that the spectral behavior for two-dimensional distributions is quite different than for two dimensional slip distributions. Haskell [30] used the properties of 2D Fourier transforms to derive several conclusions about earthquake spectra that do not apply to circular cracks. For a two-dimensional plane or anti-plane crack the spectrum decays like  $k^{-3/2}$ , such a spectrum is inadmissible in 3D because it would imply non-integrable stress distributions as we will show now.

The stress field associated with the circular shear crack slip can be computed in a straightforward way using the expressions provided by Eshelby [20] or Sneddon [70]. Let the fault be located on the plane  $(x,y)$  and slip be parallel to the axis  $x$ , (i. e.  $D(x) = \Delta u_x(x)$ ). In the spectral domain the associated stress drop can be computed from slip by

$$\Delta\sigma_{xz}(k) = -\frac{\mu}{2k} \left( \frac{2(\lambda + \mu)}{\lambda + 2\mu} k_x^2 + k_z^2 \right) D(k) \quad (55)$$

(see [64]). This is a relatively simple expression, but it is not easy to compute analytically because stress drop for the circular crack does not have cylindrical symmetry. Andrews [6] provided a simplification that we will use

here: assuming that the elastic constant  $\lambda = 0$  i. e. that is the elastic medium is incompressible), we get in Fourier domain:

$$\Delta\sigma_{xz}(k) = -\mu k D(k). \quad (56)$$

Multiplying (53) by  $-k$  and doing the inverse Hankel transform (see [70]) we get, approximately

$$\begin{aligned} \Delta\sigma_{xz}(r) &= -\Delta\sigma \quad \text{for } r < a, \\ \Delta\sigma_{xz}(r) &\cong K/\sqrt{r-a} \quad \text{for } r > a. \end{aligned} \quad (57)$$

That is, inside the crack, stress drop is constant while outside stress drop exhibits an inverse square root singularity typical of cracks as discussed in (38). This explains why the spectrum of a circular crack decays as  $k^{-2}$  in the wavenumber domain. The  $k^{-2}$  spectrum is the signature of the presence of a crack.

Mai and Beroza [51] computed the correlation lengths, fractal dimensions, Hurst exponents and wave number spectra of 42 earthquakes for which the slip distribution on the fault planes were available from inversion of seismic and geodetic data. They reached the conclusion that the high frequency decay had an average fractal dimension of  $2.29 + / - 0.23$  that implies a high wave number asymptotic decay of  $k^{1.71}$ , that is slip distributions determined from slip inversions tend to be rougher than the spectra of circular cracks. The origin of this exponent needs to be carefully scrutinized in terms of fault segmentation.

Bernard and Herrero [7] and Bernard et al. [9] proposed a model linking seismic radiation to the spectral properties of the distribution of slip on the fault. In their model rupture propagates in a single space direction at constant speed. In order to obtain an omega squared far field spectrum they assumed that rise time is essentially a delta function or that it scales with wave number, which is equivalent. This result needs to be confronted with dynamic simulations propagating at finite speeds and finite energy release rate. In the circular crack model, the high frequency spectrum was controlled by stopping phases, which do not exist in the Bernard et al. model. This problem needs careful attention, specially in the presence of geometrical heterogeneity that may produce stopping phases.

### Future Directions

Most properties of ensemble averaged seismic spectra and slip distributions can be explained by a simple circular crack model. Seismic waves as well as slip distributions determined from seismic and geodetic inversions carry

the signature of the crack models that are at the base of earthquake ruptures. Whether the earthquake can be modeled as a simple circular crack or as the complex sum of a distribution of such cracks the result is the same: slip is of  $k$ -squared type and seismic radiation is of  $\omega$ -squared type. Departure from these models can be expected if stress drop scales with fault size. There is no clear-cut evidence for such behavior because of the difficulties in estimating radiated energy mentioned several times in this review. The variations of stress drop required to explain observations may be an intrinsic variation of stresses depending on fault maturity, the position of the fault in the seismic cycle, etc.

In recent years the quality and quantity of seismic data has improved very significantly with the deployment of digital instruments in many active areas of the earth. This is a unique opportunity to test the self-similarity of earthquakes. If tests like those of Fig. 4 are applied to new data, the problem of the variability of radiated energy/moment ratio (or equivalent of apparent stress and stress drop) will be addressed more carefully. I would not be surprised if we finally concluded that apparent stress varies significantly among different earthquakes as suggested by Fig. 4.

### Acknowledgments

This research was partially funded by the SEISMULATORS contract with ANR under program Catastrophes Telluriques et Tsunamis, and by the Research training network SPICE of the 7th PCRD of the European Union. I am deeply indebted to Luis Rivera, Martin Mai and Art McGarr for their careful and patient review of an initial version of this paper.

### Bibliography

- Abercrombie RE (1995) Earthquake source scaling relationships from  $-1$  to 5 ML using seismograms recorded at 2.5 km depth. *J Geophys Res* 100:24015–24036
- Abercrombie RE, Rice JR (2005) Can observations of earthquake scaling constrain slip weakening? *Geophys J Int* 162:406–424
- Aki K (1966) Generation and propagation of G waves from the Niigata earthquake of June 16, (1964) part 2. Estimation of earthquake movement, released energy, and stress-strain drop from G wave spectrum. *Bull Earthq Res Inst Univ Tokyo* 44:23–88
- Aki K (1967) Scaling law of seismic spectrums. *J Geophys Res* 72:1217–1231
- Aki K, Richards PG (2002) *Quantitative Seismology*, 2nd edn. University Science Books, Sausalito
- Andrews DJ (1980) A stochastic fault model: 1. Static case. *J Geophys Res* 85:3867–3877
- Bernard P, Herrero A (1994) A kinematic self-similar rupture process for earthquakes. *Bull Seismol Soc Am* 84:1216–1228
- Bernard P, Madariaga R (1984) A new asymptotic method for the modeling of near-field accelerograms. *Bull Seismol Soc Am* 74:539–557
- Bernard P, Herrero A, Berge C (1996) Modeling directivity of heterogeneous earthquake ruptures. *Bull Seismol Soc Am* 86:1149–1160
- Beroza G, Kanamori H (2007) Comprehensive overview in Treatise on geophysics. In: Schubert G (ed) *Earthquake Seismology*. Elsevier, Amsterdam
- Boatwright J (1978) Detailed spectral analysis of two small New York State earthquakes. *Bull Seismol Soc Am* 68:1177–1131
- Boatwright J (1980) A spectral theory for circular seismic sources: Simple estimates of source dimension, dynamic stress drop, and radiated energy. *Bull Seism Soc Am* 70:1–26
- Boatwright J (1982) A dynamic model for far field acceleration. *Bull Seism Soc Am* 72:1049–1068
- Boatwright J (1988) The seismic radiation from composite models of faulting. *Bull Seism Soc Am* 78:489–508
- Boatwright J, Fletcher JB (1985) The partition of radiated energy between P and S waves. *Bull Seismol Soc Am* 75:361–376
- Bouchon M (1997) The state of stress on some faults of the San Andreas Fault system as inferred from near-field strong-motion data. *Bull Seismol Soc Am* 58:367–398
- Brune JN (1970) Tectonic stress and spectra of seismic shear waves from earthquakes. *J Geophys Res* 75:4997–5009
- Brune JN (1971) Correction. *J Geophys Res* 76:5002
- Brune JN, Archuleta RJ, Hartzell S (1979) Far-field S wave spectra, corner frequencies, and pulse shapes. *J Geophys Res* 84:2262–2272
- Eshelby JD (1957) The elastic field of an ellipsoid inclusion and related Problems. *Proc R Soc Lond A* 241:376–396
- Freund LB (1972) Energy flow into the tip of an extending crack in an elastic solid. *J Elast* 2:341–348
- Freund LB (1989) *Fracture Dynamics*. Cambridge University Press, Cambridge
- Gutenberg B, Richter CF (1942) Earthquake magnitude, intensity, energy, and acceleration. *Bull Seism Soc Am* 32:163–191
- Gutenberg B, Richter CF (1956) Earthquake magnitude, intensity, energy, and acceleration (second paper). *Bull Seism Soc Am* 46:105–145
- Hanks TC (1979)  $b$  values and  $w^{-g}$  seismic source models: Implications for tectonic stress variations along active crustal fault zones and the estimation of high-frequency ground motion. *J Geophys Res* 84:2235–2242
- Hanks TC (1981) The corner frequency shift, earthquake source models, and  $Q$ . *Bull Seismol Soc Am* 71:597–612
- Hanks TC (1982)  $f_{max}$ . *Bull Seismol Soc Am* 72:1867–1879
- Hanks TC, Kanamori H (1979) A moment magnitude scale. *J Geophys Res* 84:2348–2350
- Hanks T, Thatcher W (1972) A graphical representation of seismic source parameters. *J Geophys Res* 77:4393–4405
- Haskell NA (1964) Total energy spectral density of elastic wave radiation from propagating faults. *Bull Seism Soc Am* 54:1811–1841
- Haskell NA (1966) Total energy spectral density of elastic wave radiation from propagating faults: Part II. A statistical source model. *Bull Seism Soc Am* 56:125–140
- Ide S (2002) Estimation of Radiated energy of finite-source earthquake models. *Bull Seism Soc Am* 92:2294–3005

33. Ide S, Beroza GC (2001) Does apparent stress vary with earthquake size? *Geophys Res Letters* 28:3349–3352
34. Ide S, Takeo M (1997) Determination of constitutive relations of fault slip based on seismic wave analysis. *J Geophys Res* 102(27):379–391
35. Ide S, Beroza GC, Prejean SG, Ellsworth WL (2003) Apparent Break in Earthquake Scaling Due to Path and Site Effects on Deep Borehole Recordings. *J Geophys Res* 108(B5):2271. doi:10.1029/2001JB001617
36. Jost M, Busselberg LT, Jost O, Harjes HP (1998) Source parameters of injection-induced microearthquakes at 9 km depth at the KTB deep drilling site, Germany. *Bull Seismol Soc Am* 88:815–832
37. Joyner WB (1984) A scaling law for the spectra of large earthquakes. *Bull Seism Soc Am* 74:1167–1188
38. Kanamori H (1977) The energy release in great earthquakes. *J Geophys Res* 82:2921–2987
39. Kanamori H, Anderson DL (1975) Theoretical basis of some empirical relations in seismology. *Bull Seismol Soc Am* 65:2981–2987
40. Kanamori H, Rivera L (2004) Static and Dynamic Scaling Relations for Earthquakes and their implications for Rupture Speed and Stress Drop. *Bull Seismol Soc Am* 94:314–319. [http://www.gps.caltech.edu/faculty/kanamori/static\\_dynamic\\_scaling\\_relations.pdf](http://www.gps.caltech.edu/faculty/kanamori/static_dynamic_scaling_relations.pdf)
41. Kanamori H, Mori J, Hauksson E, Heaton TH, Hutton LK, Jones LM (1993) Determination of earthquake energy release and  $M_L$  using Terrascope. *Bull Seismol Soc Am* 83:330–346
42. Keilis-Borok VI (1957) Investigation of the mechanism of earthquakes. *Tr Inst Geofis Akad Nauk, SSSR* 40 (in Russian). (1960) *Sov Res Geophys Ser 4* (Engl transl)
43. Kostrov BV (1964) Self-similar problems of propagation of shear cracks. *J Appl Math Mech* 28:1077–1087
44. Kostrov BV (1974) Seismic moment and energy of earthquakes and seismic flow of rock. *Izv Earth Phys* 1:23–40
45. Kostrov B, Das S (1988) *Principles of Earthquake Source Mechanics*. Cambridge University Press, Cambridge
46. Madariaga R (1976) Dynamics of an expanding circular fault. *Bull Seism Soc Am* 66:639–666
47. Madariaga R (1977) High frequency radiation from crack (stress drop) models of earthquake faulting. *Geophys J R Astr Soc* 51:625–651
48. Madariaga R (1978) The dynamic field of Kaskell's rectangular dislocation fault model. *Bull Seismol Soc Am* 68:869–887
49. Madariaga R (1979) On the relation between seismic moment and stress drop in the presence of stress and strength heterogeneity. *J Geophys Res* 84:2243–2250
50. Mai PM, Beroza GC (2000) Source-scaling properties from finite-fault rupture models. *Bull Seis Soc Am* 90(3):604–615. <http://www.seismo.ethz.ch/staff/martin/papers/BSSA00scalingRP.pdf>
51. Mai PM, Beroza GC (2002) A spatial random-field model to characterize complexity in earthquake slip. *J Geophys Res* 107:2308. doi:10.1029/2001JB000588. <http://www.seismo.ethz.ch/staff/martin/papers/JGR02MaiSlipComplex.pdf>
52. Mayeda K, Walker WR (1996) Moment, energy, stress drop, and source spectra of western United States earthquakes from regional coda envelopes. *J Geophys Res* 101:11195–11208
53. McGarr A (1999) On relating apparent stress to the stress causing earthquake fault slip. *J Geophys Res* 104:3003–3011
54. McGarr A, Fletcher JB (2003) Maximum Slip in Earthquake Fault Zones, Apparent Stress, and Stick-Slip Friction. *Bull Seismol Soc Am* 93:2355–2362
55. Ohnaka M (2003) A constitutive scaling law and a unified comprehension for frictional slip failure, shear fracture of intact rock, and earthquake rupture. *J Geophys Res* 108:B2080. doi:10.1029/2000JB000123
56. Ohnaka M, Shen L-F (1999) Scaling of the shear rupture process from nucleation to dynamic propagation: Implications of geometric irregularity of the rupturing surfaces. *J Geophys Res* 104:817–844
57. Olsen KB, Madariaga R, Archuleta RJ (1997) Three-dimensional dynamic simulation of the 1992 Landers Earthquake. *Science* 278:834–838
58. Orowan E (1960) Mechanism of seismic faulting. *Geol Soc Am Memoir* 79:323–345
59. Prieto GA, Shearer PM, Vernon FL, Kilb D (2004) Earthquake source scaling and self-similarity estimation from stacking P and S spectra. *J Geophys Res* 109:B08310. doi:10.1029/2004JB003084. <http://pangea.stanford.edu/~gprieto/publication/scaling.pdf>
60. Rice JR (1980) The mechanics of earthquake rupture. In: Dziewonski AM, Boschi E (eds) *Physics of the Earth's Interior*. Proceedings of the International School of Physics "Enrico Fermi", Course 78, 1979, pp 555–569. North Holland, Amsterdam
61. Rivera L, Kanamori H (2005) Representations of the radiated energy in earthquakes. *Geophys J Int* 162:148–155
62. Richards PG (1973) The dynamic field of a growing plane elliptical shear crack. *Int J Solids Struct* 9:843–861
63. Richardson E, Jordan TH (2002) Seismicity in deep gold mines of South Africa: Implications for tectonic earthquakes. *Bull Seismol Soc Am* 92:1766–1782
64. Ripperger J, Mai PM (2004) Fast computation of static stress changes on 2D faults from final slip distributions. *Geophys Res Lett* 31:L18610. doi:10.1029/2004GL0594
65. Sato T, Hirasawa T (1973) Body wave spectra from propagating shear cracks. *J Phys Earth* 21:415–431
66. Savage JC (1974) Relation of corner frequency to fault dimensions. *J Geophys Res* 77:3788–3795
67. Scholz CH (2002) *The mechanics of earthquakes and faulting*. Cambridge University Press, Cambridge
68. Shearer PM, Prieto GA, Hauksson E (2006) Comprehensive analysis of earthquake source spectra in southern California. *J Geophys Res*, 111, B06303, doi:10.1029/2005JB003979. [http://pangea.stanford.edu/~gprieto/publication/scsn\\_spectra.pdf](http://pangea.stanford.edu/~gprieto/publication/scsn_spectra.pdf)
69. Singh SK, Ordaz M (1994) Seismic energy release in Mexican subduction zone earthquakes. *Bull Seismol Soc Am* 84:1533–1550
70. Sneddon IN (1951) *Fourier Transforms*. McGraw-Hill, New York
71. Spudich P, Frazer LN (1984) Use of ray theory to calculate high-frequency radiation from earthquake sources having spatially variable rupture velocity and stress drop. *Bull Seismol Soc Am* 74:2061–2082
72. Vassiliou MS, Kanamori H (1982) The energy released in earthquakes. *Bull Seism Soc Am* 72:371–387
73. Wyss M, Brune JN (1968) Seismic moment, stress, and source dimensions for earthquakes in the California-Nevada region. *J Geophys Res* 73:4681–4684

## Earthquake Source: Asymmetry and Rotation Effects

ROMAN TEISSEYRE

Institute of Geophysics, Polish Academy of Sciences,  
Warsaw, Poland

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Asymmetric Continuum and Rotation Effects](#)

[Earthquake Source: Fracture Processes](#)

[Final Remarks](#)

[Acknowledgments](#)

[Bibliography](#)

### Glossary

We use the tensor notation and the summation convention for the repeating indices:

$$T_{ss} = \sum_s T_{ss}, \quad A_k B_k = \sum_k A_k B_k.$$

In some places we underline the symmetric and antisymmetric properties of the tensors using the ( $\cdot$ ), [ $\cdot$ ] brackets for indices:

$$S_{(ik)} = S_{(ki)}, \quad S_{[ik]} = -S_{[ki]}.$$

The deviatoric part of a symmetric tensor,  $T_{(ik)}^D$ , having zero value of trace, is defined as

$$T_{(ik)}^D = T_{(ik)} - \frac{1}{3} \delta_{ik} T_{(ss)}, \quad T_{(ik)}^A = \frac{1}{3} \delta_{ik} T_{(ss)},$$

$$T_{(ss)}^D = \sum_s T_{(ss)}^D = 0,$$

where  $T_{(ik)}$  is any symmetric tensor, while  $T_{(ik)}^A$  is the axial tensor.

In some places we use for the partial differentiations the following notations equivalently:

$$\frac{\partial u_n}{\partial x_k} \leftrightarrow u_{n,k}, \quad \frac{\partial U}{\partial x_\alpha} \leftrightarrow U_{,\alpha},$$

where the indices with Roman characters run from 1 to 3, while those with Greek characters run from 1 to 4.

We use the fully antisymmetric tensor

$$\varepsilon_{lps} = \begin{Bmatrix} 1 \\ 0 \\ -1 \end{Bmatrix} \text{ for } \begin{cases} \text{even permutation of} \\ \text{repeating} \\ \text{odd permutation of} \end{cases} \text{ indices};$$

this fits to the tensor notation and helps to express some operations, e. g. curl:

$$\text{curl } \psi \leftrightarrow \varepsilon_{lps} \frac{\partial}{\partial x_p} \psi_s.$$

### Definition of the Subject

The problem of rotation waves becomes actual again due to the recent observations based on very precise instruments able to measure very small rotation time rates, and due to development and new approaches to the continuum theories.

Our aim is to present a consistent theory describing a continuum subjected to complex internal processes. First, we consider all possible kinds of point-related motions and deformations (strictly speaking, a single couple is not a point-source, as a displacement derivative only tends to a point, but a double couple enters into a family of point-deformations as it can be given by the string-string deformation, that is as a shear point-nucleus).

We define the complex rotation field which includes the spin and twist; the latter describes the angular oscillations of the shear axes and related amplitude variations. Twist point deformation can be represented by the string-string and string-membrane motions. A twist vector is defined as a vector perpendicular to a string-string plane; it becomes an important counterpart to spin and a key to presented theory, in which we shall also include the axial point deformation (e. g., the thermal one).

We believe that all point motions, displacement and rotation, and point deformations, axial and twist, shall be governed by some fundamental laws, and we intend to find the invariant forms of such relations in a frame of a modified continuum theory. Such a continuum theory may give us a new insight into the complexity of processes which can be included in a continual material description; we will demonstrate that interaction of these motions and deformations can lead us to a rich variety of internal processes.

We may mention also that the theory of continuum containing all these deformations with rotational motions (spin and twist), with the inner central motion and with defects may be projected on the intrinsic properties of non-Euclidean space. We confine ourselves only to a remark that application of differential geometry is extremely enlightening for the fundamental understanding of the nonlinear processes.

The independent rotation field can be related to the additional constitutive law joining the antisymmetric stresses and rotations, as proposed by Shimbo [1,2] in his considerations on the friction and fracturing processes; the intro-

duced motion equations are equivalent to the stress moment – angular velocity relation.

For a reference to our consideration, we recall also the Kröner theory and its modifications as introduced by Teisseyre and Boratyński [3,4].

The possible dual description of any motions by means of displacement and rotation fields are being discussed and their formal equivalence concerning the propagation processes is demonstrated, but not valid for the source phenomena. Thus, we point out the importance of physically independent rotation motions in the inner granulation and fracture processes, and different types of rotational deformations are analyzed.

Our considerations bring several results for the following subjects.

For seismology:

- New description of the source processes including the role of rotational processes, and explanation of co-action of the slip and rotation motions
- Theory of the seismic rotation waves
- Thermodynamical conditions for a seismic energy release

For continuum and fracture mechanics:

- Theory of asymmetric continuum with the balance equations for the symmetric and antisymmetric stresses
- The relations between the asymmetric stresses and dislocation field
- A new approach to fracture processes with the hypothesis of the twist-shear release following the extreme angular deformation related to the internal particles, or grains constituting a continuum
- A synchronization role of the specific wave fields in the fracture processes
- Formation of a mylonite zone adjacent to fracturing and its constitutive description

For fluid mechanics:

- Theory of the asymmetric fluid continuum with the non-vanishing rates of the asymmetric stresses
- Explanation of the extreme wave phenomena (solitons)

## Asymmetric Continuum and Rotation Effects

### Introduction

Earthquake rotation effects were observed and discussed at the time when the 19th century seismological science was formed. Some eminent scientists, e.g., Charles Lyell (1797–1875), Charles Darwin (1809–1882), Robert Mallet

(1810–1881), and Alexander von Humboldt (1769–1859), raised the problem of the vortical movements, or vortex motion, induced by earthquakes. After the Lisbon earthquakes (1755) and those of Calabria (1783), many scientists focused their attention on the effects induced by such “vortical” waves. Robert Mallet was the first who precisely explained the observed rotation effects of some surface objects, pointing out the roles of the center of adherence of these objects and their inertia moment in relation to forces twisting the objects (see: Kozak [5], and Ferrari [6]). Many scholars tried to design the instruments to record the “vortical motions”, but the first instrument prepared especially to record such motions was that constructed by Filippo Cecchi, the director of the Ximeniano Observatory of Florence, in 1875. Cecchi’s electrical seismograph used sliding smoked paper. However, at that time it was too early to construct an instrument sensitive enough to obtain any traces of such wave motions.

The problem of seismic rotation waves was apparently closed after the Gutenberg [7] statement (1926) that such waves cannot propagate as they will be immediately attenuated, even when generated at the source. Of course the rotation effects remained, with the related explanation by Mallet, as objects of studies especially in the domain of the macroseismic observations.

From the contemporary point of view, two groups of achievements shall be mentioned; first, related to continuum theories, and second, related to development of the modern very precise instruments, able to record extremely small rotation time rates.

The continuum elastic theory bears from its very origin the serious limitation that the angular motions and related moments are not included. In such a situation, there was no place for a constitutive law describing the reaction between the stress moments and rotation processes. The lack of such a law automatically denies the existence of the rotation waves. We will return to these problems further on.

### Experimental Evidence

The modern instrumentation techniques and the obtained results presenting the rotation wave seismograms need more attention. First, we can mention that maybe the first rotation seismogram (see Teisseyre [8]) was achieved in an indirect way: the azimuth array of horizontal seismographs, installed in one of the coal mines in Upper Silesia, Poland, to record the nearby tremors permitted one to deduce the rotational component of motions. However, the first, fully documented, rotation seismograms were obtained at the two geodetic fundamental stations in

Germany (Wetzell) and in Australia (Cochard et al. [9]; Schreiber et al. [10]), equipped with ring-laser interferometers based on the Sagnac principle. These stations were established, primarily, to record very small deviations and disturbances of the Earth's rotational motion. However, the instruments having sensitivity up to  $10^{-9}$  rad/s were able to record the rotation motions related to many distant earthquakes.

Latter, sensors of another type – the fiber-optic interferometers – were used by Takeo [11] especially for seismic observations; one version of his sensors included the tri-axial system. Jaroszewicz et al. [12] followed this system of rotation seismographs for the study of Silesian seismic events.

In a more traditional way, Moriya (see Moriya and Teisseyre [13]) has constructed the first rotation seismograph system consisting of a pair of anti-parallel seismographs; such a system, with the common suspension of the anti-parallel pendulums was repeated in latter constructions (Wiszniewski [14]).

Data collected by the recording systems mentioned above brought at least two important results:

- Records of different events in the very near field indicate that some events, e. g., shallow volcanic and those of explosion type, differ from the common characteristics by the extremely small rotation components (Teisseyre et al. [15]).
- Correlations between the rotation seismograms obtained from the ring-laser system and the rotation motions, curl  $\mathbf{u}$ , derived from the array of seismometers (located at the same site) show almost perfect fit (Cochard et al. [9]).

Following the Cosserat theory and the micropolar and micromorphic theories (see Subsect. “**Asymmetric Continuum Theory**”), the independent rotation field, e. g., rotation related to grains or points of a continuum, were considered by Shimbo [1,2] in relation to the friction and fracture processes; the related constitutive law, we will call it the Shimbo law, joins the antisymmetric stresses with rotations and leads us towards the asymmetric continuum theory. The Shimbo law was latter generalized for the spin and twist rotation motions (Teisseyre et al. [16]; twist motion is introduced as the equivalent to oscillations of the shear axes).

We shall remind the reader that Gutenberg [7], in the frame of the classical elasticity theory, has proved that the independent rotation waves must be immediately attenuated. Now we know that this statement is due only to the fact that in the classical theory the rotations are not related to the stress or stress moment response; we cannot intro-

duce the constitutive law joining the symmetric stresses with the antisymmetric rotations.

However, in the asymmetric continuum theory such a constitutive law is required and appears as a natural element of the theory.

### Displacements and Rotations

We cannot deny that independent rotation waves do not exist in a continuum built by the point-particles; however, this question is reduced to the magnitude of the independent rotations (that is, independent of rot  $\mathbf{u}$ ) generated in the seismic sources; the rotation wave motion is assured by the Shimbo constitutive law; this constitutive law joining the antisymmetric stresses and rotation relates directly to the friction as kind of material resistance:

$$S_{[ik]} = 2\mu^* \omega_{[ik]}, \quad S_{(ik)}^D = 2\mu E_{(ik)}^D, \quad (1)$$

where we have added the constitutive law for pure shear (further on we define the pure shear oscillations as twist), the tensors  $S_{(ik)}^D$  and  $E_{(ik)}^D$  are the deviatoric stress and strain tensors, and where  $\mu$  is the rigidity modulus and the constant  $\mu^*$  is defined as rotation rigidity, the constant entering in the antisymmetric stress – angular velocity relation (this constant is not equal to the rotation modulus in the stress moment – angular velocity relation); further on, we assume that both constants are equal,  $\mu^* = \mu$ , as it follows from the seismic wave observations.

We shall notice that both motions, displacements and rotations, are interrelated, which follows also from the fact that pure rotation,  $\omega_{[s]}$ , can be presented by means of the potentials represented by some displacement field,  $\mathbf{u}^{\text{micro}}$ ; conversely, the displacements  $U$  (excluding those related to the scalar potential, e. g., those of thermal origin or that related to explosion process, that is, we put  $\frac{\partial}{\partial x_i} U_i = 0$ ) can be described by the vector potentials represented by some rotation field  $\Omega_{[.]}$ :

$$\omega_{[.]} = \text{curl } U, \quad U = l^2 \text{curl } \Omega_{[.]} \quad (2)$$

and we arrive at the possible dual approach to the continuum mechanics.

When applying such equivalent approaches twice, e. g., from rotation field to displacement and again to rotations we obtain

$$\left\{ \begin{aligned} \omega_{[s]} &= \varepsilon_{smn} \frac{\partial U_n}{\partial x_m}, \quad u_i = l^2 \varepsilon_{iks} \frac{\partial \omega_s}{\partial x_k} \\ \rightarrow u_i &= -l^2 \Delta U_i \quad \text{at} \quad \frac{\partial}{\partial x_i} U_i = 0. \end{aligned} \right. \quad (3a)$$



Or otherwise:

$$\left\{ \begin{aligned} \omega_{[i]} &= \varepsilon_{iks} \frac{\partial U_s}{\partial x_k}, \quad U_s = l^2 \varepsilon_{smn} \frac{\partial \Omega_n}{\partial x_m} \\ \rightarrow \omega_{[i]} &= -l^2 \Delta \Omega_i \quad \text{at} \quad \frac{\partial}{\partial x_i} \Omega_i = 0, \end{aligned} \right. \quad (3b)$$

where  $l$  represents the basic intrinsic length measure.

The intrinsic length (Cosserat characteristic length) plays an important role in material properties; there is extensive literature related to this subject, however we limit ourselves to the remark that the displacement and rotation motions could be completely independent only for the case with  $l = 0$ , but such a case is excluded by quantum mechanics with the minimal Planck's length of the order of  $10^{-34}$  m.

Excluding the axial motions, our consideration leads us to an apparent equivalence of the two descriptions, those by means of displacements and rotations. However, there remains the problem of the scales of these motions generated at the different fracture modes in the seismic source and also the scales of these motions observed at the Earth's surface.

The observed rotation fields and effects are usually much smaller than those related to the displacement field. There are, however, some exceptions leading to situations in which rotations play an important role, e. g.: the tilt motions, the rocking and tilting components related to building structures hit by strong ground motions and, as it will be discussed further on, the rotations generated by fracture which occurred under the prevailing compression load.

These equivalent descriptions of the motions in a continuum can be combined in the asymmetric theory, but we shall note that it could be constructed as well, the continuum theory neglecting completely the displacement fields and using only the rotation motions – such a case can be called a degenerated continuum.

Therefore, keeping in mind the above remarks that the rotations contribute to displacement derivatives and that displacements may contribute to rotations, we can state that these motions are interrelated; this statement is empirically supported by the above mentioned almost perfect fit between the derived rotations, curl  $\mathbf{u}$ , and the rotations obtained from the ring-laser system data. Therefore, we think that the problem related to existence of rotation waves appears as an irrelevant question. However, we shall stress that the displacements and rotation motions differ in general, especially when considering their physical origins and effects. Instead of that problem, we propose to consider the classification of rotation motions from the point of view of their origins, scales and effects produced.

We propose the following classification:

- The micro-rotations or rotations,  $\omega$ , as related to the wave motions based on the internal friction processes (rotation rigidity), as well as to slip motions with friction/fracture processes
- The meso-rotations related to material granulation and formation of the mylonite zones under the shear load fracturing processes
- The total rotations,  $\omega^T$ , (the nomenclature introduced by Kröner [17]) related to the displacement field,  $\mathbf{u}$ ;
- The macro-rotations as related to fragmentation of material at the fracturing under compression load
- The mega-rotation effects related to the ground tilts and tilting of high objects on the ground

The important counterpart of rotational processes in the mechanics of fracturing and the related energy release shall be underlined (Teisseyre et al. [18]). Both under confining pressure and under external shears, the role of micro-fracturing in the bond breaking process is similar; however, we observe here the essential differences for rotations in larger scales.

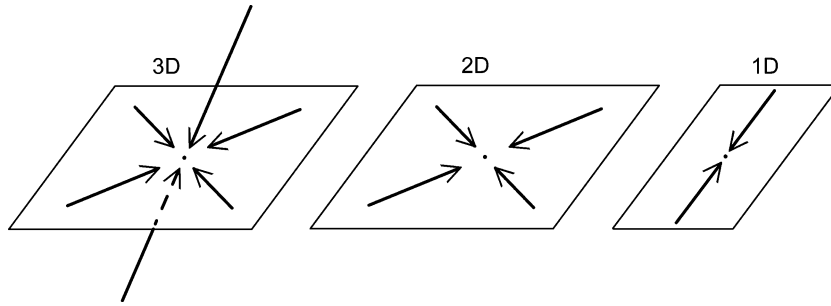
The confining load condition leads to formation of the induced opposite arrays of dislocations, resulting in fragmentation processes and related macro-rotations. On the other hand, shear load leads to more concentrated fracturing along some planes. In the thermodynamical fracture band theory, see Subsect. “**Earthquake Thermodynamics**”, we consider the additional super-lattice formed by dislocations and the properly defined vacant dislocations; with this advanced approach we express an effective role of dislocation band structure in the shear fracture thermodynamics. Similarly, the fragmentation and macro-rotation processes become more effective for the fracturing under confining pressure. Thus, we try to find a continuum description for a number of processes leading us from an elastic solid to that undergoing successive deterioration by crushing, granulation and fragmentation.

These considerations give us ground for the classification of the basic motions.

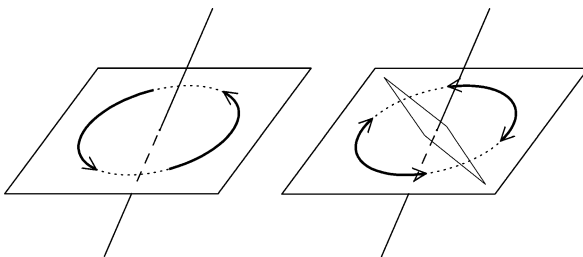
### Basic Deformations and Simple Motions in Asymmetric Continuum

Basic and simple motions could be defined as those which may be reduced to the 3D point motion in the Cartesian or Riemann spaces, or those deformations conceived as the respective curvatures.

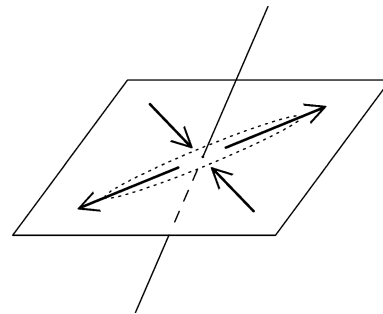
Considering basic motions, we can distinguish the simple motions. First the translation described by vector  $\mathbf{u}$  and the independent rotation, called spin; in a non-homogeneous continuum, the grains having different material parameters, can rotate due to an interaction of a displace-



Earthquake Source: Asymmetry and Rotation Effects, Figure 1  
Axial basic deformations (3D, 2D and 1D)



Earthquake Source: Asymmetry and Rotation Effects, Figure 2  
Rotational motions: spin and twist



Earthquake Source: Asymmetry and Rotation Effects, Figure 3  
String-string nucleus

ment field with the related moment of inertia of the grains. Then, we shall pass to the tensorial motions/deformations:

Any antisymmetric tensor can be related to the vectorial field, e. g., to spin motion; thus, we come again to the equivalent vector field. However, this simple spin motion,  $\omega_{[.]}$ , shall be treated basically as independent of the displacement rotation, however, both contribute to the total spin field,  $\omega_{[.]} + \text{curl } \mathbf{u}$ , observed, e. g. in seismology. We have already mentioned that any symmetric tensor can be split into the axial and deviatoric tensors. The axial deformation tensor relates to the point deformations representing compression/dilatation nuclei, e. g., related to thermal anomaly.

These total axial oscillation motions include the equal translation motions along all three axes and relate to that part of the displacement field which can be derived from a scalar potential (see Fig. 1).

For the point-like continuum with the axial deformations, we would obtain either the Riemannian curvature or, for more complicated cases, the Riemannian torsion tensor.

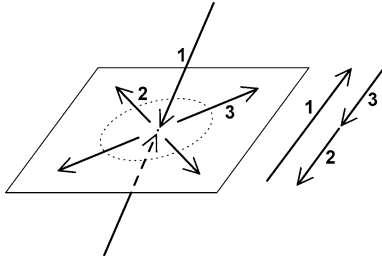
There remains the deviatoric field to consider. This field relates to pure shear deformations; it is possible to show that the deviatoric field may be used to define the new antisymmetric tensor related to the simple deforma-

tions representing another kind of rotation motion – the twist deformation. These deviatoric deformations, for continua formed by particles/grains relate to pure shear oscillations (see Fig. 2; the left side shows a spin motion and the right side a twist one).

However, considering the point motions, it is better to relate these shear deformations to the equivalent twist tensor,  $\omega_{(ks)}$ ; such motions contribute to the observed, e. g. in seismology, total shear/twist field. This field is directly related to the S-waves ( $\mathbf{u}^S = l\omega$ , where  $l$  is the effective radius of grains/particles forming the continuum and  $\omega$  is the related rotation).

In the limit related to the point-like deformations, we would arrive at the string-string type motions (see Fig. 3) leading to the another representation of twist as vector,  $\omega_{(k)}$ , perpendicular to the string-string plane and having an appropriate magnitude of a string-string deformation (invariant representation of the string-string vector is discussed in Subsect. “Spin and Twist Motions”, while determination of its amplitude Subsect. “Recording Spin and Twist Angle Variation”).

A combination of the axial oscillation motions and those related to twist deformations leads to different forms



Earthquake Source: Asymmetry and Rotation Effects, Figure 4  
String-membrane

of the point deformations: Fig. 4 relates to the string-membrane oscillations.

We shall repeat that all motions, except axial deformation, may be described by the displacement field, but this is not true for the point related spin and string-string deformation originated as the independent source motions.

Of course, there remains a number of the first, and higher order moments of these basic motions and deformations; such source models tend in a limit to a point source, but cannot be treated as exact point nuclei.

### Asymmetric Continuum Theory

We shall add to the above considerations some arguments forming the fundamentals of the asymmetric continuum theory which is based on the asymmetric stresses:

- When studying the elastic field of an edge dislocation, we find some asymmetry in relation to its components in the plane perpendicular to its line (wedge direction); in confrontation with the symmetry of shears, this fact results in the asymmetry of stresses for a continuous distribution of dislocations (for screw dislocations such a contradiction does not exist). Therefore, a direct differential relation between any density of dislocations and the related stresses cannot be adequately found in a symmetric continuum.
- Fracture usually reveals its asymmetric pattern with a main slip plane; we shall believe that the premonitory processes, as described by deformations in a continuum with defects, develop also in an asymmetric way.
- In the classical continuum, the balance of angular momentum holds only if the stresses are symmetric; here the angular motions can be introduced only artificially with the help of a length element and a reference rotation point. This classical theory has also many other limitations and therefore many trials have been undertaken to generalize it. The asymmetric theory of elasticity with asymmetric stresses and couple-stresses was founded by Nowacki [19]. However, a first generalization to include the moments in a continuum is due to Voigt in 1887 and a complete theory, including the asymmetry of stress and strain, is that known as the Cosserat theory of elasticity with the displacement vector and rotation vector [20]. Micropolar and micromorphic elastic theories were developed by Eringen and his co-workers and Mindlin (see: Eringen and Suhubi [21] and Mindlin [23]). Teisseyre [22] proposed a simpler version of the asymmetric theory which includes asymmetric stresses, strains and rotations, but in which the equations for the antisymmetric stresses differ from those of the couple moments in the Nowacki theory; their roles are interchanged, but both systems remain almost equivalent.
- Usually, when searching the fault slip solutions, we rely on classical elasticity with the friction constitutive laws introduced additionally in accordance with the experimental data. The obtained results well explain the observational data. Instead, we consider the consistent elastic continuum with asymmetric stresses and defects; such an approach enables one to study the defect interactions and elastodynamic solutions describing a slip propagation along a fault, including friction effects and related seismic radiation.
- The asymmetry of fields follows also from the notion of antisymmetric stresses considered by Shimbo [1,2] in relation to the friction processes and rotation of grains. Fracture processes develop usually along the main fault plane; hence there appears the initial asymmetry of the fracture pattern [24]; because of friction, the rotation of grains adjacent to the main slip plane causes an appearance of the antisymmetric part of the stresses and twist deformations. Following Shimbo [1,2], we introduced the constitutive law joining the antisymmetric stresses with the rotation nuclei (self-rotation field); without such a constitutive law any theory reduces both the rotation motions (except the rotation of displacements) and the related rotation waves to zero.
- In the asymmetric continuum, defined as that including both the symmetric stresses and the antisymmetric stresses, there appear also the rotational motions/deformations which split into pure spin and twist motions, the latter relate to the shear deformations of the grains; when considering the point-like nuclei, the twist deformation passes into 3D space torsion (Riemannian space).
- Experimental evidence for an appearance of spin and twist motions in a seismic field is based on the records of seismic rotation fields. For spin motion we shall be aware that the recorded rotation contains two el-

ements: a rotation of displacements and an independent spin motion. Both these elements co-act in motion propagation and represent its dual description, but differ in their origin, depending on the source processes and material properties. At fracturing under a confining load, we deal rather with a high rotation release process and therefore the spin motion for the very near seismic events usually distinctly overpasses the displacement rotation. For some events of an explosive nature, or for some near-surface volcanic events, both the pure spin motion and rotation of displacement almost disappear; some observed effects might be related to a nearby  $P \Leftrightarrow S$  conversion. For the strong motions which include a tilting component, the rotation of displacements exceeds a spin motion. In engineering seismology we observe that the rotation of displacements may exceed the pure spin motion; such an effect is due to magnification of a horizontal rotation of displacements and to the appearance of a rocking/tilting component of displacement rotation caused by the geometry of construction, especially for high buildings.

### Self-Field Nuclei: Deviations from Classical Elasticity

Any continuum could be described using the Kröner approach [17] based on a concept of internal fields excited by a density of defects and internal nuclei; stresses and strains are related by the unique constitutive law for the ideal elasticity. This approach is equivalent to another approach in which we change the constitutive law in a way appropriate to describe the plastic, viscous and relaxation effects. In the Kröner continuum with a density of the internal point-like nuclei, the elastic strains, rotations and stresses can be expressed as differences between total and self-fields.

Following the Kröner approach, we can keep the ideal elastic relation for the stresses and strains, supplemented by the constitutive law joining the antisymmetric stresses with rotations, and we introduce the self/inner stresses, strains and rotations as related to the internal nuclei or defects:  $\mathbf{S}^S$ ,  $\mathbf{E}^S$ ,  $\omega^S$ . We distinguish between the total stresses, strains and rotations related to the displacement field:  $\mathbf{S}^T$ ,  $\mathbf{E}^T$ ,  $\omega^T$ , and the asymmetric elastic stresses, strains and rotations  $\mathbf{S}$ ,  $\mathbf{E}$ ,  $\omega$ :

$$\begin{aligned} \mathbf{E} &= \mathbf{E}^T - \mathbf{E}^S, & \omega &= \omega^T - \omega^S, & \beta &= \beta^T - \beta^S, \\ \mathbf{S} &= \mathbf{S}^T - \mathbf{S}^S, \end{aligned} \quad (4a)$$

$$\mathbf{E}_{ki}^T = \mathbf{u}_{(i,k)}, \quad \omega_{i,k}^T = \mathbf{u}_{[ik]}. \quad (4b)$$

The elastic and self-deformations, strains and rotations, and stresses can be, in general, asymmetric ones (see: Teis-

seyre and Boratyński [4]); under the conditions that the antisymmetric parts of the stresses and strains, as well as the symmetric parts for elastic and self-rotations, be mutually compensated for:

$$\mathbf{E}_{[ik]} + \mathbf{E}_{[ik]}^S = 0, \quad \mathbf{S}_{[ik]} + \mathbf{S}_{[ik]}^S = 0, \quad \omega_{(ik)} + \omega_{(ik)}^S = 0. \quad (5)$$

However, referring to our earlier papers (see: Teisseyre and Boratyński [4]) we shall then assume that the respective self-parts of the asymmetric strain and rotation are equal to each other:

$$E_{[ik]}^S = \omega_{[ik]}^S, \quad \omega_{(ik)}^S = E_{(ik)}^S, \quad (6)$$

where symmetric rotation is related to the shear axes oscillations (comp.: twist definition and Figs. 2 and 3). The elastic fields  $\mathbf{S}$ ,  $\mathbf{E}$ ,  $\omega$  represent the physical fields, while the total fields  $\mathbf{S}^T$ ,  $\mathbf{E}^T$ ,  $\omega^T$  relate, according to the compatibility condition, to the displacement motions  $u_i$ , and the self-fields relate to the internal nuclei, defect densities and continuum structure.

Any deviations from the symmetry properties of fields, and any deviations from the ideal elasticity, can be described by suitable forms of the self-field, represented by the internal nuclei for both the defects and interaction fields.

The defects, dislocation and disclination densities can be defined, following Kossecka and De Witt [25], by considering the total disclosure and twist along a closed circuit (the Burgers vector and the Frank vector) and the appropriate form of the twist-bend tensor:

$$\begin{aligned} B_l &= -\oint [E_{(kl)} - \varepsilon_{lqr} \chi_{kq}^S x_r] dl_k, \\ \Omega_q &= -\oint \chi_{kq}^S dl_k = \theta_{pq} ds_p \end{aligned} \quad (7)$$

and the definitions of the dislocation and disclination densities,  $\alpha$  and  $\theta$ , become based on the self-fields  $E_{(kl)}^S$  and  $\chi_{kq}^S$ :

$$\begin{aligned} \alpha_{pl} &= -\varepsilon_{pmk} \left( \frac{\partial E_{(kl)}^S}{\partial x_m} + \varepsilon_{klq} \chi_{mq}^S \right), \\ \theta_{pq} &= -\varepsilon_{pmk} \frac{\partial \chi_{kq}^S}{\partial x_m}. \end{aligned} \quad (8)$$

After Teisseyre [26], the total twist-bend tensor can be defined as follows:

$$\chi_{mq}^T = \varepsilon_{ksq} \frac{\partial \omega_{mk}^T}{\partial x_s}, \quad \chi_{mq}^T = \chi_{mq} + \chi_{mq}^S \quad (9)$$

where, for the continuum with the asymmetric part of the stresses, we are not restricted to the compatibility condition for the twist-bend tensor (Kleman [27]).

The compatibility conditions for the asymmetric stresses and strains lead us to the physical equations for the dislocation and disclination densities in relation to the elastic fields of strain  $E_{(kl)}$  and twist-bend  $\chi_{kq}$ :

$$\alpha_{pl} = \varepsilon_{pmk} \left( \frac{\partial E_{(kl)}}{\partial x_m} + \varepsilon_{klq} \chi_{mq} \right), \quad \theta_{pq} = \varepsilon_{pmk} \frac{\partial \chi_{kq}}{\partial x_m}. \tag{10}$$

Further on we will not rely on the Kröner approach, instead we will confine ourselves to a simpler approach given by the standard asymmetric continuum theory.

### Asymmetric Continuum: Standard Theory

In opposition to the Kröner approach presented above, we may construct the asymmetric standard theory entirely related to the displacement field. Such a theory shall be based both on the symmetric and asymmetric stresses and on the related constitutive laws and motion equations. The asymmetric deformations contain the symmetric strain and antisymmetric rotation. Thus, our theory is based on two groups of relations; for the symmetric and antisymmetric fields:

$$\begin{aligned} S_{kl} &= S_{(kl)} + S_{[kl]}, & E_{kl} &= E_{(kl)}, & \omega_{kl} &= \omega_{[kl]}, \\ D_{ks} &= E_{ks} + \omega_{ks}, \end{aligned} \tag{11a}$$

where  $D_{ks}$  means the asymmetric deformation tensor.

However, when introducing the new material parameters (material structure indices):  $e^0, \chi^0$ , we may join these deformation fields in an independent way, with some reference displacement motion:

$$E_{kl} = e^0 \frac{1}{2} \left( \frac{\partial u_l}{\partial x_k} + \frac{\partial u_k}{\partial x_l} \right), \quad \omega_{kl} = \chi^0 \frac{1}{2} \left( \frac{\partial u_l}{\partial x_k} - \frac{\partial u_k}{\partial x_l} \right). \tag{11b}$$

For an internal energy stored in such a medium we obtain:

$$E = S_{(ks)} E_{ks} + S_{[ks]} \omega_{ks}.$$

The indices  $e^0, \chi^0$  are not new constitutive constants, but they define families of solutions describing the complexity of deformation processes in continua; their ratio determines the phase shift between strain and rotation tensors. In this sense, the strain and rotation can be shifted in phase as follows from the particular deformations considered.

For the particular cases of these index values,  $e^0, \chi^0$ , we have:

- The classic elasticity, obtained for  $\chi^0 = 0$ ;
- For  $e^0 = 0$  we obtain a granular/crushed medium filled with rigid spheres interacting by friction; when applying a torque load on its surface boundary, e.g., a cylindrical one, we would obtain only some angular deformation, and torque energy stored given as  $E = S_{[ks]} \omega_{ks}$ ;
- The cases with  $e^0 = \chi^0$  relate to the elastic continua with friction and different kinds of internal defects – different kinds of dislocation densities and the granulated materials;
- A continuum densely filled with the edge dislocations is described by the case  $e^0 = -1, \chi^0 = -1$ ; while that of a partial content of that density  $\alpha^E = \{0, 1\}$  by  $e^0 = 1 - 2\alpha^E, \chi^0 = \alpha^E$ ;
- A continuum filled densely with the screw dislocations would be given by  $e^0 = 2, \chi^0 = 2$ ; while that of a partial content of that density  $\alpha^S = \{0, 1\}$  by  $e^0 = 1 + \alpha^S, \chi^0 = 2\alpha^S$ .

Further on, we will consider a more general continuum with the constitutive laws, including also the time rates processes; for such cases we might discuss in a similar way the different particular cases of the material structure indices including dynamic objects.

For the symmetric part of stresses we can assume the classical constitutive relation:

$$S_{(kl)} = \lambda \delta_{kl} E_{ss} + 2\mu E_{kl}. \tag{12}$$

But there is no problem to include in it the appropriate linear deviations related to visco-plastic effects.

To construct the asymmetric theory, we assume, after Shimbo [1,2], the appropriate constitutive law for the antisymmetric part of stresses. It joins the friction/fracture rotations with the antisymmetric stresses:

$$S_{[kl]} = 2\mu \omega_{kl}, \tag{13}$$

where rigidity constant  $\mu$  plays the role of rotation rigidity entering in the antisymmetric stress-angular velocity relation (while the rotation modulus enters in the stress moment-angular velocity relation and may be considered as product of the rigidity  $\mu$  and the Cosserat characteristic length  $l$ ).

The motion equation for antisymmetric stresses  $S_{[ni]}$  shall replace the balance law for the stress moments. To this end, we take the divergence of the rotation force moment acting on a body element due to the antisymmetric stresses (rotational moment of forces per infinitesimal arm length corresponding to stress moments), and, on the other hand, the balancing term – the acceleration related

to angular momentum [3]:

$$\begin{aligned} \varepsilon_{lki} \frac{\partial^2}{\partial x_k \partial x_n} S_{[ni]} &= \rho \frac{1}{2} \varepsilon_{lki} \frac{\partial^2}{\partial t^2} \left( \frac{\partial u_i}{\partial x_k} - \frac{\partial u_k}{\partial x_i} \right) \\ &+ \varepsilon_{lki} \rho K_{[ki]}, \end{aligned} \quad (14)$$

where we have put  $e^0 = 1$  and we have introduced the body couple  $K_{[ki]}$  equivalent to body moment  $K_{[l]} = \varepsilon_{lki} \rho K_{[ki]}$ .

With the compatibility condition introduced in a similar way as for the symmetric strains:

$$I_{[ij]} = \varepsilon_{imk} \varepsilon_{jns} \frac{\partial^2}{\partial x_m \partial x_n} \omega_{ks} = 0$$

we obtain from Eqs. (11b), (13) and (14):

$$\begin{aligned} \frac{\partial^2 S_{[ki]}}{\partial x_s \partial x_s} &= 2\rho \frac{\partial^2 \omega_{ki}}{\partial t^2} + 2\rho K_{[ki]}, \quad \text{or} \\ \mu \frac{\partial^2 \omega_{ki}}{\partial x_s \partial x_s} - \rho \frac{\partial^2 \omega_{ki}}{\partial t^2} &= \rho K_{[ki]}, \end{aligned} \quad (15a)$$

where we have introduced also the body couple  $K_{[ki]}$  or body moment  $K_{[l]} = \varepsilon_{lki} \rho K_{[ki]}$ .

Otherwise, we can write:

$$\mu \frac{\partial^2}{\partial x_k \partial x_k} \omega_{[l]} - \rho \frac{\partial^2}{\partial t^2} \omega_{[l]} = \rho K_{[l]}, \quad (15b)$$

where the left-hand side of this form presents the basic expression for the resulting stress moment divergence.

These relations are equivalent to the following ones:

$$\begin{aligned} \frac{1}{l^2} \frac{\partial}{\partial x_k} M_{lk} &= \varepsilon_{lki} \frac{\partial^2}{\partial x_k \partial x_n} S_{[ni]} = \varepsilon_{lki} \frac{\partial}{\partial x_n} \frac{\partial}{\partial x_n} S_{[ki]}, \\ \frac{1}{l^2} M_{lk} &= \varepsilon_{lki} \frac{\partial}{\partial x_n} S_{[ni]} \end{aligned}$$

or defining the angular moment  $\Xi_i$ , we obtain:

$$\frac{\partial}{\partial x_k} M_{ik} = 2\mu \Xi_i, \quad \Xi_i = l^2 \varepsilon_{iks} \frac{\partial}{\partial x_n} \frac{\partial}{\partial x_s} \omega_{[kn]}.$$

From the motion equation for the symmetric part of stresses

$$\frac{\partial}{\partial x_k} S_{(kl)} = \rho \frac{\partial^2}{\partial t^2} u_l + F_l$$

and using the scalar and vector potentials

$$\begin{aligned} u_l &= l^2 \frac{\partial}{\partial x_l} \Phi + l^2 \varepsilon_{lps} \frac{\partial}{\partial x_p} \Psi_s, \\ F_l &= l^2 \frac{\partial}{\partial x_l} \Phi + l^2 \varepsilon_{lps} \frac{\partial}{\partial x_p} \Psi_s \end{aligned}$$

we obtain:

$$\begin{aligned} (\lambda + 2\mu) \frac{\partial^2}{\partial x_k \partial x_k} \varphi &= \rho \ddot{\varphi} + \Phi, \\ \mu \frac{\partial^2}{\partial x_k \partial x_k} \psi_s &= \rho \ddot{\psi}_s + \Psi_s, \end{aligned} \quad (16)$$

where according to Eq. (11b) we have introduced the index  $e^0$  and we assume  $\frac{\partial}{\partial x_s} \psi_s = 0$ ,  $\frac{\partial}{\partial x_s} \Psi_s = 0$  and where we have introduced the intrinsic length unit  $l$ .

Here the potential  $\psi_s$  may be interpreted as rotation vector motions in another scale than that defined by relation  $\text{curl } v = \omega$ :

$$\varepsilon_{mql} \frac{\partial}{\partial x_q} u_l = -l^2 \frac{\partial^2}{\partial x_k \partial x_k} \psi_m,$$

where  $\omega$  means the micro-rotations, and  $\psi$  – the meso-rotations related to the granulated material (mylonite) and shear processes.

The strain tensor and its trace can be presented with the help of the introduced potentials as follows:

$$\begin{aligned} E_{lq} &= l^2 \frac{\partial^2}{\partial x_l \partial x_q} \varphi + \frac{1}{2} l^2 \varepsilon_{lps} \frac{\partial^2}{\partial x_p \partial x_q} \psi_s \\ &+ \frac{1}{2} l^2 \varepsilon_{qps} \frac{\partial^2}{\partial x_p \partial x_l} \psi_s. \end{aligned}$$

We can divide this expression into the axial and deviatoric parts:

$$\begin{aligned} E_{kk} &= l^2 \frac{\partial^2 \varphi}{\partial x_s \partial x_s}, \\ E_{lq}^D &= l^2 \left( \frac{\partial^2 \varphi}{\partial x_l \partial x_q} - \frac{\delta_{lq}}{3} \frac{\partial^2 \varphi}{\partial x_s \partial x_s} \right. \\ &\quad \left. + \frac{1}{2} \frac{\partial}{\partial x_p} \left( \varepsilon_{lps} \frac{\partial}{\partial x_q} + \varepsilon_{qps} \frac{\partial}{\partial x_l} \right) \psi_s \right). \end{aligned} \quad (17)$$

Returning to our wave Eqs. (16) we arrive at the wave equations for the axial and deviatoric strain parts:

$$(\lambda + 2\mu) \Delta E_{kk} - \rho \frac{\partial^2 E_{kk}}{\partial t^2} = l^2 \Delta \Phi, \quad (18a)$$

$$\begin{aligned} (\lambda + \mu) \left( \frac{\partial^2 E_{ss}}{\partial x_l \partial x_q} - \frac{\delta_{lq}}{3} \frac{\partial^2 E_{ss}}{\partial x_k \partial x_k} \right) \\ + \mu \frac{\partial^2 E_{lq}^D}{\partial x_k \partial x_k} - \rho \frac{\partial^2 E_{lq}^D}{\partial t^2} \\ = l^2 \left( \frac{\partial^2 \Phi}{\partial x_l \partial x_q} - \frac{\delta_{lq} \Delta \Phi}{3} \right. \\ \left. + \frac{\varepsilon_{lps}}{2} \frac{\partial^2 \Psi}{\partial x_p \partial x_q} + \frac{\varepsilon_{qps}}{2} \frac{\partial^2 \Psi}{\partial x_p \partial x_l} \right). \end{aligned} \quad (18b)$$

In terms of the potentials we obtain

$$\begin{aligned} & \left( \mu \Delta - \rho \frac{\partial^2}{\partial t^2} \right) \left( \left( \frac{\partial^2}{\partial x_l \partial x_q} - \frac{\delta_{lq}}{3} \Delta \right) \varphi \right. \\ & \quad \left. + \frac{\partial}{2 \partial x_p} \left( \varepsilon_{lps} \frac{\partial}{\partial x_q} + \varepsilon_{qps} \frac{\partial}{\partial x_l} \right) \psi_s \right) \\ & = \left( \left( \frac{\partial^2}{\partial x_l \partial x_q} - \frac{\delta_{lq}}{3} \Delta \right) \Phi \right. \\ & \quad \left. + \frac{\partial}{2 \partial x_p} \left( \varepsilon_{lps} \frac{\partial}{\partial x_q} + \varepsilon_{qps} \frac{\partial}{\partial x_l} \right) \Psi_s \right) \end{aligned}$$

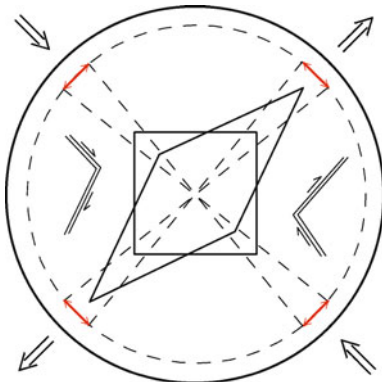
### Spin and Twist Motions

The spin motion is governed by Eq. (15a), or equivalently by its vector form Eq. (15b). We may consider the system related to the main shear axes or that related only to the off-diagonal components; in the latter case the motion equation for the deviatoric strains Eq. (18b) can be presented in the form of the rotation vector motion – the twist,  $\omega_{(s)}$ :

$$\{\omega_{(s)}\} = \{E_{23}^D, E_{31}^D, E_{12}^D\}. \quad (19a)$$

The defined twist motion,  $\omega_{(s)}$ , means the rotational oscillation of the off-diagonal shear axes of the deviatoric tensor (corresponding to oscillation of the main shear axes),  $E_{lq}^D$ , accompanied by the changes of the shear magnitude; such perturbation of the shear load may be caused by the internal fracturing processes (see Fig. 5).

Once having defined the twist vector field we can maintain its form due to the invariant properties of the



Earthquake Source: Asymmetry and Rotation Effects, Figure 5  
Twist motion: rotational oscillations of the off-diagonal shear axes and internal fractures as the sources of perturbations; in the center we present an external shear deformation, while arrows along the circle give possible oscillations of the shear axes as influenced by some intrinsic processes, e. g. the fractures marked inside

Dirac tensors applied to the symmetric off-diagonal tensor  $\omega_{(ik)}$  in its 4D form:

$$\begin{aligned} \omega_{(\lambda\kappa)} & = \omega_{(1)}\gamma^1 + \omega_{(2)}\gamma^2 + \omega_{(3)}\gamma^4\gamma^2\gamma^3 \\ & = \begin{bmatrix} 0 & -\omega_{(3)} & -\omega_{(2)} & -\omega_{(1)} \\ -\omega_{(3)} & 0 & \omega_{(1)} & -\omega_{(2)} \\ -\omega_{(2)} & \omega_{(1)} & 0 & -\omega_{(3)} \\ -\omega_{(1)} & -\omega_{(2)} & -\omega_{(3)} & 0 \end{bmatrix}, \end{aligned} \quad (19b)$$

where

$$\begin{aligned} \gamma^1 & = \begin{bmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{bmatrix}, \\ \gamma^2 & = \begin{bmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix}, \\ \gamma^3 & = i \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{bmatrix}, \\ \gamma^4 & = i \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}, \end{aligned} \quad (19c)$$

and

$$\gamma^4\gamma^2\gamma^3 = \begin{bmatrix} 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \end{bmatrix}.$$

In a similar way, we may define the external off-diagonal part of the right-side expression of Eq. (18b):

$$\begin{aligned} \mathbf{Y}_{(lq)} & = \\ & l^2 \left( \frac{\partial^2}{\partial x_l \partial x_q} \Phi + \frac{\partial}{2 \partial x_p} \left( \varepsilon_{lps} \frac{\partial}{\partial x_q} + \varepsilon_{qps} \frac{\partial}{\partial x_l} \right) \Psi_s \right). \end{aligned}$$

For its 4D form we can write:

$$\begin{aligned} \mathbf{Y}_{(\lambda\kappa)} & = \mathbf{Y}_{(12)}\gamma^1 + \mathbf{Y}_{(13)}\gamma^2 + \mathbf{Y}_{(23)}\gamma^4\gamma^2\gamma^3 \\ & = \begin{bmatrix} 0 & -\mathbf{Y}_{(12)} & -\mathbf{Y}_{(13)} & -\mathbf{Y}_{(23)} \\ -\mathbf{Y}_{(12)} & 0 & \mathbf{Y}_{(23)} & -\mathbf{Y}_{(13)} \\ -\mathbf{Y}_{(13)} & \mathbf{Y}_{(23)} & 0 & -\mathbf{Y}_{(12)} \\ -\mathbf{Y}_{(23)} & -\mathbf{Y}_{(13)} & -\mathbf{Y}_{(12)} & 0 \end{bmatrix}. \end{aligned}$$

Using these definitions for the off-diagonal form Eq. (19b) we obtain

$$\mu \frac{\partial^2 \omega_{(\lambda\kappa)}}{\partial x_k \partial x_k} - \rho \frac{\partial^2 \omega_{(\lambda\kappa)}}{\partial t^2} = \mathbf{Y}_{(\lambda\kappa)}. \quad (20)$$

The defined 4D twist motion,  $\omega_{(\lambda\kappa)}$ , means the rotational oscillation of the off-diagonal shear axes of the deviatoric tensor,  $E_{lq}^D$ , accompanied by the changes of the shear magnitude; such perturbation of the shear load may be caused by internal fracturing processes (Fig. 5).

The spin and twist motions form the complex rotation field defined as:

$$\omega_s = \omega_{[s]} + i\omega_{(s)} \tag{21}$$

From the balance relation (see: Subsect. “Recording Spin and Twist Angle Variation”) we obtain the relations joining the spin and twist motions.

**Defects: Dislocation and Disclination Densities**

The classical approach to the dislocation and disclination densities is based on the Kröner description of continuum with the self-fields (compare Subsect. “Self-field Nuclei: Deviations from Classical Elasticity”). In the asymmetric homogeneous continuum, the defect density can be introduced using the modified definition of disclosure,  $B_l$ , and the following definition of the twist-bend vector (compare Eqs. (7-9)) we define:

$$B_l = \oint [E_{kl} - \omega_{kl}] dl_k, \quad \Omega_q = \oint \chi_{kq}^T dl_k = \iint \theta_{kq} ds_k, \tag{22a}$$

where for

$$\chi_{mq}^T = \varepsilon_{ksq} \frac{\partial \omega_{mk}}{\partial x_s} \tag{22b}$$

the disclination density vanishes due to the compatibility conditions:

$$\theta_{pq} = \varepsilon_{pmk} \frac{\partial \chi_{kq}^T}{\partial x_m} = \chi^0 \varepsilon_{pmk} \varepsilon_{qns} \frac{\partial^2 \omega_{ks}}{\partial x_m \partial x_n} = 0.$$

For the dislocation field we obtain (compare Eq. (8)):

$$\begin{aligned} \alpha_{pl} &= \varepsilon_{pmk} \left( \frac{\partial E_{kl}}{\partial x_m} - \frac{\partial \omega_{kl}}{\partial x_m} \right) \\ &= \varepsilon_{pmk} \frac{\partial}{\partial x_m} \left( \frac{e^0}{2} \left( \frac{\partial u_l}{\partial x_k} + \frac{\partial u_k}{\partial x_l} \right) - \frac{\chi^0}{2} \left( \frac{\partial u_l}{\partial x_k} - \frac{\partial u_k}{\partial x_l} \right) \right). \end{aligned} \tag{23}$$

With the help of the constitutive relations (12) and (13) we arrive at the relation between the dislocation density and asymmetric stresses:

$$\alpha_{pl} = \frac{\varepsilon_{pmk}}{2\mu} \frac{\partial}{\partial x_m} \left( (S_{(kl)} - \frac{\nu}{1+\nu} \delta_{kl} S_{ii}) - S_{[kl]} \right). \tag{24}$$

We shall note that the material constants,  $e^0$  and  $\chi^0$ , define the types of defects and types of rotation nuclei; the complex constants will mean the constant phase shift between the fields.

Note that in the classic theory with defects, we distinguish also the different definitions for a dislocation field, e. g., the Burgers and Nye dislocations (comp: [28]).

For some particular case,  $e^0 = 1, \chi^0 = -1$ : we obtain a vanishing of defects, like:

$$B_l = \oint [E_{kl} - \omega_{kl}] dl_k = \oint \frac{\partial u_l}{\partial x_k} dl_k = 0, \quad \alpha_{pl} = 0. \tag{25}$$

This case may represent an extreme shear deformation.

We give also relations for another simple case,  $e^0 = \chi^0$ , which leads to dislocation density:

$$\begin{aligned} B_l &= \oint [E_{kl} - \omega_{kl}] dl_k = \chi^0 \oint \frac{\partial u_k}{\partial x_l} dl_k, \\ \alpha_{pl} &= \chi^0 \varepsilon_{pmk} \frac{\partial^2 u_k}{\partial x_m \partial x_l}. \end{aligned} \tag{26}$$

We can consider two particular cases, the first giving a relation between the asymmetric stresses and the edge type dislocations ( $e^0 = -1, \chi^0 = -1$ ):

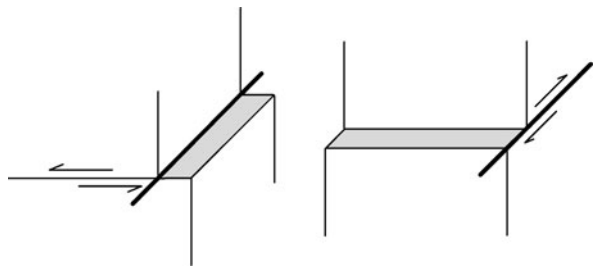
$$\alpha_{pl} = -\varepsilon_{pmk} \frac{\partial^2 u_k}{\partial x_m \partial x_l} \tag{27}$$

and the other which may describe the relation between the screw-type dislocations and asymmetric stresses ( $e^0 = 2, \chi^0 = 2$ ):

$$\alpha_{pl} = 2\varepsilon_{pmk} \frac{\partial^2 u_k}{\partial x_m \partial x_l}, \tag{28}$$

where  $p = l = s$ , no summation over indices  $p$  and  $l$ .

Both of the considered cases could relate to the formation of the respective slip-discontinuities – Fig. 6.



Earthquake Source: Asymmetry and Rotation Effects, Figure 6 The edge and screw dislocation types



We find that for a suitable choice of the disclosure definition, we may arrive at different definitions of the dislocation and disclination densities; in particular, we note that for the edge and screw dislocations we shall consider different rotation nuclei.

The case (Eq. (25)) presents the extreme deformation while the further cases present the standard source models and related rotations.

### Balance Laws for the Rotation Field and the EM Analogy

The complex rotation field (21),  $\omega_s = \omega_{[s]} + i\omega_{(s)}$ , may be presented in the tensor form:

$$\begin{aligned} \varepsilon_{kis}\omega_s &= \omega_{ki} \\ &= \begin{pmatrix} 0 & \omega_{[3]} + i\omega_{(3)} & -\omega_{[2]} - i\omega_{(2)} \\ -\omega_{[3]} - i\omega_{(3)} & 0 & \omega_{[1]} + i\omega_{(1)} \\ \omega_{[2]} + i\omega_{(2)} & -\omega_{[1]} - i\omega_{(1)} & 0 \end{pmatrix}. \end{aligned} \quad (29)$$

We can write the balance condition as

$$\iint \varepsilon_{kps} \frac{\partial}{\partial x_p} \omega_s ds_k = \iint \left( \frac{\partial}{iV\partial t} \omega_k + \frac{4\pi}{\nu} J_k \right) ds_k, \quad (30)$$

where we introduce the current field  $J_k$  and velocity  $V$ .

Hence, we obtain the field equations for the complex rotation motions:

$$\varepsilon_{kps} \frac{\partial}{\partial x_p} \omega_s - \frac{\partial}{iV\partial t} \omega_k = \frac{4\pi}{\nu} J_k \quad (31a)$$

or for spin and twist motions explicitly:

$$\varepsilon_{kps} \frac{\partial \omega_{[s]}}{\partial x_p} - \frac{1}{V} \dot{\omega}_{(k)} = \frac{4\pi}{V} J_k, \quad \varepsilon_{kps} \frac{\partial \omega_{(s)}}{\partial x_p} + \frac{1}{V} \dot{\omega}_{[k]} = 0. \quad (31b)$$

These equations lead us to the related wave forms:

$$\begin{aligned} \Delta \omega_{[n]} - \frac{1}{V^2} \ddot{\omega}_{[n]} &= \frac{4\pi}{V} \varepsilon_{npk} \frac{\partial}{\partial x_p} J_k, \\ \Delta \omega_{(n)} - \frac{1}{V^2} \ddot{\omega}_{(n)} &= 4\pi \frac{\partial}{\partial x_n} \varepsilon - \frac{4\pi}{V^2} \dot{J}_n, \end{aligned} \quad (32)$$

where  $\omega_{[s],s} = 0$  and  $\omega_{(s),s} = 4\pi\varepsilon$ , and under the condition that the velocity,  $V$ , is simultaneously transformed according to relativistic rules for a sum of velocities.

The obtained wave equations coincide with those derived previously (compare Eqs. (15, and 18b) with the definition for twist – Eq. (19)).

We shall note that in the asymmetric elastic continuum, the bonds related to rotational deformations are considered as comparable to those related to elastic rigidity moduli. More complicated situations with the material constant appears in the micropolar and micromorphic theories with the infinitesimally small nuclei (Eringen [21]). In the asymmetric continuum theory, presented in our treatise, the displacements and rotations appear as equally and similarly treated independent fields. Here enter also in a natural way, the axial deformation fields, with a structure similar to that of a thermal field (comp. Eq. (18a)).

Finally, we shall note that these wave fields correspond with  $\nu \rightarrow c$  to the EM fields,  $\omega_{[n]} \rightarrow B_n$ ,  $\omega_{(n)} \rightarrow E_n$ . The form of the rotation complex tensor (Eq. (29)) is fully analogous to the definition of the complex electromagnetic field  $F_s = B_s + iE_s$ .

### Recording Spin and Twist Angle Variation

We shall find the suitable links between the defined fields and experimental data.

The spin motion can be precisely recorded by means of the Sagnac type interferometers (up to  $10^{-9}$  rad/s); there are different types of such systems, e. g., ring laser and fiber optic, to record spin motion.

The angular twist oscillations and shear-twist motions we can record using a system of rotation seismometers. Such a system is based on the rotation seismographs that can record simultaneously the spin and twist angular motions [14,15,29].

In order to obtain the rotation motions, e. g., spin and twist, around the vertical axis we need the data from two parallel horizontal pendulums of opposite orientation. The observations collected clearly indicate that both the mean values of the spin and those of twist angular motions show the seismic oscillations with the same order of magnitudes [13,15,29].

We stress that the twist field, measured in this way gives only the angular variations of the off-diagonal axes of shears (19a); however, we may note that both the spin motion and the twist variation are mutually joint (see Eq. (31)) and therefore, we might theoretically derive knowledge of the shear state from the spin observations.

We shall add that when measuring the shear deformations with the help of a system of strainmeters, we can achieve more reliable and independent data on the shear-twist variations. Moreover, the strainmeter system can measure also the axial deformations.

Finally, we shall reply to the question of how we could compare the invariant twist field (19a and 19b) with the

observed shear variations. An exact procedure requires the following: the 6 components of the shear strain determined in an observation site system shall be transformed, at each time moment, into the off-diagonal system:

$$\{E_{11}, E_{22}, E_{33}, E_{23}, E_{31}, E_{12}\} \rightarrow \{E_{23}^D, E_{31}^D, E_{12}^D\} \\ = \{\omega_{(s)}\}.$$

## Conclusions

In the standard asymmetric continuum theory the defects defined are not the material defects, but only those related to the structural deformations. This standard asymmetric theory permits one to find the differential relation between the dislocation density and the asymmetric stress field. Moreover, in this theory we may consider also other deviations related to other defects or interaction fields; to this end we could apply the Kröner approach with the elastic, self- and total fields [4].

The other important conclusion is that the influence of rotational processes in earthquake sources spreads outward, because these waves are not attenuated strongly, as it was believed according to classical ideal elasticity.

## Earthquake Source: Fracture Processes

### Introduction

We start our considerations with the thermodynamical conditions related to seismic energy release, and then we consider the rotation counterpart in the fracturing.

We shall be aware that the rotation processes of different nature and scale take part in such extremely complicated fracture phenomena, in which the dynamic processes proceed together with the simultaneous changes of material properties (see Teisseyre [30]). We shall recall the special role of rotations in the energy release effectiveness under different load conditions, and further on we shall include the rotation impact on the granulation processes accompanying the material crushing.

The constitutive laws must undergo simultaneously considerable changes, from the rigid elastic to plastic, and further, to mylonite-type material (in tectonics the mylonite means the crushed, granulated and even partly melted material in zone adjacent to fracture plane). In the narrow zones adjacent to fracturing, the shear stresses break the molecular bonds and in the crashed rock material the stresses immediately drop to a much lower level, while together with the advancing material granulation, we shall include a rapid increase of the stress and strain rates. Finally, in that narrow zone adjacent to fracturing, the stresses and strains may be gradually neglected

and progressively replaced by their time-rates. To describe these processes we shall simultaneously introduce the changes into the related constitutive relations. In result, the rock properties in this zone may even approach those characteristic for fluid. Such conditions may permit one to include in the fracture description the transport Navier–Stokes relations. The fracturing transport process, the bond breaking and granulation processes force us to include in the fracturing description, the hypothesis that the twist-shear deformations leading to the bond breaking precede the rebound rotation motion by  $\pi/2$  in phase; this means that the difference between the shear motion and spin motion shall reach minimum when the latter is shifted by  $\pi/2$  in phase.

We shall underline that the considered conditions in the mylonite zone can serve as the basis to formulate the asymmetric fluid theory with the extreme motion phenomena and dynamic defect objects.

A counterpart to the rotations and rotation energy release at fracture processes (e.g., in an earthquake source) explains fragmentation and spall processes and makes it possible to estimate the efficiency of different fracturing modes. Again we shall underline that in any theoretical approach, the elastic rotation energy can be considered only when assuming the constitutive law joining rotations with the antisymmetric stresses or stress moments.

Teisseyre et al. [18] have reexamined Dietrich's compression experiments [31], coming to the conclusion that under the compression load, there arise at some centers in the source region the induced precursory shear stresses; at a fracturing event we would arrive at the coseismic rebound compensation leading to a release of the induced stresses by the rebound process. Similarly, the precursory rotations associated with the newly formed dislocations or cracks shall have an opposite orientation to that related to the coseismic process. At the precursory stage these repeated processes lead to micro-fracturings, while during the seismic event there will occur under compression load, the fracturing with the rock fragmentation and the rebound macro-rotations at the inner centers where the precursory induced stresses accumulate.

### Earthquake Thermodynamics

Basic thermodynamic relations for line defects (dislocations and vacant dislocations) are derived under the assumption of a dense network of defects forming a kind of super-lattice [32,33,34,35]. The thermodynamic functions of line defects can be associated with defects in the super-lattice. Let us confine our considerations to the irreversible (plastic) deformations of solids.

To distinguish the thermodynamic functions used here from those used under pressure conditions, we will use the symbols with hat and we will consider only a pure shear work under shear load  $S_{(\dots)}$  and under induced friction stress moment  $S_{[\dots]}$  (see: Eqs. (13) and (14)) – we consider deformations at a constant volume; the work  $d\hat{W}$  done on a body (per unit volume), the internal energy change  $d\hat{U}$  and the heat received in an exchange with the surrounding  $dQ$  are related:

$$\begin{aligned} d\hat{W} &= SdE = S_{(\dots)}dE_{(\dots)} + S_{[\dots]}d\omega_{[\dots]} \geq 0, \\ d\hat{U} &= dQ + SdE, \end{aligned} \tag{33}$$

where  $dE_{(\dots)}$  and  $d\omega_{[\dots]}$  are increments of strain and spin.

For the Helmholtz free energy  $\hat{F}$  and Gibbs free energy  $\hat{G}$  we have:

$$\begin{aligned} \hat{F} &= \hat{U} - T\hat{S}, & \hat{G} &= U - SE - T; \\ Td\hat{S} &\geq dQ; & d\hat{S} &\geq 0, \end{aligned} \tag{34}$$

where  $T$  is the absolute temperature and  $\hat{S}$  is the entropy;  $d\hat{S}$  would be the entropy production due to the irreversible processes occurring inside the system.

The local formulation of the second law of thermodynamics requires that the entropy production be positive wherever an irreversible process occurs [34]. It is postulated that even outside equilibrium, the entropy depends only on the same variables as at equilibrium. In order to derive the expression for the entropy production, Prigogine [36] introduced some additional assumptions. Namely, he assumed that the entropy production can be determined for conditions near equilibrium.

Formation of a dislocation gives negative contribution to the Gibbs energy [37] and therefore it is not possible to find a minimum of the Gibbs function with respect to the number of dislocations. Thus, the dislocation distribution cannot exist as a thermodynamically stable system, since the Gibbs free energy has no minimum of any equilibrium concentration of dislocations.

However, for a dense dislocation distribution there enter the repulsive interactions between dislocations, and a kind of dislocation super-lattice can be considered [32,34,35]. The “ideal super-lattice” can be treated as a reference state and the real super-lattice, in the case of dense distribution of dislocations, can be in the equilibrium state. We define the vacant dislocations in the following way: to the randomly formed network of dislocations we shall add a number of line vacancies – the vacant dislocations – in such a way that as a result we obtain the super-lattice filled by dislocations and vacant dislocations. In this situation, a real distribution of dislocations can be

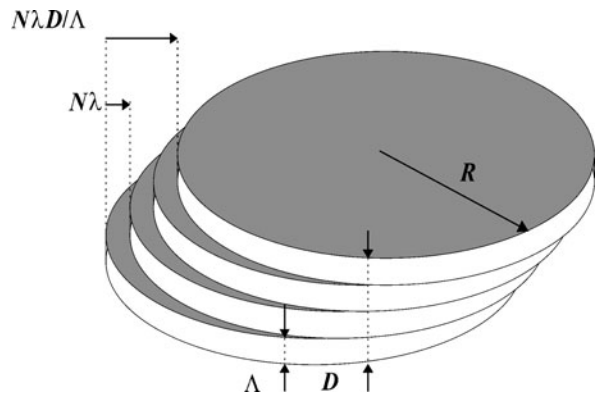
described as a departure from the state of ideal super-lattice, given by the amount of vacant dislocations.

These processes are accompanied by an internal friction related to displacement formed by dislocations and hence a spin motion appears as inherently present there.

The Gibbs energy minimum can now exist as the equilibrium number of the vacant dislocations. We can consider the structure of a cross-zone consisting of bands of layerlets; such a structure favors the appearance of some macroscopic dislocations under conditions of shearing deformation. The particular values of the Burgers vector become related to particular layer thicknesses. In this sense we suppose that a fine band structure could play the role of a quantization kind factor; this problem is related to the earthquake shear band model.

Consider a continuum that contains a regular (cubic) super-lattice of dislocation lines with a certain super-lattice parameter  $\Lambda$  ( $\Lambda \gg \lambda$ ;  $\lambda$  relates to a basic rock lattice). The notion of the super-lattice is directly related to the shear band model of fracturing [34], see Fig. 7.

We associate the thermodynamic functions of line defects with the defects in a super-lattice; the Gibbs free energy may have a minimum corresponding to the equilibrium concentration of the vacant dislocations in the super-lattice. Many results can be now transferred from the thermodynamics of point defects (Varotsos and Alexopoulos [38]). The regular super-lattice, which includes the dislocations and vacant dislocations, may be described in a very rough approximation by a characteristic distance  $\Lambda$  (super-lattice constant). For the ideal super-lattice (no vacant dislocations), the mean value of distances following from distribution of dislocations defines the reference dislocation density  $\alpha^0 = \lambda/\Lambda^2$ , while for a real body with  $n$  dislocations we may add to it other  $\hat{n}$  vacant dislo-



Earthquake Source: Asymmetry and Rotation Effects, Figure 7 Shear band model

cations in such a way that the whole set  $n + \hat{n} = N$  (dislocations and vacant dislocations) fits to a regular super-lattice with the smallest error. For the density of dislocations  $\alpha$ , and vacant dislocations  $\hat{\alpha}$ , we can write [34]:

$$\alpha = \left(1 - \frac{\hat{n}}{N}\right) \frac{\lambda}{\Lambda^2}, \quad \hat{\alpha} = \frac{\hat{n}\lambda}{N\Lambda^2} \frac{\lambda}{\Lambda^2} \exp\left(-\frac{\hat{g}^f}{kT}\right), \quad (35)$$

where the number  $\hat{n}$  can be identified with an equilibrium value in relation to the formation energy of vacant dislocation  $\hat{g}^f$  per length of the crystal lattice  $\lambda$ .

The stress field and the resistance stress (e. g., the drag resistance in a dislocation motion and the friction stress in a crack motion) are defined as [37]:

$$S = \frac{\partial \hat{W}}{\partial E}, \quad S_F \equiv \frac{\partial \hat{F}}{\partial E},$$

while the Gibbs function for a crystal containing the vacant dislocations can be written as

$$\hat{G} = \hat{G}^0 + \hat{n}\hat{g}^f - T\hat{S}_c,$$

where  $\hat{S}_c$  is the configuration entropy.

Near the equilibrium state under a constant local shear  $S$  and temperature  $T$  the Gibbs energy is close to its minimum and the equilibrium values could be found as follows:

$$\begin{aligned} \left. \frac{\partial \hat{G}}{\partial \hat{n}} \right|_{S,T} = 0, \quad \hat{n}^{\text{eq}} &= N \exp\left(-\frac{\hat{g}^f}{kT}\right), \\ \hat{\alpha} &= \frac{\lambda}{\Lambda^2} \exp\left(-\frac{\hat{g}^f}{kT}\right), \quad \hat{S}_c = \hat{n} \left(k + \frac{\hat{g}^f}{T}\right), \end{aligned} \quad (36a)$$

while the Gibbs energy function becomes

$$\hat{G} = \hat{G}^0 - \hat{n}kT. \quad (36b)$$

The equilibrium free energy is less than that for an ideal super-lattice  $\hat{G}^0$ ; the difference is  $kT$  per line vacancy, per length of crystal lattice.

For the point defect thermodynamics, Varotsos and Alexopoulos [38] have introduced the so-called  $CB\Omega$  theory approximating the contribution to the Gibbs energy from the formation of point defects.

For the line vacancies, a change of the Gibbs energy depends on the stress level and resistance stress. Therefore, we postulate for the approximative value of such change per unit element (formation energy of vacant dislocation) the following expression defining the  $C\mu b\lambda^2$  model:

$$\hat{g}^f = C\mu b\lambda^2, \quad \hat{n}^{\text{eq}} = N \exp\left(-\frac{C\mu b\lambda^2}{kT}\right), \quad (37)$$

where  $C$  is constant;  $\hat{g}^f$  becomes here independent on stress load and resistance,  $\mu$  is the rigidity,  $b$  is the Burgers vector of dislocation.

Concluding, a body containing some number of dislocations cannot be in a state of equilibrium; there is no minimum of the Gibbs function, because when reducing the number of dislocations we always get a smaller value of the free energy. For a dense distribution of dislocations we can assume, due to their interaction, that there exists a certain super-lattice composed of dislocations.

The equilibrium density of the vacant dislocations may be written now with help of Eq. (37)

$$\hat{\alpha} = \frac{\lambda}{\Lambda^2} \exp\left(-\frac{C\mu b\lambda^2}{kT}\right) \quad (38)$$

and becomes useful, when looking for the most probable density value of defects after the energy release in a fracturing process. The density  $\alpha^0 = \lambda/\Lambda^2$  may be identified here with the reference density.

We can assume that before an earthquake a super-lattice is almost completely filled in by dislocations ( $n \approx N$  and  $\hat{n} \approx 0$ ). The maximum number of dislocations in arrays could reach the value  $(\Lambda/\lambda)^2$  per area  $\Lambda^2$ . The total moment for an area  $\Delta s = N\Lambda^2$  affected by the arrays of dislocation along the slip planes becomes:

$$\bar{M} = \mu\lambda\Delta s = \mu\lambda N\Lambda^2 \left(\frac{\Lambda}{\lambda}\right) = \mu N\Lambda^3.$$

After an earthquake, the number of vacant dislocations  $\hat{n}$  shall increase, probably to the equilibrium value (37) and hence we can express the seismic moment by the number of coalescence processes related to surface element  $\Lambda^2$  as equal to  $\Delta\hat{n} = \frac{\Lambda}{\lambda}\hat{n}^{\text{eq}}$ ; the factor  $\Lambda/\lambda$  expresses a maximum concentration of dislocations in the arrays.

We obtain for the seismic moment

$$\begin{aligned} \bar{M}_0 &= \bar{M}\Delta\hat{n} = \mu N\Lambda^3 \left(\frac{\Lambda}{\lambda}\right) \Delta\hat{n} \\ &= \mu N\Lambda^3 \left(\frac{\Lambda}{\lambda}\right) \exp\left(-\frac{C\mu\lambda\Lambda^2}{kT}\right), \end{aligned}$$

where  $C$  is constant for given structure.

Using the expression for a change of the free energy values we may include the formation of dislocation arrays along the glide planes and we put

$$G = G^0 + \Delta\hat{n} \left(\frac{\Lambda}{\lambda}\right) kT$$

According to these results, the total energy release  $\Delta E$  and seismic moment are:

$$\begin{aligned}\Delta E &= G - G^0 = \Delta \hat{n} \left( \frac{\Lambda}{\lambda} \right) kT \\ &= \left( \frac{\Lambda}{\lambda} \right) NkT \exp \left( -\frac{C\mu\lambda\Lambda^2}{kT} \right) \text{ and} \\ \tilde{M}_0 &= \mu\Lambda^3 \frac{\Delta E}{kT}.\end{aligned}\quad (39)$$

This formula is an important relation between the energy release density and seismic moment density; for instance, for a given  $\Delta E$  the elementary seismic moment  $\tilde{M}_0$  decreases with temperature. Free energy related to defect formation,  $\hat{g}^f$ , is proportional to  $\mu\lambda\Lambda^2$  being constant for a given structure; with growing value of  $\Lambda$  the seismic moment becomes greater.

Neglecting the term related to the formation entropy, we can write for entropy density change:

$$\Delta \tilde{S} = kN \left( \frac{\Lambda}{\lambda} \right) \left( 1 + \frac{C\mu b\lambda^2}{kT} \right) \exp \left( -\frac{C\mu\lambda\Lambda^2}{kT} \right).$$

All of these relations concern the quantities referred to the multiple of the cubic volume  $N\Lambda^3$  thus, we can correct these quantities to that related to a given source volume by introducing the factor  $\pi R^2 D / N\Lambda^3$ :

$$\begin{aligned}M_0 &= \mu\pi R^2 D \left( \frac{\Lambda}{\lambda} \right) \exp \left( -\frac{C\mu\lambda\Lambda^2}{kT} \right), \\ \Delta E &= \pi R^2 D \frac{kT}{\Lambda^2 \lambda} \exp \left( -\frac{C\mu\lambda\Lambda^2}{kT} \right) \text{ and} \\ \frac{M_0}{E^{\text{rad}}} &= \frac{\mu\Lambda^3}{\eta kT}, \\ \Delta \tilde{S} &= \pi R^2 D \frac{k}{\Lambda^2 \lambda} \left( 1 + \frac{C\mu\lambda\Lambda^2}{kT} \right) \exp \left( -\frac{C\mu\lambda\Lambda^2}{kT} \right),\end{aligned}\quad (40)$$

where  $\eta\Delta E = E^{\text{rad}}$ ,  $\eta$  is the seismic efficiency;  $E^{\text{rad}}$  is radiated energy.

In the above consideration we took into account both energies related to slip and friction processes, and thus, the total released energy includes that related to stress drop and that related to heat caused by friction processes.

Further on, we will consider the fracturing processes in an earthquake source.

### Synchronization and Fracturing

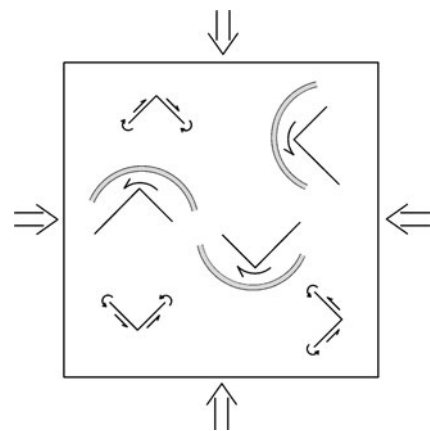
The inner stress accumulation relates to the formation of defect densities; due to the interaction between dislocations, we arrive at stress concentration at the first blocking dislocations of the formed dislocation arrays.

In the compression case with no initial shear field and due to the lower value of shear resistance, we have to assume that inside a body there appear regions with induced shear stresses of opposite signs, and induced antisymmetric stresses. The earthquake process and its energy release relate to a coalescence of dislocation arrays of opposite signs and related rotation release motion. Of course, we shall consider a fracture process as a chain of events; let us consider the micro-fracture centers formed on two perpendicular plane fragments; the induced shear stresses will be opposite on those plane fragments, but will have the common orientation of a spin motion – see Fig. 8; hence, the shears will be almost compensated for, while the spin field will remain unchanged.

The spin field,  $\omega_{[s]}$ , propagates and influences the processes in the adjacent regions; we believe that this propagation synchronizes the spin motions in the adjacent centers, in such a way that the sense of spin motion becomes the same over the whole fracture region. That means, the spin propagation assures a synchronization of fracture processes, especially under compression load where the energy release relates to the fragmentation revealed by rotation and granulation processes.

Reversely, under the shear load, while common shear deformation,  $E_{kn}^D$  (or expressed as twist  $\omega_{(s)}$ ), progresses, the spins on the main fracture differ from those on the adjacent perpendicular fragments and attenuate fracture progress on those fragments.

Accordingly, we can believe that at the compression load, the total shear stress drop will be relatively small, while the rebound rotations will release an important amount of rotation energy. At the shear load the release of shear stresses will prevail.



Earthquake Source: Asymmetry and Rotation Effects, Figure 8  
Compression load: induced shear centers and formation of fragments with related rotations

Concluding, the rotation processes in fragmentation and fracturing under compression load play an essential role. Under prevailing shear load the rebound process releases shear load with the regional stress drop, while the rotation processes play a minor role.

Further on, we will discuss the importance of the granulation processes related to rotations in meso-scale, which we can place between the bond breaking processes in the micro-scale and material fragmentation in the macro-scale.

**Granulation and Formation of Mylonite Zones**

The fracturing process, especially under the action of shearing load, is accompanied by material granulation adjacent to shear fracture planes; thus it becomes spectacular at the formation of narrow, long mylonite zones. In this process we shall take into account a special role of rotations – the meso-rotations of different scales; these rotations are related to bond breaking and friction processes.

Co-action of the spin and twist-shear motions in bond breaking, granulation and formation of mylonite material can effectively help us to explain the fracture process; the simultaneous formation of the adjacent mylonite zone appears due to such a co-action of spin and shears and of the fracture transport phenomena.

Based on the standard asymmetric continuum theory, as presented in the first part, we would like to consider the material undergoing a progressive crushing process. We may arrive even at the conditions more similar to fluid material, and thus finally shall enter into our considerations the Navier–Stokes transport equations.

Starting with the description of the rock continuum following from the standard asymmetric theory of continuum ( $S_{ik} = S_{(ik)} + S_{[ik]}, E_{ik} = E_{(ik)}, \omega_{ik} = \omega_{[ik]}$ ), we approach the final stage of the crushing/granulation process in zones adjacent to fracture planes. In these zones, simultaneously with dynamic processes, there occur changes of material properties from hard rocks to mylonite granulated material.

Approaching the final stage, the stresses, strains and rotations presented in the description of the standard asymmetric continuum become gradually neglected and progressively replaced by the constitutive relations for time-rates of stresses and strains.

The constitutive laws for rock asymmetric continuum – the relations (1) written for the deviatoric fields and for the antisymmetric fields

$$S_{(kl)}^D = 2\mu E_{kl}^D, \quad S_{[kl]} = 2\mu\omega_{kl} \tag{41}$$

– will gradually change during fracturing to those including the time dependent processes:

$$\begin{aligned} \sigma S_{(ik)}^D + \tau \dot{S}_{(ik)}^D &= 2\mu E_{ik}^D + 2\eta \dot{E}_{ik}^D, \\ \sigma S_{[ik]} + \tau \dot{S}_{[ik]} &= 2\mu E_{ik} + 2\eta \dot{\omega}_{ik}. \end{aligned} \tag{42}$$

The introduced material constants are related to magnitudes of the slip,  $u$ , and slip rate,  $v$ .

When in a narrow zone the huge shear stresses break the molecular bonds, the stresses crushing rock material immediately drop down to the low values and in the crushed mylonite material we observe the immediate increase of the stress and strain rates to such degree that the stresses and strains may be neglected in the respective constitutive relations for that narrow zone. Finally, these changes will lead to the constitutive laws for the melt and granulated parts of mylonite material in which, practically, will remain only the field time rates:

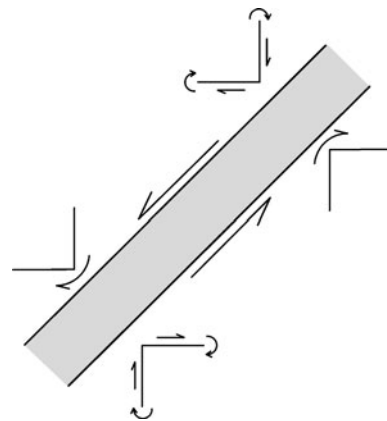
$$\dot{S}_{(ik)} = 2\eta \dot{E}_{ik}, \quad \dot{S}_{[ik]} = 2\eta \dot{\omega}_{ik}. \tag{43}$$

The direct observation of the gauge zone of the Kobe (Japan, 1995) earthquake at the Avaji island suggests that the size of an inner completely melted part of the mylonite zone ranges around couple of centimeters (private communication W. Debski).

Further on, we will assume for the sake of simplicity, that during the fracturing the mylonite material remains incompressible.

In such a way, the nucleation progresses and fracture propagates simultaneously with the granulation process in the intact material (or in the compact zone previously crushed) – see Fig. 9.

In this new description, the shear rates create the dynamic angular deformations then lead to the bond



Earthquake Source: Asymmetry and Rotation Effects, Figure 9  
Mylonite zone and neighboring deformations

breaking processes, and finally to the fracturing transport process.

We pass to the final stage; for the crushed incompressible mylonite or sand, similarly to incompressible fluids, where tensor  $\hat{\omega}_{ik}^T$  is related to spin motion (not to rotation of displacement), the mylonite viscosity is  $\eta$ , and the mylonite relaxation time is denoted by  $\tau$ .

The relations (42) define the ideal quasi-viscous mylonite for an incompressible crushed material.

Further, we assume the coincidence/identity of rotation of velocity field  $\dot{\mathbf{u}}$  with the point rotation field  $\omega$ . So, we assume that rotations of particles (micro-rotation  $\omega$ ) coincide with macro-rotations (rot  $\mathbf{u}$ ). For mylonite, such a coincidence between the micro-rotation and macro-rotation seems reasonable, and thus, our assumption that viscosity  $\eta$  coincides with rotation viscosity  $\eta^*$  may be correct.

For our narrow mylonite zone, existing already near the pre-slip planes or just simultaneously formed, we may, further on, apply the Navier–Stokes transport equation. Referring to our former considerations on the asymmetric continuum theory (see Subsect. “Spin and Twist Motions”) we may add to the spin rotational motions the oscillations of the strain shear rates (called twist motion). Such motions, especially in an earthquake source zone, are due to the friction processes.

We may note that when including these complex rotational motions in the theory we may replace the friction constitutive laws, as based on the experimental data, by the constitutive law joining the asymmetric stresses with spin and shears field oscillations or otherwise with spin and twist.

### Slip Propagation and Spin Release Hypothesis

While searching the fault slip solutions, we use the classical elasticity tools with an additional friction constitutive law based on experimental data. When instead of it we consider the asymmetric elastic continuum, we are able to include the defect interaction and we can derive the elastodynamic fault solution describing slip propagation with fracturing process and related seismic radiation.

The angular deformations preceding the bond breaking process lead to the efficient rise of the angular moments around material grains. In the narrow mylonite zone, we arrive at the equivalence between this expression and the laws introduced in the considerations on the friction resistance and slip.

The co-action of the rotation (rot  $\mathbf{u}$ , or spin  $\omega_{[.]}$ ) and shear ( $\mathbf{E}^D(\mathbf{u})$ , or twist –  $\omega_{(.)}$ ) motions can lead further to the slip fracturing motion. We assume that the bond

breaking process and granulation of material precede the slip movement: just after the bond breaking micro-process there, we would have the released rebound spin motion retarded in phase.

This hypothesis is supported by the following solution of the homogeneous wave equations for the twist and spin in a mylonite zone (see Eq. (31)):

$$\begin{aligned}\omega_{(s)} &= i\omega_{[s]}; & \omega_{[s]} &= \omega_{[s]}^0 \exp[i(k_i x_i - \omega t)], \\ \omega_{(s)} &= \omega_{(s)}^0 \exp[i(k_i x_i - \omega t)],\end{aligned}\quad (44)$$

where the six constants in  $\omega_{[s]}^0 = \text{abs}(\omega_{[s]}^0) \exp(i\psi_s)$ ,  $\omega_{(s)}^0 = \text{abs}(\omega_{(s)}^0) \exp(i\varphi_s)$  shall fulfill the six conditions.

We may consider the following 2D solution of Eq. (44) in the systems  $\{r, \varphi, z\}$ :

$$\omega_{(\varphi)}(r) = i\omega_{[\varphi]}(r), \quad \square\omega_{[\varphi]}(r) = 0.$$

The related solution corresponds to a turbulence structure. Thus, from a dislocation-slip structure formed in the earthquake premonitory domain (see: Subsect “Earthquake Thermodynamics”), gradually destroyed during a fracture process by a spin release motion, we can arrive at a turbulence structure appearing in a melted or fully granulated material.

With the introduced waves,  $\omega_{(s)} = i\omega_{[s]}$  we arrive at the possibility to study the dynamic defect objects and to explain the synchronization of the micro-fracturing processes due to an influence of the propagating waves. For the fracture processes under compression such a synchronization will assure the common sense of the induced twist and spin motions, while under shear load – the formation of a long shearing fracturing. In the last case, the spin waves related to a given slip on the main fracture plane attenuate those with the opposite spins generated at the perpendicular fragments, and due to the conjugate solution (see Eq. (44) reduce the slip motions on those fragments.

The presented conjugate solution Eq. (44) suggests that the spin rebound motion is delayed in phase by  $\pi/2$  (as we have  $\exp[i(k_i x_i - \omega t)] = \exp[ik_i x_i - i(\omega t - \pi/2)]$ ); when slip starts due to breaking of bonds, the micro-spin motions are released.

Following this assumption we expect that such a correlation between the recorded twist motions and spin motions shifted by  $\pi/2$  in phase can exist in some wavelets.

Now we can propose the following description of the fracture process.

- First, according to external load conditions the stresses rise while the disclosure and dislocation field can be neglected – the case given by relations (25).

- Next, approaching the fracture process we may observe the “accumulation” phase with the co-action of the twist and spin – the case given by relations (26).
- Finally, fracturing processes start and when entering into the time rate domain we can describe the “release” phase of the process as follows (compare: Eqs. (21–26 and 43):

$$\begin{aligned} \dot{B}_l &= \oint [E_{(kl)} + i\omega_{[kl]}] dl_k, \\ \dot{\alpha}_{pl} &= \varepsilon_{pmk} \left( \frac{\partial \dot{E}_{kl}}{\partial x_m} + i \frac{\partial \omega_{kl}}{\partial x_m} \right) \\ \dot{\alpha}_{pl} &= \frac{\varepsilon_{pmk}}{2\mu} \frac{\partial}{\partial x_m} \left( (\dot{S}_{(kl)} - \frac{\nu}{1+\nu} \delta_{kl} \dot{S}_{ii}) + i \dot{S}_{[kl]} \right), \end{aligned} \tag{45} \tag{46}$$

where we have the dynamic disclosure and  $\nu$ -dislocation density under the conditions formed by solution (see Eq. (44)), supplemented with the relations between the asymmetric stress rates and the dynamic dislocation objects ( $\nu$ -dislocations).

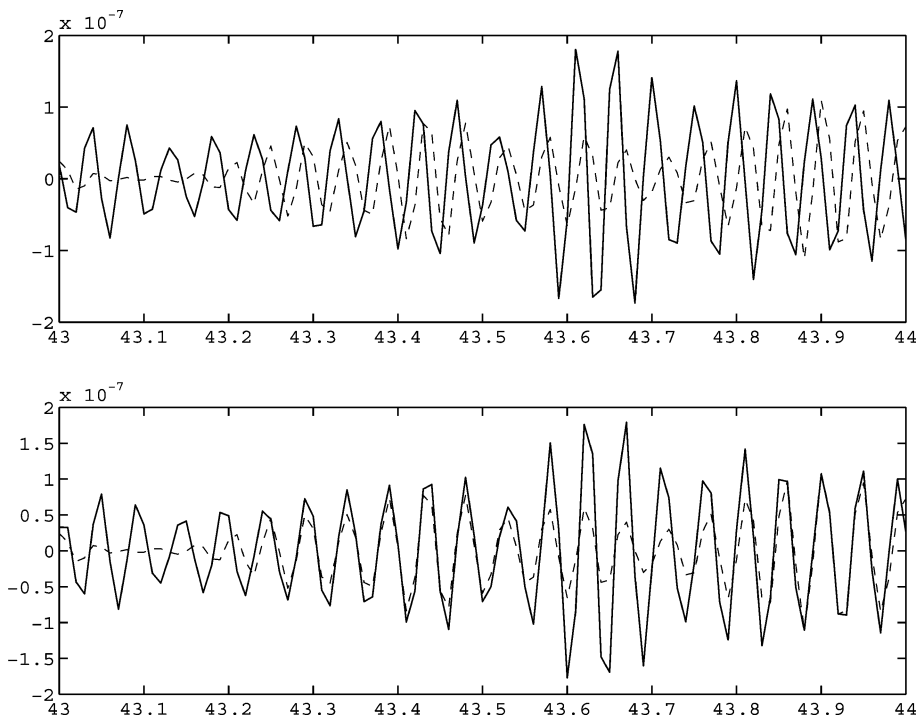
- This case presents a formation of dynamic discontinuities and the related dynamic processes in which the slip and bond breaking leads to the rebound spin motions delayed in phase by  $\pi/2$ .

The co-action of the spin and twist motions leads to the “accumulation” phase, while the conjugate solution (see Eq. (44)) presents a fracture process – “release” phase.

We might suppose that the fracture process could proceed with the consecutive accumulation and release micro-processes; in such a situation the related twist and spin motions will appear consecutively as pairs of wavelets in phase (or anti-phase) and those differing in phase by  $\pi/2$ . Figure 10 presents an example of the coincidence of the spin and twist motions after the Hilbert transformation shifting the angular twist record ahead in phase by  $\pi/2$ .

The presented theory, due to its simplicity, could be very useful for some problems, among others those in which macro-rotation takes an important role in the asymmetric fluid dynamics.

We can call the solution (44) as the fracture synchronization waves.



Earthquake Source: Asymmetry and Rotation Effects, Figure 10

Example of coincidence of the spin and twist angular motions (rad/s versus s) after the Hilbert transformation shifting the twist record ahead in phase by  $\pi/2$ ; upper part – the original records, lower part – the twist record transformed (from the original seismic record obtained by the system of the rotation seismometers; L’Aquila Observatory, 17.02.2006; the continuous line – twist, the broken line – spin)



Finally, we shall notice that similar solutions may exist for the electric and magnetic induction vectors:

$$D_s = iB_s \leftrightarrow D_s^0 = iB_{[s]}^0, \tag{47}$$

where this relation shall be assured by the appropriate material constants.

**Towards Asymmetric Fluid Theory and Extreme Phenomena**

Approaching the conclusions, we shall specify how we could formulate the asymmetric theory of the fluid continuum in which the stress, strain and rotation fields vanish, but their rates exist as related to the velocity field  $v$ :

$$\begin{aligned} \dot{E}_{ik} &= e^0 \frac{1}{2} \left( \frac{\partial v_k}{\partial x_i} + \frac{\partial v_i}{\partial x_k} \right), \\ \dot{\omega}_{ik} &= \chi^0 \frac{1}{2} \left( \frac{\partial v_k}{\partial x_i} - \frac{\partial v_i}{\partial x_k} \right). \end{aligned} \tag{48}$$

The velocity field  $v$  shall obey the Navier–Stokes **transport equation**.

For the sake of simplicity, let us consider the incompressible fluid ( $\dot{E}_{ss} = 0, \dot{E}_{ik} = \dot{E}_{ik}^D$ ), the basic constitutive laws are assumed similarly to those for the asymmetric continuum:

$$\dot{S}_{(ik)} = \eta \dot{E}_{ik}, \quad \dot{S}_{[ik]} = \eta \dot{\omega}_{ik}, \tag{49}$$

where  $\eta$  is viscosity.

In a similar manner, the structural dynamic objects can then be defined as defects in the standard asymmetric continuum.

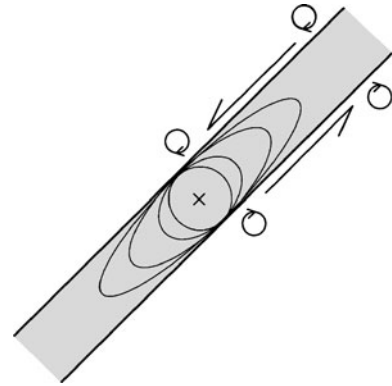
The extreme motion phenomena, related to the shear rate and spin,  $\dot{E}_{(kl)} + \dot{\omega}_{(kl)}$ , could be expected for the following case (compare: Eqs. (22–26 and 43):

$$\begin{aligned} \dot{B}_l &= \oint [\dot{E}_{kl} + \dot{\omega}_{kl}] dl_k = \oint \frac{\partial v_l}{\partial x_k} dl_k = 0, \\ \dot{\alpha}_{pl} &= 0, \quad \dot{\Omega}_q = 0, \end{aligned} \tag{50}$$

while formation of the dynamic defect objects can be described as:

$$\begin{aligned} \dot{B}_l &= \oint [\dot{E}_{kl} - \dot{\omega}_{kl}] dl_k =, \\ \dot{\alpha}_{pl} &= \frac{1}{2\mu} \varepsilon_{pmk} \frac{\partial}{\partial x_m} [\dot{S}_{(kl)} - \dot{S}_{[kl]}]. \end{aligned} \tag{51}$$

The former case (see Eq. (50)) presents an extreme shear rate deformation, like soliton waves, while this case (see



Earthquake Source: Asymmetry and Rotation Effects, Figure 11 Extreme motions: soliton wave will be related to a given deformation of the circle as follows for the parameter  $C = e^0$ ; Amplitude plot (Mathematica 5.0): with AspectRatio  $\rightarrow$  Automatic:  $A = \text{Plot}[y = -Cx/2 \pm 0.5\sqrt{[(Cx) \wedge 2 - 4x \wedge 2 + 4]}], \{x, -\sqrt{[4/(4 - C \wedge 2)]}, \sqrt{[4/(4 - C \wedge 2)]}\}$

Eq. (51)) would relate to a formation of the  $v$ -slip-discontinuity.

In 2D we can show an effect of the co-action of the macro spin and twist motions in the following way; let us put  $C = e^0$  and  $\chi^0 = 1$

$$\begin{aligned} \dot{E}_{12} + \dot{\omega}_{12} &= Cv_{(2,1)} + v_{(2,1)} \\ &= C \frac{1}{2} \left( \frac{\partial v_k}{\partial x_i} + \frac{\partial v_i}{\partial x_k} \right) + \frac{1}{2} \left( \frac{\partial v_k}{\partial x_i} - \frac{\partial v_i}{\partial x_k} \right). \end{aligned}$$

For  $C = 0$  we would have only the spin motion, while a full coincidence will occur at  $C = 1$ .

The effect of such a superposition of the spin and twist motions is presented on Fig. 11.

We believe that this approach might explain some extreme fluid phenomena related to atmosphere and oceans.

The balance equations for field rates (48) at  $e^0 = \chi^0$  may lead us to the wave equations for the related spin and twist rate fields,  $\dot{\omega}_{[s]}$  and  $\dot{\omega}_{(k)}$ , as defined similarly to the relations derived in Subsect. “Balance Laws for the Rotation Field and the EM Analogy” (see: Eqs. (29 and 31):

$$\begin{aligned} \varepsilon_{kps} \frac{\partial}{\partial x_p} \dot{\omega}_{[s]} - \frac{1}{V} \frac{\partial}{\partial t} \dot{\omega}_{(k)} &= \frac{4\pi}{V} j_k, \\ \varepsilon_{kps} \frac{\partial}{\partial x_p} \dot{\omega}_{(s)} + \frac{1}{V} \frac{\partial}{\partial t} \dot{\omega}_{[k]} &= 0. \end{aligned} \tag{52}$$

The appearance of such coupled waves transversal to the transport motion brings physical background for diffraction in fluids as explained usually by the Huygens principle.

## Conclusions

Under both the confining pressure and external shear, the role of micro-fracturing in the bond breaking process is similar; however, we observe the essential differences for rotations in larger scales.

The confining condition leads to formation of induced opposite arrays of dislocations, resulting in fragmentation processes and chaotically oriented macro-rotations, leading therefore to a rotation release process.

The shear condition leads to more concentrated fracturing along some planes, high shear strain release and correlated rotations.

Both cases include formation of narrow mylonite zones adjacent to the fracturing planes or their fragments, but these processes prevail rather under shear conditions.

We draw attention to the importance of the rotations in meso-scales – between the micro-scale bond breaking process and that related to macro-rotation at material fragmentation. The meso-scale rotations are related to material granulation and become observed in any fracturing process; such motions may be revealed in the spectacular formation of the narrow, long mylonite zones under shear load conditions. Coincidence and co-action of the spin and twist-shear motions in bond breaking and formation of mylonite material help one to understand the fracture motion; the simultaneous formation of the mylonite zones appears due to common action of these motions and to fracture transport phenomena.

Rotations at source zones help to understand geometry of fracturing and releases of stress and rotation counterparts as a result of precursory and rebound processes.

We have presented also a new idea how to construct the asymmetric fluid theory with the asymmetric stress rate field.

## Final Remarks

With the additional constitutive law joining the rotations with antisymmetric part of stresses, we have proved that the rotation waves exist, even in a homogeneous elastic continuum. We have defined the twist motion as the rotational oscillation of the main shear axes including the shear magnitude variations. The derived wave equations for the twist and spin motions have been considered in relation to the processes in seismic source.

Only in the presented approach with the standard asymmetric theory may we study the co-action of the independent motions and deformations; this is due to the new relations joining the spin and slip motions. A possible phase shift between these motions leads to different families of deformation and related solutions.

We have shown how the rotations at source zones help us to understand physics and geometry of fracturing and release of stresses in the precursory and rebound processes.

The derived wave equations for spin and twist motions are similar to the EM wave equations.

Our considerations show the importance of the simultaneous recording of the translational and rotational earthquake motions, and also the strains (at least the deviatoric strains).

Finally, we have shown how to construct the asymmetric fluid theory in which, with a help of the asymmetric stress rates, the various extreme phenomena, including soliton waves, can be theoretically explained.

## Acknowledgments

Work done under the support of the INTAS Project 05-1000008-7889.

## Bibliography

### Primary Literature

1. Shimbo M (1975) A geometrical formulation of asymmetric features in plasticity. *Bull Fac Eng Hokkaido Univ* 77:155–159
2. Shimbo M (1995) Non-Riemannian geometrical approach to deformation and friction. In: Teisseyre R (ed) *Theory of earthquake premonitory and fracture processes*. PWN, Polish Scientific Publishers, Warszawa, pp 520–528
3. Teisseyre R, Boratyński W (2003) Continua with self-rotation nuclei: evolution of asymmetric fields. *Mech Res Commun* 30:235–240
4. Teisseyre R, Boratyński W (2006) Deviations from symmetry and elasticity: Asymmetric continuum mechanics. In: Teisseyre R, Takeo M, Majewski E (eds) *Earthquake source asymmetry, structural media and rotation effects*. Springer, Berlin, pp 31–42
5. Kozak JT (2006) Development of earthquake rotational effect study. In: Teisseyre R, Takeo M, Majewski E (eds) *Earthquake source asymmetry, structural media and rotation effects*. Springer, Berlin, pp 3–10
6. Ferrari G (2006) Note on the historical rotation seismographs. In: Teisseyre R, Takeo M, Majewski E (eds) *Earthquake source asymmetry, structural media and rotation effects*. Springer, Berlin, pp 367–376
7. Gutenberg B (1926) *Grundlagender Erdbebenkunde*. Univ. Frankfurt a/M, Frankfurt
8. Teisseyre R (1973) Earthquake processes in a micromorphic continuum. *Pure Appl Geophys* 102:15–28
9. Cochard A, Igel H, Schuberth B, Suryanto W, Velikoseltsev A, Schreiber U, Wassermann J, Scherbaum F, Vollmer D (2006) Rotational motions in seismology: Theory, observation, simulation. In: Teisseyre R, Takeo M, Majewski E (eds) *Earthquake source asymmetry, structural media and rotation effects*. Springer, Berlin, pp 391–411
10. Schreiber KU, Stedman GE, Igel H, Flaws A (2006) Ring laser gyroscopes as rotation sensors for seismic wave studies. In: Teis-

- seyre et al (eds) Earthquake source asymmetry, structural media and rotation effects. Springer, Berlin, pp 377–389
11. Takeo M (2006) Rotational motions excited by earthquakes. In: Teisseyre R, Takeo M, Majewski E (eds) Earthquake source asymmetry, structural media and rotation effects. Springer, Berlin, pp 131–156
  12. Leszek R, Jaroszewicz LR, Krajewski Z, Solarz L (2006) Absolute rotation measurement based on the Sagnac effect. In: Teisseyre R, Takeo M, Majewski E (eds) Earthquake source asymmetry, structural media and rotation effects. Springer, Berlin pp 413–438
  13. Moriya T, Teisseyre R (2006) Design of rotation seismometer and non-linear behaviour of rotation components of earthquakes. In: Teisseyre R, Takeo M, Majewski E (eds) Earthquake source asymmetry, structural media and rotation effects. Springer, Berlin, pp 439–450
  14. Wiszniowski J (2006) Rotation and twist motion recording – couple pendulum and rigid seismometers system. In: Teisseyre R, Takeo M, Majewski E (eds) Earthquake source asymmetry, structural media and rotation effects. Springer, Berlin, pp 451–470
  15. Teisseyre R, Suchcicki J, Teisseyre KP, Wiszniowski J, Palangio P (2003) Seismic rotation waves: basic elements of the theory and recordings. *Ann Geophys* 46:671–685
  16. Teisseyre R, Bialecki M, Górski M (2006) Degenerated asymmetric continuum theory. In: Teisseyre R, Takeo M, Majewski E (eds) Earthquake source asymmetry, structural media and rotation effects. Springer, Berlin, pp 43–56
  17. Kröner E (1981) Continuum theory of defects. In: Balian R, Klemman M, Poirer JP (eds) *Physique des défauts/physics of defects* (Les Houches, Session XXXV, (1980)). North Holland Publ Com, Dordrecht
  18. Teisseyre R, Górski M, Teisseyre KP (2006) Fracture–band geometry and rotation energy release. In: Teisseyre R, Takeo M, Majewski E (eds) Earthquake source asymmetry, structural media and rotation effects. Springer, Berlin, pp 169–184
  19. Nowacki W (1986) Theory of asymmetric elasticity. PWN, Warszawa and Pergamon Press, Oxford, New York, Toronto, Sydney, Paris, Frankfurt, p 383
  20. Cosserat E, Cosserat F (1909) *Theorie des Corps Déformables*. A. Hermann, Paris
  21. Eringen AC, Suhubi ES (1964) Non-linear theory of simple micro-elastic solids. I *Int J Eng Sci* 2:189–203
  22. Mindlin RD (1965) On the equations of elastic materials with microstructure. *Int J Solids Struct* 1:73
  23. Teisseyre R (2005) Asymmetric continuum mechanics: Deviations from elasticity and symmetry. *Acta Geophys Polon* 53:115–126
  24. Teisseyre R, Kozak JT (2003) Considerations on the seismic rotation effects. *Acta Geophys Polon* 51:243–256
  25. Kossecka E, DeWitt R (1977) Disclination kinematic. *Arch Mech* 29:633–651
  26. Teisseyre R (2001) Evolution, propagation and diffusion of dislocation fields. In: Teisseyre R, Majewski E (eds) Earthquake thermodynamics and phase transformations in the earth's interior. Academic Press (Vol. 76 of International Geophysical Series), San Diego, San Francisco, New York, Boston, London, Sydney, Tokyo, pp 167–198
  27. Klemman M (1980) The general theory of disclinations. In: Nabarro FRN (ed) *Dislocations of solids*, vol 5. Other effects of dislocations: Disclinations. North-Holland Publ. Comp., Amsterdam, pp 243–297
  28. Nabarro FRN (1967) Theory of crystal dislocations. Clarendon Press, Oxford
  29. Moriya T, Marumo T (1998) Design for rotation seismometers and their calibration. *Geophys Bull Hokkaido Univ* 61:99–106
  30. Teisseyre R (1996) Shear band thermodynamical earthquake model. *Acta Geophys Polon* 44:219–236
  31. Dietrich JHJ (1978) Preseismic fault slip and earthquakes prediction. *J Geophys Res* 83(B8):3940–3954
  32. Teisseyre R, Majewski E (1990) Thermodynamics of line defects and earthquake processes. *Acta Geophys Polon* 38:355–373
  33. Teisseyre R, Majewski E (1995) Earthquake thermodynamics. In: Teisseyre R (ed) Theory of earthquake premonitory and fracture processes. PWN, Warszawa, pp 586–590
  34. Teisseyre R, Majewski E (2001) Thermodynamics of line defects and earthquake thermodynamics. In: Teisseyre R, Majewski E (eds) Earthquake thermodynamics and phase transformations in the earth's interior. Academic Press (Vol. 76 of International Geophysical Series), San Diego, San Francisco, New York, Boston, London, Sydney, Tokyo, pp 261–278
  35. Majewski E, Teisseyre R (1997) Earthquake thermodynamics. *Tectonophysics* 227:219–233
  36. Prigogine I (1979) Irreversibility and randomness. *Astron Phys Space Sci* 65:371–381
  37. Kocks UF, Argon AS, Ashby MF (1975) Thermodynamics and kinetics of slip. Pergamon Press, Oxford, New York, p 288
  38. Varotsos PA, Alexopoulos KD (1986) Thermodynamics of point defects and their relation with bulk properties. North-Holland, Amsterdam, New York, p 474

### Books and Reviews

- Bridgman W (1950) The thermodynamics of plastic deformation and generalized entropy. *Rev Mod Phys* 22:56–63
- Eringen AC (1999) Microcontinuum field theories I: Foundations and solids. Springer, Berlin, p 325
- Eringen AC (2001) Microcontinuum field theories II: Fluent media. Springer, Berlin, p 340
- Drazin PG (1983) *Solitons*. Cambridge University Press, Cambridge
- Infeld E, Rowlands G (2000) *Nonlinear waves, solitons and chaos*. Cambridge University Press, Cambridge
- Muskhelishvili NT (1953) Some basic problems of the elasticity. Noordhoff, Groningen
- Newell A (1985) *Solitons in mathematics and physics*. Society for Industrial and Applied Mathematics, Philadelphia
- Prigogine I (1978) *Thermodynamics of irreversible processes*, 3rd edn. Wiley, New York
- Teisseyre R (1974) Symmetric micromorphic continuum: wave propagation, point source solutions and some applications to earthquake processes. In: Thoft-Christensen P (ed) *Continuum mechanics aspects of geodynamics and rock fracture mechanics*. D. Reidel Publ. Comp., Dordrecht-Holland/Boston, pp 201–244
- Teisseyre R (ed) (1995) Theory of earthquake premonitory and fracture processes. PWN, Warszawa, p 648
- Teisseyre R, Czechowski L, Leliwa-Kopystynski J (eds) (1993) *Dynamics of the earth's interior*. PWN, Warszawa, p 469
- Teisseyre R, Majewski E (eds) (2001) Earthquake thermodynamics

- and phase transformations in the earth's interior. Academic Press (Vol. 76 of International Geophysical Series), San Diego, New York, Boston, London, Sydney, Tokyo, p 670
- Teisseyre R, Majewski E (2002) Physics of earthquakes. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International handbook of earthquake and engineering seismology, Part A. Academic Press, Amsterdam, pp 229–235
- Teisseyre R, Takeo M, Majewski E (eds) (2006) Earthquake source asymmetry, structural media and rotation effects. Springer, Berlin, p 582
- Thoft-Christensen P (ed) (1974) Continuum mechanics aspects of geodynamics and rock fracture mechanics, NATO Advance Study Institutes Series C, vol 12. D. Reidel Publ. Comp., Dordrecht-Holland/Boston, p 273

## Earthquake Source Parameters, Rapid Estimates for Tsunami Warning

BARRY HIRSHORN, STUART WEINSTEIN  
NOAA/NWS/Pacific Tsunami Warning Center,  
Ewa Beach, USA

### Article Outline

Glossary  
Definition of the Subject  
Introduction  
Tsunami Warning Center Operations  
Seismic Methods  
Earthquake Source Parameters  
Future Directions  
Acknowledgments  
Bibliography

### Glossary

**CMT centroid moment tensor** The CMT represents the displacement of the Earth's crust that best reproduces the observed wave-field generated by an earthquake and gives the average location in time and space of the earthquake energy release. The seismic moment can be determined from the CMT.

**Convolution** Convolution is a type of integral transform combining two signals to form a third signal or output. It is the single most important technique in Digital Signal Processing. In the case of Seismology, the two signals can be e. g., the ground motion as a function of time and the response of the seismometer, and the output is the seismogram.

**Deconvolution** Does the reverse of convolution. In the case of Seismology, one uses deconvolution to remove the instrument response from the seismogram to recover the actual ground motion.

**Deep earthquake** An earthquake characterized by a hypocenter located more than 100 km below the Earth's surface.

**Hypocenter** The point within the Earth where the earthquake rupture starts. The epicenter is the projection of the hypocenter onto the Earth's surface.

**Local tsunami** A tsunami that has little effect beyond 100 km from its source.

**Magnitude**  $m_B$ : The "broad-band" body-wave magnitude, generally based on measurements of the amplitude of P-waves with periods in the 2 to 20 s range.

$M_S$ : The surface-wave magnitude.  $M_S$  is generally based on measurements of the amplitude of the surface

(Love or Rayleigh) waves with periods of about 20 s. The US tsunami warning centers have applied a correction to the IASPEI formula that allows the estimation of  $M_S$  closer to the epicenter at a period of 20 s.

$M_E$ : The "energy magnitude" scale, derived from velocity power spectra.

$M_m$ : The mantle wave magnitude, based on the measurement of the amplitude of surface waves with periods of 50–400 s.

$M_W$ : The moment magnitude or the "work magnitude" is based on the estimation of the scalar seismic moment,  $M_0$ .

$M_{wp}$ : The moment magnitude based on the initial long period P-waves.

$M_L$ : The Local magnitude scale, based on the measurement of the maximum peak-to-peak amplitude observed on a Wood-Anderson seismogram, corrected for the decrease in amplitude with increasing epicentral distance. Generally based on the analysis of Sg, Lg or Rg surface waves oscillating with periods observed out to about 600 km. from the earthquake's epicenter.

pMag: A magnitude scale based on the average of the absolute values of the first three half cycles of the P-waves recorded at local distances.

**Marogram** A recording of sea-level variations obtained by tide gauges.

**Regional tsunami** A tsunami that has observable effects up to 1000 km from its source.

**Seismic body waves** Waves that propagate through the interior of an unbounded continuum. Primary waves (P-waves) are longitudinal body waves that shake the ground in a direction parallel with the direction of travel. Secondary body waves (S-waves) are shear waves that shake the ground in a direction perpendicular to the direction of travel. There are other types of arrivals (also known as phases) visible on seismographs corresponding to reflections of P- and S-waves from the earth's surface: The pP phase is a P-wave that travels upwards from the hypocenter and reflects once off the surface and the PP phase is a P-wave that travels downwards from the epicenter and reflects once off of the surface. The definitions of the S-wave phases follow in the same manner.

**Seismic moment** The seismic moment  $M_0$ , (expressed in units of force times distance; e. g. Newton-meters, or dyne-cm) is the moment of either couple of an equivalent double couple point source representation of the slip across the fault area during the earthquake. Mathematically, the Seismic Moment,  $M_0 = \mu Ad$ , where  $\mu$  denotes the shear rigidity, or resistance of the faulting material to shearing forces,  $A$  represents the area of the

fault plane over which the slip occurs, and  $d$  represents the average co-seismic slip across  $A$ .

**Seismic waves** Elastic waves generated by movements of the earth's crust that propagate as radiated seismic energy,  $E_R$ .

**Seismic surface waves** Waves that propagate along the surface boundary of a medium, e. g. along the surface of the earth.

**Shallow earthquake** An earthquake characterized by a hypocenter located within 100 km of the Earth's surface.

**Teletsunami** A tsunami that has observable effects on coastlines more than 1000 km away from its source.

**Tsunami** A series of water waves generated by any rapid, large-scale disturbance of the sea. Most are generated by sea floor displacements from large undersea earthquakes, but they can also be caused by large submarine landslides, volcanic eruptions, calving of glaciers and even by meteorite impacts into the ocean.

**Tsunami earthquake** An earthquake that generates a much larger tsunami than expected given its seismic moment.

**Tsunami warning system** A tsunami warning system consists of a tsunami warning center such as the Pacific Tsunami Warning Center (PTWC), a formal response structure that includes Civil Defense authorities and Government Officials, and an education program that brings a minimum level of awareness and education to the coastal populations at risk.

## Definition of the Subject

Tsunamis are among nature's most destructive natural hazards. Typically generated by large, underwater earthquakes near the Earth's surface, tsunamis can cross an ocean basin in a matter of hours. Although difficult to detect, and not dangerous while propagating in deep water, tsunamis can unleash awesome destructive power when they reach coastal areas. With advance warning, populations dwelling in coastal areas can be alerted to move to higher ground and away from the coast saving many lives. Unfortunately, due to the lack of a tsunami warning system in the Indian Ocean, the Sumatra earthquake of Dec. 26, 2004 killed over 250 000 people with thousands of lives lost as far as away as East Africa many hours after the earthquake occurred. Had a tsunami warning system been in place many lives could have been saved [77].

As fast as tsunami waves are, seismic waves can travel at speeds more than 40 times greater. Because of this large disparity in speed, scientists rely on seismic methods to detect the possibility of tsunami generation and to warn

coastal populations of an approaching tsunami well in advance of its arrival. The seismic P-wave for example, travels from Alaska to Hawaii in about 7 min, whereas a tsunami will take about 5½ h to travel the same distance. Although over 200 sea-level stations reporting in near-real time are operating in the Pacific it may take an hour or more, depending on the location of the epicenter, before the existence (or not) of an actual tsunami is confirmed. In other ocean basins where the density of sea-level instruments reporting data in near real-time is less, the delay in tsunami detection is correspondingly longer. In addition, global, regional, and local seismic networks, and the infrastructure needed to process the large amounts of seismic data that they record, are widespread. For these reasons, tsunami warning centers provide initial tsunami warnings to coastal populations based entirely on seismic data.

## Introduction

A tsunami can be produced by any mechanism that causes a sudden displacement of the ocean's surface affecting a significant volume of water. Tsunamis can be generated by undersea earthquakes, landslides and volcanic explosions, calving of icebergs, and even meteorite impacts. However, the majority of tsunamis are generated by earthquakes. Not uncommon are earthquakes that trigger landslides so that both the displacement of the crust due to the earthquake, and the landslide, each contribute to the generation and size of the tsunami. Tsunamis are a devastating, natural, high fatality hazard [18]. In the absence of a proper tsunami warning system, a destructive tsunami will cause death and destruction as it encounters coastal areas while propagating across an ocean basin as it did in the case of the Sumatra tsunami of December 2004.

Although tsunamis propagate in deep water with speeds exceeding 900 km/h they are hard to detect in the open ocean. For instance, the first wave of the great Sumatra tsunami had a wave height of only one meter in deep water ( $> 500$  m) [28], and a wavelength on the order of several hundred kilometers. Consequently, people aboard ocean vessels did not feel the accelerations caused by the Sumatra tsunami as it passed under them. However, as the speed,  $v$ , of a tsunami is governed by the simple relation  $v = \sqrt{gh}$  where  $g$  is the acceleration of gravity (in m/s), and  $h$  is the thickness of the water column (in m), the tsunami will slow down as it propagates into shallow waters. At this point, the wave speed and wavelength decrease, causing the wave height to increase. Depending on the nature of the tsunami, and the shape and bathymetry

of the coastal area, the tsunami wave height can be greatly amplified, thus magnifying its destructive power.

Because most tsunamis are generated by earthquakes, and seismic waves travel more than 40 times faster than tsunamis, the first indication that a tsunami may have been generated is the earthquake itself. Depending on the earthquake's location (undersea or inland), depth (shallow or deep) in the Earth's crust, and magnitude, a warning center may be required to issue an official message product. If the earthquake is a shallow, under sea earthquake, the severity of the message will depend upon the magnitude of the earthquake. The more rapidly and accurately the tsunami warning center can characterize the earthquake source, the quicker the initial evaluation of the tsunamigenic potential of the earthquake can be disseminated.

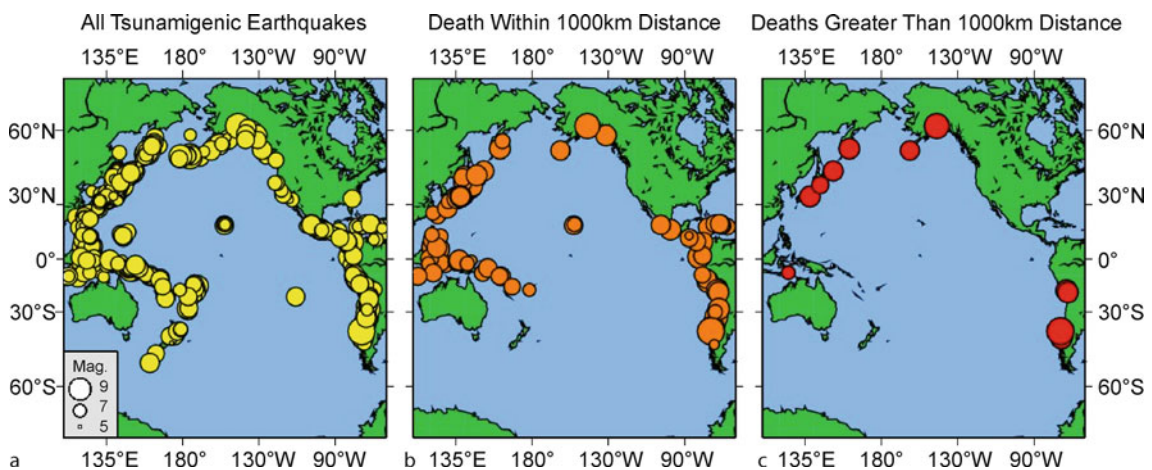
While some tsunamis are destructive, most are rather small, producing few if any casualties and little or no damage, although they are easily observable on marograms (Fig. 1). On the basis of how widespread their effects are, tsunamis can be classified as local (within 100 km of the epicenter), regional (up to 1000 km from the epicenter) or teletsunamis (greater than 1000 km from the epicenter). In the Pacific Basin there are warning centers designed to respond to tsunami threats on each of these scales.

The Richard H. Hagemeyer Pacific Tsunami Warning Center (PTWC) provides basin-wide warnings to the coastal areas of the Pacific basin. PTWC also functions as a local tsunami warning center for the Hawaii region. Other local warning centers include CPPT (French Polynesia Tsunami Warning Center) which is based in Tahiti,

GFZ-Indonesia which is based in Jakarta provides local warnings for Indonesia, and Japan's JMA (Japan Meteorological Agency) which operates Japan's tsunami warning centers provides local tsunami warnings to Japan. Examples of regional warning centers include JMA which provides regional tsunami warnings to the Northwest Pacific and the West Coast and Alaska Tsunami Warning Center (WC/ATWC) which provides regional and local warnings to the US mainland coasts and the west coast of Canada.

The tsunami warning centers themselves are not a complete tsunami warning system, they are simply the first of line of defense within the warning system. The warning system consists of three main components a) the tsunami warning centers, b) emergency management/civil defense authorities who receive tsunami warning center message products and c) a public in coastal areas that is educated in how to respond to tsunami emergencies. If any of these three components are lacking, the tsunami warning system can fail. Unfortunately, none of these components existed in the Indian Ocean at the time of the December 2004 Sumatra earthquake.

The greatest challenge for a tsunami warning system, particularly in the near field, is the slow (in terms of rupture speed) or "tsunami" earthquake. Tsunami earthquakes are so-called because they generate much larger than expected tsunamis given the size of the seismic moment of the earthquake [47]. In a well functioning tsunami warning system, residents in coastal areas are educated to immediately move inland and onto higher ground if they feel strong ground shaking and not wait for an offi-



Earthquake Source Parameters, Rapid Estimates for Tsunami Warning, Figure 1

Epicenters of tsunamigenic earthquakes occurring in the Pacific since 1 A.D. Of those earthquakes that do produce a tsunami a, most tsunamis cause no damage. Most events that cause casualties and/or damage do so within 1000 km of the epicenter b, leaving only a few great earthquake sources that generated tsunamis which caused casualties and/or damage more than a 1000 km from the epicenter, c. Data provided by the NOAA National Geophysical Data Center (NGDC), ([www.ngdc.noaa.gov/hazard/tsu.html](http://www.ngdc.noaa.gov/hazard/tsu.html))

cial tsunami alert [24]. However, because a tsunami earthquake produces much less radiated high frequency body-wave energy than normal, even a large (in terms of moment magnitude) tsunami earthquake may not be strongly felt in the near field so that this strategy of having people self-evacuate upon feeling strong ground shaking will not work. This was, unfortunately, made dramatically clear by the Java earthquake of July 17, 2006. The tsunami generated by the Java earthquake killed  $\sim 500$  people as many residents in coastal areas near the earthquake did not feel strong shaking [83]. Tsunami warning centers need to be able to properly detect the occurrence of these tsunami earthquakes.

### Tsunami Warning Center Operations

Tsunami warning center functions are much like those of a seismic observatory, i. e.: detecting, locating and characterizing the source of major earthquakes occurring around the world as fast as possible. Depending on the earthquake's location (underwater vs. inland), depth below the surface, and magnitude, tsunami warning centers may issue an official message product to advise Civil Defense/Emergency Management authorities within the warning centers AOR (area of responsibility), of the occurrence of a large earthquake and its potential for generating a tsunami. The PTWC, located in Ewa Beach, Hawaii, provides advance warning of the generation of a destructive tsunami for the Pacific Ocean Basin, and on an interim basis, for the Indian, and Caribbean ocean basins.

After consultation with the member states of the Pacific Tsunami Warning System (PTWS), the PTWC has agreed to issue tsunami bulletins for the Pacific Ocean Basin according to the criteria shown in Fig. 2.

**An Observatory (Earthquake) Message:** The PTWC sends observatory messages to certain seismological observatories and organizations for any earthquake in or near the vicinity of the Pacific, Indian, or Caribbean ocean basins for seismic events when the magnitude is larger than about 5.5. This unofficial message contains only the earthquake's epicentral location, origin time, depth, magnitude, and a list of stations used in computing these parameters. These messages contain no evaluations regarding seismic or tsunami hazard, as the magnitude of the earthquake is far too small to have a significant tsunami generation potential.

**A Tsunami Information Bulletin (TIB):** The PTWC issues this message product for any earthquake in or near the vicinity of the Pacific Basin with a magnitude in the range  $6.5 \leq M_W \leq 7.5$ . A TIB states that a destructive tsunami is not expected outside the area of the epicenter.

<b>Mw less than 6.5</b> (Mw: Moment Magnitude)	<b>Earthquake Message Only</b>
<b>Mw 6.5 to 7.5</b>	<b>Tsunami Information Bulletin</b>
<b>Mw 7.6 to 7.8</b>	<b>Regional Tsunami Warning</b>
<b>Mw &gt; 7.8</b>	<b>Expanding Warning / Watch</b>
<b>Confirmed Teletsunami</b>	<b>Pacific-Wide Warning</b>

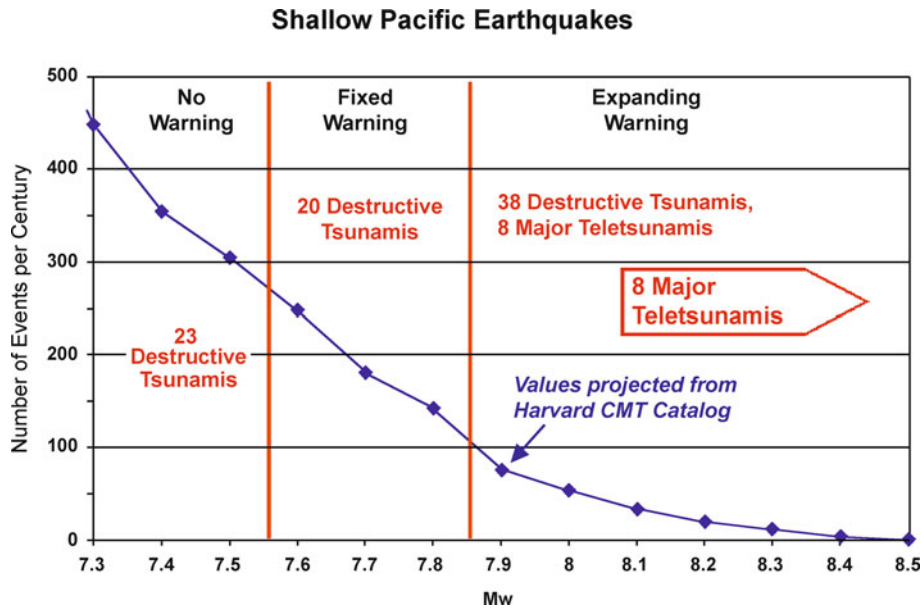
Earthquake Source Parameters, Rapid Estimates for Tsunami Warning, Figure 2  
PTWC Bulletin Criteria for the Pacific Basin

However, it does warn of the possibility of a destructive tsunami along coastlines within 100 km of the epicenter.

**A Fixed Regional Warning Bulletin:** The PTWC issues this message product for shallow underwater earthquakes (depth  $< 100$  km) with a magnitude in the range  $7.6 \leq M_W \leq 7.8$ . This bulletin warns of the possibility of a regionally destructive tsunami within 1000 km of the epicenter. All regions within 1000 km of the epicenter are thus initially placed in a warning.

**An Expanding Watch/Warning Bulletin:** The PTWC issues this message for shallow underwater earthquakes with magnitude  $M_W \geq 7.9$ . This criteria is similar to the fixed watch/warning with the exception that this bulletin also warns of the possibility of a destructive tsunami traveling greater than 1000 km away from its source area. The use of the term "expanding" stems from the fact that the watch and warning regions expand across the Pacific as time progresses until the watch/warning is canceled. The extensions of the watch/warning area are referenced to the leading edge of the tsunami waves at the time the bulletin is issued. Areas within 3 to 6 h tsunami travel-time from the predicted current leading edge of the tsunami are placed in a watch. Areas within 3 h tsunami travel-time are placed in a warning. All other areas are placed in an advisory. Because of the expanding nature of the watch/warning, areas that were initially only placed in an advisory may eventually come to fall into the watch or warning region. If no potentially destructive tsunami is detected by sea-level stations, the watch/warning is canceled. On the other hand, if the data provided by sea-level stations provide evidence that a destructive tsunami





Earthquake Source Parameters, Rapid Estimates for Tsunami Warning, Figure 3  
Retroactive performance based on current bulletin criteria

is moving across the Pacific, the PTWC may upgrade to its most severe message, the Pacific-Wide Warning. A Pacific-Wide Warning is a tsunami warning for all coasts in the Pacific Basin. Before issuing a Pacific-Wide Warning the scientist on duty must confirm the presence of a potentially dangerous tsunami on sea-level instruments.

Figure 3 summarizes our response in retrospect, had these criteria been in place over the 20th century, after applying them to the earthquakes and tsunami that occurred during this period. The application of these criteria would have resulted in the issuance of a TIB, but no warning for 23 destructive locally generated tsunamis occurring over the last century in the Pacific Basin. At larger distances from the earthquake, the PTWC would have issued a *Fixed Regional Warning Bulletin* ahead of 20 destructive tsunamis generated in the last century, an *Expanding Watch/Warning Bulletin* for thirty-eight destructive tsunamis, as well as eight major Pacific Basin wide tsunami warnings within the same time period.

Coastlines close to the earthquake epicenter can experience tsunami waves within two to fifteen minutes after the earthquake; hence a local tsunami warning needs to be issued within a few minutes to be effective. This requires access to real-time data provided by a dense local network of seismic stations near the epicenter that allows both, the rapid location, and source characterization of the earthquake. In the case of the Hawaii region, the PTWC uses data from its own seismic network, and from the dense

seismic network maintained by the USGS Hawaii Volcano Observatory (HVO) to rapidly detect Hawaii earthquakes. These data enabled PTWC to issue an information bulletin to the state of Hawaii, and to the Pacific Basin for the Kiholo Bay earthquake ( $M_w$  6.7) within 3 min of the origin time of the earthquake [42]. However, without access to dense local seismic networks around the Pacific rim, the PTWC is unable to issue timely warnings for populations in the immediate vicinity of large earthquakes outside of Hawaii. As a result, the PTWC does not have the capability of functioning as a local warning center for areas outside of Hawaii.

The WC/ATWC of the US National Weather Service has access to dense local seismic networks on the US mainland, Puerto Rico, and Canada, and can therefore provide rapid warnings to the US and Canadian West Coast as well as to Puerto Rico and the Virgin Islands. Japan has a similar capability for its coastlines, and spurred on by the December 2004 Sumatra earthquake, several other nations such as Indonesia and New Zealand, for example, are rapidly developing and improving their seismic networks in an effort to improve their tsunami warning capabilities.

### Seismic Methods

To rapidly detect, locate, and characterize the source of earthquakes occurring around the world, tsunami warning centers rely on the Global Seismic Network (GSN

USGS/IRIS) which has many contributors in the US and worldwide. It is this unfettered access to real time seismic data supplied by a number of different networks that makes a basin-wide tsunami warning center possible. To rapidly deal with the threat posed by locally generated tsunamis to the state of Hawaii, PTWC processes seismic data from about 70 stations located in the Hawaiian Islands. The USGS HVO's dense network supplies most of this data. The US tsunami warning centers use the Earthworm software developed by the USGS to import and export seismic data [46].

PTWC duty scientists can receive automatic pages at any time, for any earthquake with magnitude  $M_W$  above  $\sim 5.5$ . The system generating these pages combines Evan's and Allen's [23] teleseismic event detection algorithm, adapted for broadband data by Wither's [84], and Whitmore's [81] teleseismic picker and associator. In the Hawaiian Islands, the application of Hirshorn and Lindh's [40] algorithm notifies duty scientists for earthquakes with magnitudes larger than about 3.5 within 10 to 20 s of the earthquakes origin time. Other software automatically locates the event, and provides a first estimate of the earthquake's magnitude, and other source parameters in real time [3,39,40,45]. PTWC duty scientists then refine and supplement the software's automated real-time hypocenter location and magnitude estimates. Determining the earthquake's depth is particularly important as earthquakes occurring at depths greater than about 100 km generally don't cause tsunamis.

Earthquakes are located using P-wave arrival times recorded at a number of seismic instruments. As both the locations of seismic instruments and P-wave travel-times as a function of distance are well known, a process analogous to triangulation is used to locate the earthquake. The pickers and associators perform these functions automatically on a continuous basis.

While the depth of the earthquake can be estimated on the basis of P-times alone, a more robust result often requires the addition of depth phases such as pP which is a P-wave that travels directly up to the earth's surface from the earthquake source and reflects once off of the Earth's surface before arriving at the seismometer. The duty scientists use pP arrival times to refine hypocentral depths of distant earthquakes (teleseisms). For earthquakes in Hawaii, observed at local distances, the S-wave arrival time would be useful for constraining earthquake depth. However, automatic picking of S-wave arrival times is not yet robust, and manual picking and incorporation of S-wave arrival times into the analyzes by a duty scientist would take too long in the local earthquake case. In addition, as the deepest earthquakes in Hawaii have a hypocenter lo-

cated about 50 km deep, (Fig. 6) which is comparable to the rupture length of an earthquake over about magnitude seven, the PTWC's local tsunami warning criteria are currently based on magnitude only.

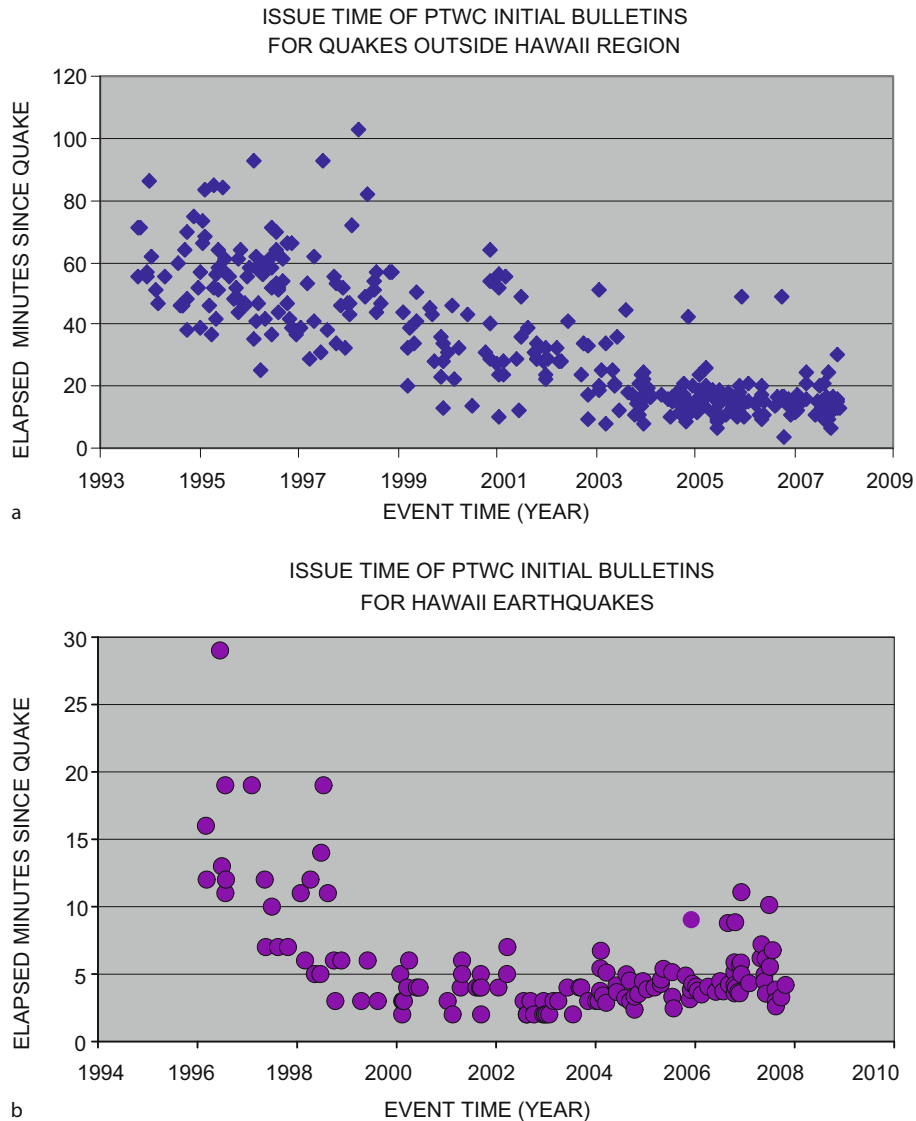
Seismologists use a panoply of different magnitudes to characterize the seismic source. These different methods examine different parts of the seismic wave train, such as short and long period body waves (seismic waves that travel through the earth's interior like the P- and S-waves) and long period surface waves (slower seismic waves that are constrained to travel along the earth's surface). Most of these magnitude scales were developed to estimate the same quantity; the energy released by the earthquake as radiated seismic wave energy,  $E_R$ . Traditional magnitude measures such as  $M_L$  [66], and  $m_b$  (a shorter period variant of  $m_B$  Gutenberg [30,31] that examines high frequency body waves). The Gutenberg surface wave magnitude  $M_S$  [9], modified by Vanek [75] as well as the newer  $M_m$  Okal and Talandier [63], or mantle magnitude are derived from the surface waves. A relatively new and quick method,  $M_{wp}$  analyzes long period P-waves [73,74,82]. The  $M_{wp}$  magnitude is now the magnitude used in the decision process for deciding which if any official message product to issue, supplanting the  $M_S$  method which had been used for over 50 years. For large earthquakes, duty scientists also routinely estimate  $M_m$ , a very long period surface wave magnitude based on mantle waves with periods in the range 50–410 s [63]. The relationship between these magnitudes, each looking at different parts of the seismic wave spectrum of an earthquake, can be used to characterize the earthquake source [2,16,17].

When evaluating the tsunamigenic potential of an event, PTWC duty scientists also compute the quantity  $\log_{10}(E_R/M_0)$ , known as "Theta",  $\Theta$ , where  $M_0$  is the seismic moment [1]. Newman and Okal [62] showed that  $\Theta$  is anomalously small for tsunami earthquakes.

Since about the mid 1990's the two US Tsunami Warning Centers response times to potentially tsunamigenic teleseisms has decreased dramatically due to the much larger amounts of seismic data that they now receive, and to the switch from the slower  $M_S$  magnitude method to the faster  $M_{wp}$  moment magnitude method for their initial messages (Fig. 4a). For local events, using the real-time associator binder\_agl [45], and the very fast  $pMag$  scale [40], has brought the PTWC's response time down to less than 3 min (Fig. 4b).

## Earthquake Source Parameters

A fundamental problem with traditional magnitude estimates such as  $M_L$ ,  $m_b$ , and  $M_S$ , is that they are based



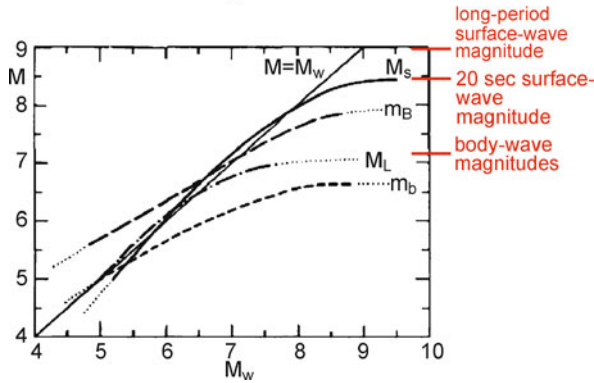
Earthquake Source Parameters, Rapid Estimates for Tsunami Warning, Figure 4  
 Elapsed time from earthquake origin time to issuance of first official message product for **a** earthquakes outside the Hawaii Region and **b** for earthquakes within the Hawaii Region

on the amplitudes of relatively short period seismic waves with periods usually less than 3 s for  $m_b$  and  $M_L$ , and about 20 s for  $M_S$ . When the largest rupture dimension of the earthquake exceeds the wavelength of these seismic waves, which is about 50 km for the 20 s period surface waves used for  $M_S$  [48,49], these magnitude values will start to “saturate”. Saturation in this case means that these magnitude measures will underestimate the true size of the earthquake when the periods of the amplitudes on which they are based are shorter than the corner period of the earthquake’s seismic wave spectrum [2,16,17] (see Fig. 5).

Another, equivalent explanation is that these magnitude methods, which look at waves with periods of a fraction of a second to a few tens of seconds, cannot sample enough of the energy released by an earthquake whose source duration (the length of time over which the rupture occurs) is many times larger than the periods used by these methods. As the earthquake becomes very large, one needs to examine longer period waves to avoid saturation.

K. Aki used a spectral representation to establish that earthquakes of varying size had spectra of similar shape, differing primarily in the low frequency ampli-

### Saturation of Magnitude Scales



Earthquake Source Parameters, Rapid Estimates for Tsunami Warning, Figure 5

Saturation of different classical magnitude scales with respect to non-saturating moment magnitude according to Kanamori [50]. Note that  $m_B$  refers to the original Gutenberg–Richter [32,33] body-wave magnitude scale based on amplitude measurements made on medium-period broadband instruments. It saturates at larger magnitudes when compared to the short-period based  $m_b$ ,

tude, proportional to seismic moment, and the location of the “characteristic frequency” (corner frequency of the source spectrum) which he related to the characteristic length scale of an earthquake [2]. Subsequent studies by Brune [16,17], and Savage [68] also related the corner frequency to the dimensions of the fault plane.

To circumvent the saturation problem, Kanamori defined a new magnitude scale, the moment magnitude  $M_w$ , [48], in terms of a minimum estimate of the total co-seismic strain energy drop,  $W_0$ , via Gutenberg and Richter’s energy-magnitude relationship [33]. The  $M_w$  scale, more properly a magnitude that describes the total “work” required to rupture the fault, is computed from the seismic moment,  $M_0$ , assuming 1) that the stress changes associated with large earthquakes are approximately constant, and 2) that the stress release during an earthquake is about the same as the kinetic frictional stress during faulting.  $M_w$  and its agreement with the  $M_L$  and  $M_s$  magnitude scales in their unsaturated ranges was discussed by T.C. Hanks and H. Kanamori [34] while Kanamori [50] discusses additionally the average relationship of  $M_w$  with  $m_b$  and  $m_B$ , also in the range where these magnitudes saturate (see Fig. 5).

### Traditional Amplitude Based Magnitudes at the PTWC

**Local Earthquake Magnitude Methods** Hirshorn and Lindh [40] developed a short period P-wave magnitude scale, called  $pMag$ , which is based on the average of the ab-

solute values of the amplitudes of the first three half-cycles of the initial p-waves recorded, at local distances, on short period seismometers [40,45]. The  $pMag$  scale is based on the assumption that the decrease of locally recorded initial P-wave amplitudes with increasing hypocentral distance shares a common decay curve in a given geographic area, independent of the magnitude of the earthquake. Lindh and Hirshorn incorporated  $pMag$  into Carl Johnson’s [46] local p-wave associator, binder\_agl, enabling automatic pages, containing the hypocentral parameters and a the lower bound magnitude estimate provided by  $pMag$ , within about 10 to 20 s of an events origin time. At the PTWC, the System for Processing Local Earthquakes in Real Time (SPLERT) is based on this software.

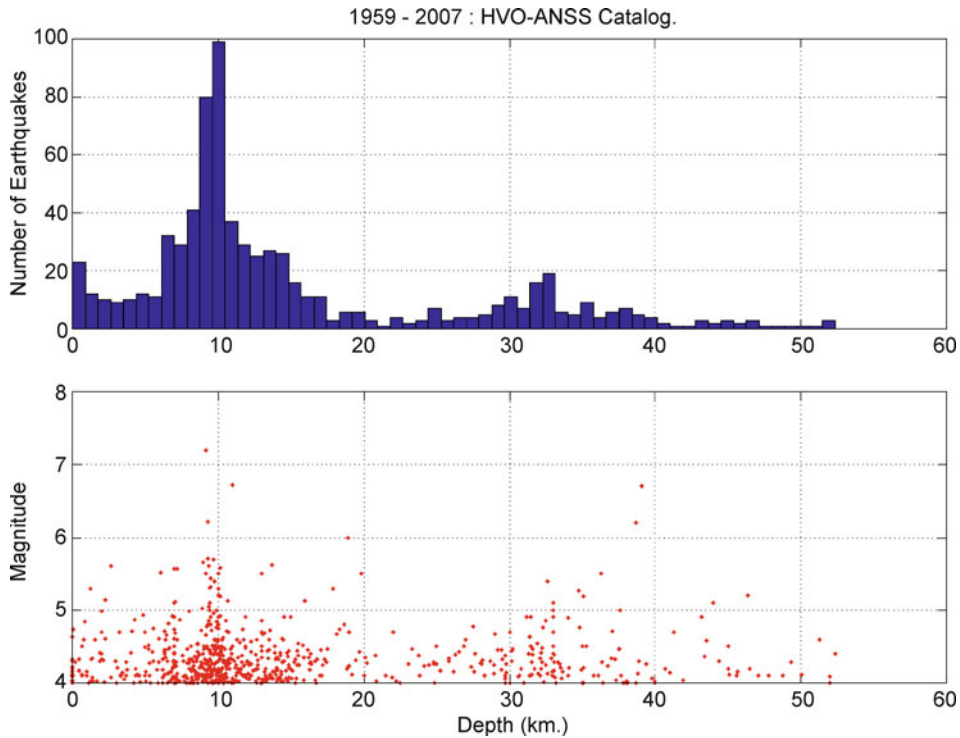
PTWC also uses a very band-limited  $M_L$  scale, based on the maximum amplitudes measured on the horizontal components of short period seismograms recorded at local hypocentral distances from earthquakes that occur in Hawaii. These short period waves attenuate about 6 times less along the path between the hypocenter and recording stations for the “deeper” population of Hawaiian earthquakes (with hypocenters (Fig. 6) located between 20 and 50 km depth), then they do along the path from source to receiver for the “shallower” event population in the lower oceanic crust (about 10–20 km below the sea surface). For this reason, PTWC adds 0.8 magnitude units to the  $M_L$  values obtained for events with hypocentral depths  $\geq 25$  km.

Because of the bimodal depth distribution of Hawaii earthquakes (Fig. 6.) our  $M_L$  calculation requires only a good enough depth estimate to discriminate between these two populations.

**Teleseismic Magnitude Methods** PTWC’s body wave magnitude method is called  $bMag$  and has similarities to the intermediate period broadband body wave  $m_B$  magnitude as defined by IASPEI. The IASPEI  $m_B$  [11,16] is based on Gutenberg’s [30,31] and Gutenberg and Richter’s  $m_B$  [32,33].  $bMag$  uses a 90 s window of broadband vertical component seismogram starting 30 s prior to the arrival of the P-wave. This window is band-pass filtered between .3 s and 5 s. The largest amplitude and its period found in the 60 s after the first P-wave arrival are chosen for use in the magnitude formula. In PTWC’s implementation, the formula used is the same as Gutenberg and Richter’s [32,33] relation, adopted by IASPEI for  $m_B$ :

$$bMag = \log(A_{\max}/T_{\max}) + Q(\Delta, z),$$

where  $\Delta$  is the epicentral distance ( $15^\circ \leq \Delta \leq 90^\circ$ ),  $z$  is the hypocentral depth,  $A_{\max}$  is the maximum wave amplitude obtained from the band-pass filtered record, and  $T_{\max}$



Earthquake Source Parameters, Rapid Estimates for Tsunami Warning, Figure 6

This figure shows the bimodal depth distribution of Hawaii earthquakes of  $M \geq 4$ , taken from the USGS Hawaiian Volcano Observatory (HVO) ANSS Catalog. The *top figure* is a histogram of all events binned by hypocentral depth. The *bottom figure* shows the magnitude vs. hypocentral depth for each event

is the period of the wave with the maximum amplitude. Gutenberg and Richter's [33] table of  $Q(\Delta, z)$  is used to provide the distance and depth correction. The largest amplitude found in the 30 s prior to the P-wave arrival time is used as the basis for the signal to noise ratio.  $bMag$  differs from the IASPEI  $m_B$  in three respects,

1.  $m_B$  uses the largest amplitude wave in the P-wave coda up to the arrival of the PP phase,
2.  $m_B$  uses a slightly different distance range ( $15^\circ \leq \Delta \leq 90^\circ$ ), and
3. for  $m_B$ , the seismogram is band-pass filtered using the band .3 s to 5 s.

$bMag$  will saturate at lower magnitudes than  $M_S$  does, so it is of limited use for large earthquakes. However,  $bMag$ , is still useful for three main reasons

1. unlike  $M_S$ ,  $bMag$  has a correction for the depth of the event's hypocenter,
2. it is useful for determining the magnitude of moderate earthquakes that occur as aftershocks of much larger earthquakes, e.g. when longer period energy is

still present in the signal from an earlier, larger event, that can adversely affect magnitude methods based on longer periods, and

3. by comparison with magnitudes based on longer periods, such as  $M_w$ , it can also provide a way to detect slow or tsunami earthquakes.

In computing  $M_S$  (first proposed by Gutenberg [9] and later revised by Vanek et al. [75]) at the PTWC, we band-pass filter 14 min of the broadband velocity seismogram with a 7 s band, from 16 to 23 s, starting 3 min before the expected arrival time of the surface waves. We then apply the following equation, similar to the IASPEI [11,75] formula:

$$M_S = \log_{10}(A_{\max}/T_{\max}) + 1.66 \log_{10}(\Delta) + 3.3 + \text{correction}.$$

The *correction* term is 0 for epicentral distances,  $\Delta$ , greater than  $16^\circ$ , and  $0.53 - 0.033\Delta$  for  $\Delta$  less than  $16^\circ$ . Note that the IASPEI implementation considers a much greater range in periods from 3 to 60 s for  $T_{\max}$  and has no need for the *correction* term. The *correction* term allows the US TWC's to compute the  $M_S$  magnitudes from stations as

close as 600 km to the epicenter at a period of 20 s. This method is susceptible to saturation effects as the magnitude reaches the high 7's.

Although  $M_S$  is no longer used as the basis for issuing bulletins, it is still helpful in diagnosing deep earthquakes, and for comparing the amount of 20 s radiated energy to the amounts of radiated energy at other periods. Deep earthquakes do not excite large surface waves. Hence if  $bMag > M_S$  the hypocenter is likely to be deep.

### The $M_{wp}$ Method

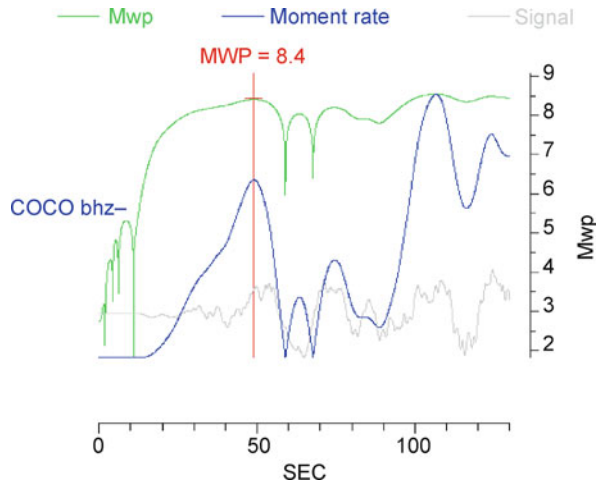
The broadband P-wave moment magnitude,  $M_{wp}$ , has replaced  $M_S$  as the magnitude upon which the US TWCs initial tsunami messages are based [73,74,82]. This is because  $M_{wp}$  uses P-waves, recorded at any epicentral distance, up to about  $90^\circ$ , when the observed initial P-waves are affected by refraction due to the earth's outer core.  $M_{wp}$  is obtained much quicker than  $M_S$ , which is based on the slower traveling surface waves, and because  $M_{wp}$  examines much longer period waves than the 20 s surface used by  $M_S$  making it less susceptible to the saturation effects discussed above.  $M_{wp}$ , as implemented at the PTWC, uses the first 120 s of the vertical component, broadband velocity seismogram, beginning at the P-wave arrival time (gray trace in Fig. 7).

The derivation of  $M_{wp}$  assumes that we can obtain the seismic moment,  $M_0$ , from the far-field P- and/or pP-wave portion of the vertical broadband displacement waveform,  $u_z(x_r, t)$ , using

$$M_0 = (4\pi\rho\alpha^3 r/F^P) \text{Max} \left| \int u_z(x_r, t) dt \right| ,$$

where  $\rho$  and  $\alpha$  are the density and P-wave velocity averaged along the propagation path,  $r$  is the epicentral distance, and  $F^P$  is the earthquake source radiation pattern [73,74,82]. At the PTWC, we follow Tsuboi [73], approximating  $\text{Max} \left| \int u_z(x_r, t) dt \right|$  by the first significant or "big" peak in the absolute value of the integrated displacement record. We prefer to use velocity seismograms,  $v(t)$ , from STS-1 or KS54000 broadband seismometers as there is then no need to deconvolve the instrument response from the data. We simply scale the data by a gain factor, because we can assume that the instrument response function is flat in the frequency band of interest. For the STS-1, or the KS54000, which both have a flat velocity up to about 350 s, this works for all but the very largest or slowest earthquakes.

We first remove any pre-event offset from  $v(t)$ , ending before the P-waves from the earthquake arrive, inte-



Earthquake Source Parameters, Rapid Estimates for Tsunami Warning, Figure 7

The first 2 min of the broadband vertical velocity seismogram (gray) recorded by the GSN USGS/IRIS broadband station COCO, on Cocos Island, about 15 degrees south of the epicenter of the  $M_W$  9.2 Sumatra Earthquake of Dec. 26, 2004. Note that this portion of the broadband velocity seismogram is not clipped. This instrument, a KS54000, has a flat frequency response to velocity to a period of about 350 s. The blue trace is the integrated displacement record (doubly integrated velocity), and the green trace is  $M_{wp}$  as a function of time

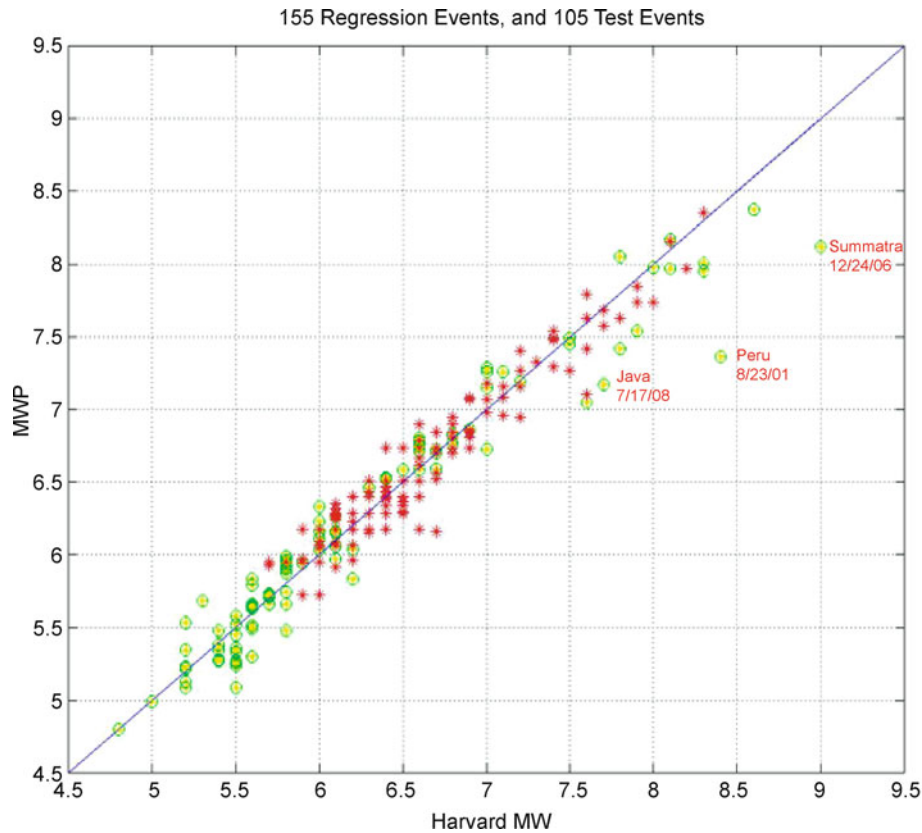
grate  $v(t)$  twice, and then multiply the absolute value of each data point by  $4\pi\rho\alpha^3$  to obtain  $M_0(t)$  in N-m (the blue trace in Fig. 7). We then apply the standard IASPEI moment magnitude formula [34]:

$$M_W = (\log_{10} M_0 - 9.1)/1.5$$

to  $M_0(t)$ , to calculate  $M_W(t)$  (the green trace in Fig. 7). To correct for the radiation pattern,  $F^P$ , we then add 0.2 to the average of the individual  $M_{wp}$  values, each obtained at different azimuths and distances from the epicenter. This is because  $\int (F^P)^2 d\Omega = 4/15$ , where  $\Omega$  is the azimuthal angle of the observation around the epicenter, and  $\sqrt{4/15} = 0.52$ . Therefore, we multiply the averaged  $M_0$  by 2, which is equivalent to adding 0.2 to  $M_{wp}$ . Finally, we apply the Whitmore et al. [82] magnitude dependent correction,  $M_{wp} = (M_{wp} - 1.3)/0.843$ , to get a final value for  $M_{wp}$ .

Figure 8 compares these final  $M_{wp}$  values resulting from this procedure with the Harvard moment magnitude  $M_W$  estimated from their CMT [19] solutions.

For some complex earthquakes, such as the  $M_W$  8.4 [19] Peru earthquake of June 21, 2001, or the great  $M_W$  9.2 [5,65,71] Sumatra earthquake of December 2004,  $M_{wp}$  (7.4 and 8.1, respectively) will underestimate  $M_W$ ,



Earthquake Source Parameters, Rapid Estimates for Tsunami Warning, Figure 8

A Scatter plot of average  $M_{wp}$  versus the  $M_W$  [19] for a set of 260 earthquakes with magnitudes in the interval  $4.8 \leq M_W \leq 9.2$  occurring from 1994 through October of 2007. Whitmore et al. [82] found that with the application of an empirical correction made to the results, satisfactory results could be obtained. Our linear, least squares fit to the 155 earthquakes (red stars) yielded a slope of 0.83, close to the Whitmore et al. [82] slope of 0.84. Since April of 2002, we have used this corrected relationship (green circles filled in yellow) to calculate  $M_{wp}$

when the first moment release is not the largest. In contrast, the PTWC's final estimate of  $M_{wp}$  8.4 for the  $M_W$  8.6 [19] Nias event of March 28, 2005 was acceptable, as it was for 9 other earthquakes in the range  $8.0 \leq M_W \leq 8.4$  (Fig. 8).

The PTWC also uses  $M_{wp}$  for large local earthquakes, occurring in the Hawaiian Islands [41,42].  $M_{wp}$  is based on the far field formulation for P-wave displacements [73]. For earthquakes whose largest source dimension is small compared to the distances at which the P-waves are observed, this assumption is satisfied. Up to approximately  $M_W$  7.5,  $M_{wp}$  calculated from these locally recorded, far field P-waves agrees well with the Harvard  $M_W$  values for the same events [41,42]. For example the PTWC calculated a value of  $M_{wp}$  6.5 for the  $M_W$  6.7 [19] Kiholo Bay event within two minutes of the initiation of rupture at the hypocenter [41,42].

### The Mantle Magnitude ( $M_m$ ) Method

Emile Okal and J. Talandier developed the  $M_m$  method in 1989 [63]. The mantle magnitude is related to the moment magnitude via the simple expression  $M_W = M_m/1.5 + 2.6$ . This work was inspired by the need to develop a magnitude method for tsunami warning centers that would not suffer the saturation problem of  $M_S$  [64]. Not only may  $M_S$  saturate as the magnitude becomes large ( $> 8$ ) but slow earthquakes can cause  $M_S$  to be seriously deficient and  $bMag$  even more so. Severely underestimating the magnitude of an earthquake can lead to a failure to warn. PTWC's implementation of the  $M_m$  method is based on analyzing Rayleigh waves obtained on vertical component seismograms.

$M_m$  being based on slow traveling long period surface waves, is available too late to be used in the decision pro-

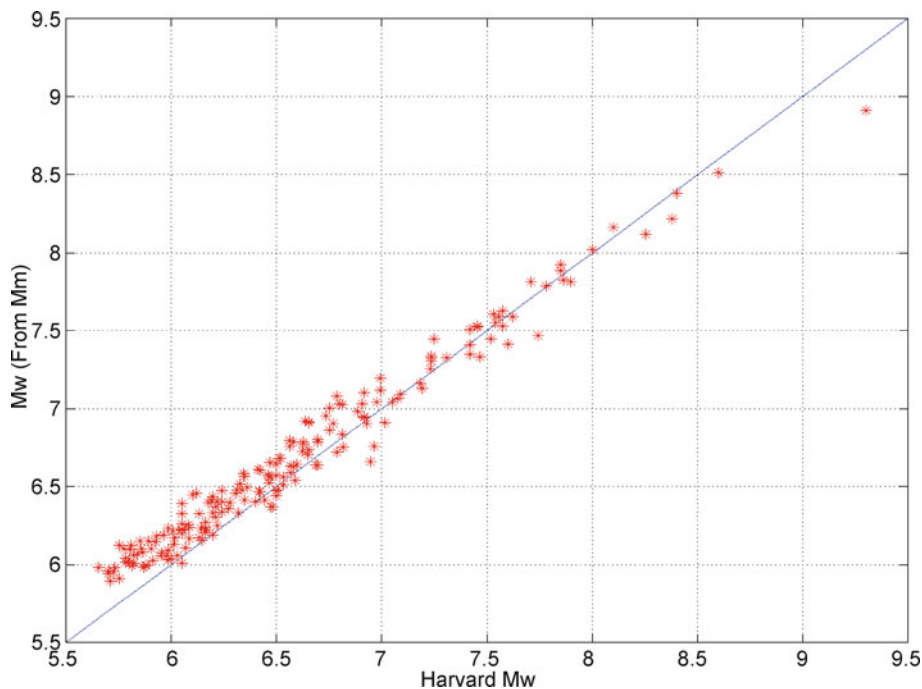
cess for issuing an initial bulletin. Notwithstanding, it does provide a useful check on the magnitude obtained from the  $M_{wp}$  method and if there is a discrepancy between  $M_{wp}$  and  $M_W(M_m)$  on the order of 2–3 tenths or more in the 7+ magnitude range, the duty scientist may instead use the results of the  $M_m$  method in subsequent bulletins. The  $M_m$  method overcomes the limitation of saturation because it is a variable period magnitude. Multiple values of  $M_m$  are routinely computed for a number of fixed periods ranging from 50 to 270 s for each station. Because  $M_m$  may saturate at the smaller periods for great earthquakes, while at longer periods  $M_m$  will be unsaturated, Okal and Talandier's [63] procedure was to choose the largest  $M_m$  mitigating the effects of saturation.

$M_m$  is more complicated than the other methods described here as it uses frequency domain deconvolution. This can cause problems due to deconvolution noise at low magnitudes, where the amplification of noise by the deconvolution process at long periods may result in spurious magnitudes. Thus  $M_m$  works best with very long period broadband seismometers such as the KS54000's, KS36000's and STS-1's. While STS-2 seismometers tend to do well, however, the shorter period broadband seismometers tend to behave poorly at the longest periods [79]. Using the maximum  $M_m$  obtained for each sta-

tion proved to be suboptimal due the heterogeneous distribution of instruments coupled with the total automation of the procedure at the PTWC. Weinstein and Okal [79] devised a sampling method that alleviates most of these difficulties in PTWC's implementation of  $M_m$ .

The December 2004 Sumatra earthquake showed that for earthquakes with an unusually long source duration (in this case  $\sim 600$  s), even  $M_m$  at 270 s will saturate. Hence PTWC's  $M_m$  implementation will now automatically extend the period range to 410 s when the magnitude exceeds 8.0. At 410 s  $M_W(M_m)$  is 8.9 [79] for the December 2004, Sumatra earthquake, still deficient, but a marked improvement over the moment magnitude 8.5 obtained by PTWC and 8.2 obtained by the USGS (NEIC Fast Moment Tensor) on Dec. 26, 2004.  $M_m$  normally uses a 660 s window of the surface wave train, but when  $M_m$  exceeds 8.0, the window expands to 910 s. Given the mix of instruments and their distribution used at PTWC, and the effects of broadband deconvolution noise, Weinstein and Okal [79], found that the  $M_m$  method was not useful for  $M_W < 6.0$ .

Figure 9 compares  $M_W(M_m)$  values obtained for more than 200 recent earthquakes with the respective  $M_W$  values of HRV/GCMT [19] for the same events. PTWC's implementation of  $M_m$  of still tends to over estimate  $M_W$  by about .15 magnitude units for  $M_W < 7.0$ .



Earthquake Source Parameters, Rapid Estimates for Tsunami Warning, Figure 9  
Scatter plot of  $M_W(M_m)$  vs. Harvard/GCMT [19]  $M_W$  for over 200 Earthquakes



**Rupture Slowness Estimation (Theta Program)**

One way in which the occurrence of a tsunami earthquake may be indicated is if  $bMag$  and/or  $M_S$  are significantly smaller than  $M_W$  obtained from the longer period P-waves, or longer period mantle waves. This is made clear in Fig. 10. Note the population of 4 tsunami earthquakes that fall well off the trend. The short period magnitudes may also simply be deficient simply due to the size of the earthquake. Measuring the “rupture slowness” of the earthquake can further aid the warning centers in deciding between the two possibilities. As can be determined from Fig. 10, the body wave magnitude for July 2006, Java earthquake was deficient by nearly 1.5 magnitude units.

As mentioned earlier, a fundamental characteristic of a tsunami earthquake is the slowness of the rupture speed. Newman and Okal [62] showed that the log ratio of the radiated energy  $E_R$  [9,10], to the seismic moment  $M_0$ ,  $\text{Log}_{10}(E_R/M_0)$  (also denoted by Theta, or “ $\Theta$ ”) is anomalously small for tsunami earthquakes. A number of factors can affect this ratio such as rupture velocity, stress-drop/apparent stress, fault plane geometry, maximum strain at rupturing, and directivity (bi-lateral vs. unilateral rupture). However, for shallow thrust, low stress-

drop subduction zone earthquakes, unusually slow rupture velocity may have the largest influence on the value of  $\Theta$ .

Newman and Okal [62] showed that for tsunami quakes, the value of  $\Theta$  is usually about  $-6.0$  or less. For an earthquake with a unilateral rupture with nominal speed ( $\sim 3$  km/s), theory suggests that  $\Theta$  is about  $-4.9$  [26,69,76]. Weinstein and Okal [79] extended the original dataset of Newman and Okal [62] by including an additional 118 earthquakes. The mean value of all  $\Theta$  values is approximately  $-5.1$ . However, when averaged by event, the distribution of  $\Theta$ 's peaks precisely at  $-4.9$ , in accordance with theoretical expectations. Given the standard deviation of  $0.39$  for all  $\Theta$ 's (for the 118 earthquakes), Weinstein and Okal [79] found that values of  $\Theta$  around  $-6.0$  or below are more than  $2\sigma$  off the mean and hence clearly anomalous.

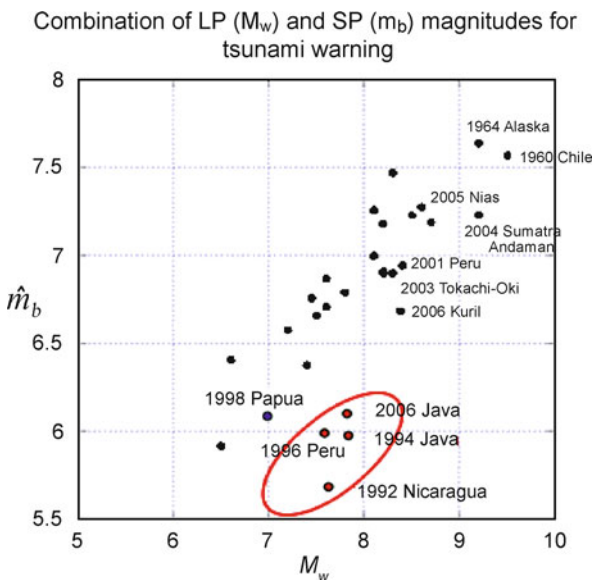
The PTWC uses broadband vertical component seismograms, obtained in the distance interval  $25^\circ \leq \Delta \leq 90^\circ$ , to compute  $\Theta$ . A window of  $75$  s is used starting approximately  $5$  s before the P-wave arrival. This is done to insure that the first arrivals are not missed by the integration. This window is deconvolved with the instrument response and the radiated energy contained between  $.1$  Hz and  $2$  Hz is computed.

In general it is thought that anomalously slow rupture speed is due to either low rigidity sediments in the fault or faulting through an accretionary prism [6,8,25,47,52,67]. In either case, the small shear rigidity associated with weak materials retards the rupture speed. As to why “slow” earthquakes produce more destructive than expected tsunamis, one can look at the well-known relation for moment magnitude:

$$M_0 = \mu Ad,$$

where  $\mu$  is the shear rigidity,  $A$  is the fault plane area and  $d$  is the average slip over the fault plane. Thus for two quakes with the same  $M_0$  and all other properties equal except for  $\mu$  and hence rupture speed, the slow quake requires a correspondingly larger slip,  $d$ , in order to achieve the same moment as the earthquake with a nominal value of  $\mu$  and hence normal rupture speed.

One problem with  $\Theta$ , is that it can be misleading and occasionally yield false indications of rupture slowness. This was made apparent by the Peru earthquake of June 23, 2001. This earthquake began with a initial event that had a moment magnitude of approximately  $7.4$ , followed almost  $60$  s later by a much larger event, which had a moment magnitude of almost  $8.4$ . [7,27,55]. Due to the  $60$  s delay, the  $\Theta$  computation used mainly P-wave coda from the first shock, and little if any energy from the main



Earthquake Source Parameters, Rapid Estimates for Tsunami Warning, Figure 10

Comparison between short-period  $\hat{m}_b$  [43] and  $M_W$  for earthquakes with  $M_W > 6$ . Note the cluster of red dots representing tsunami earthquakes. This illustrates the diagnostic potential of short-period/long-period magnitude ratios to identify unusually slow earthquakes with high tsunami potential (Kanamori 2007, Talk at the PTWC in April of 2007)

shock. As a result the PTWC initially obtained a  $\Theta$  of  $-6.1$ , using a moment based on  $M_m$ , making this earthquake appear very slow indeed. However, this result is spurious and was due to the complexity of the earthquake itself and not to actual slowness of the rupture.

Weinstein and Okal [79] found that by sliding the window over which  $\Theta$  is computed forward in time,  $\Theta$  would increase as the  $\Theta$  window overlapped with the occurrence of the main event of the Peru 2001 earthquake. Indeed for a window offset of 70 s,  $\Theta$  increases to  $-5.6$ , which is a strong trend to slowness, but not a slow or tsunami quake. This was further borne out by the size of the tsunami, which while detected on sea-level instruments around the Pacific (more than 2 m peak-to-peak in Chile), was not destructive outside of Peru.

Weinstein and Okal [79] explored the windowing technique and found that in actuality, it was a more comprehensive method than the single determination of  $\Theta$  (zero offset). Computing theta in a succession of windows separated in time by 10 s (each window spanning 70 s) up to 100 s post P-wave arrival yields a better method of detecting slowness (see Fig. 11). What Weinstein and Okal [79] found is that for true tsunami earthquakes, the variation of  $\Theta$  with offset time was small, generally no more than 0.1 log units over the entire 100 s. It is this flat trend that is probably the best discriminant for tsunami earthquakes.

In effect, the curve resulting from the window-offset technique tells us something about the source duration of the earthquake. Gigantic earthquakes have long source



Earthquake Source Parameters, Rapid Estimates for Tsunami Warning, Figure 11  
 The variation of  $\Theta$  with offset for a “normal” earthquake, b a “slow earthquake” (Java, 2006), and c a complex earthquake (Peru, 2008) respectively. In these plots  $\Theta$  is de-meaned (the mean is found on the bottom right of the plot in black) and the number next to the dots indicates how many stations were used in computing that value. These plots are taken from PTWC’s operational software

durations, and slow earthquakes have anomalously long source durations for their seismic moment. Therefore  $\Theta$  can be viewed as a measurement of how anomalous the source duration is in terms of whether the earthquake is anomalously slow, or simply anomalously large. It turns out that in the case of the Sumatra earthquake of December 2004,  $\Theta$  has little variation, even when the integration window is increased to 200 s and the offset carried out to 300 s. The magnitude of  $\Theta$  based on PTWC's  $M_W(M_m)$  of 8.5 was  $\sim -5.6$ , a trend to slowness, but not slow. Using the  $M_W$  based on normal mode studies,  $\Theta$  is  $\sim -6.1$  (with a 200 s integration window!) and discussion continues to the current day as to whether or not the Sumatra earthquake of 2004 was slow, simply had aspects of a tsunami earthquake, or none at all [5,20,53,57,70].

### Future Directions

Given the availability of high quality broadband seismic data, the tsunami warning centers can determine basic earthquake source parameters rapidly. However, the source characterization at the warning centers has rested largely on scalar measures of earthquake magnitude and slowness. The reasons for this are historical and practical. The warning centers have not always received the quantity of seismic data they do now, and in the interest of speed, the calculation of scalar measures can be accomplished with the data at hand in a small amount of time. One issue the PTWC faced during the 2004 Sumatra earthquake was that no near real-time magnitude method existed at the time that would correctly estimate the size of the Sumatra earthquake. Since then, new techniques have been developed to determine the magnitude of great earthquakes. Among these are techniques based on P-wave broadband signals [12,13,15,22,35,36,59,60] and the W-phase [21,52,55,58].

Hara [35,36] and Lomax et al. [59], both use techniques that involve estimating the source duration from the P-wave coda. This estimate is obtained by applying a high pass filter to the velocity seismogram, squaring the result and smoothing it. This procedure results in a relatively smooth curve or envelope function that tracks the variation in the velocity-squared time-series. The source duration estimate is then obtained by measuring the time from the beginning of the P-wave to the point when the envelope function falls below a certain percentage of its maximum value. However, these studies also show that the use of source duration alone is not a completely satisfactory basis for a moment magnitude estimate.

Lomax et al. [59] also measures the radiated energy of the P-wave in the interval between the P- and S-arrivals,

and uses both the radiated energy and source duration estimate to formulate a magnitude scale based on the relation  $M_0 \approx E^{1/2} \cdot T^{3/2}$  [76] where  $E$  is the high frequency radiated energy and  $T$  is the source duration estimate. Hara [35] uses the estimated source duration and the maximum displacement measured in the interval of the estimated source duration from a number of earthquakes to construct an empirical formula for the magnitude. Hara [36] showed that this technique also works well for tsunami earthquakes.

Lomax et al. [60] has derived a duration-amplitude procedure for determination of a moment magnitude,  $M_{wpd}$ , for large earthquakes within 20 min of the event origin time using teleseismic P-wave recordings. Their procedure determines apparent source durations,  $T_0$ , from high frequency, P-wave records, and estimates seismic moments via integration of broadband seismograms over the interval  $t_p$  to  $t_p + T_0$ , where  $t_p$  is the P-wave arrival time. The characteristics of this method make it an extension of  $M_{wp}$ .

De Kool et al. [22] present a variation of the  $M_{wp}$  method which estimates the asymptotic behavior of the integrated displacement seismogram caused by the P-wave arrivals. Their results for  $M_{wp}$  show less scatter than do the PTWC's  $M_{wp}$  values, described above. In addition, they have automated their method.

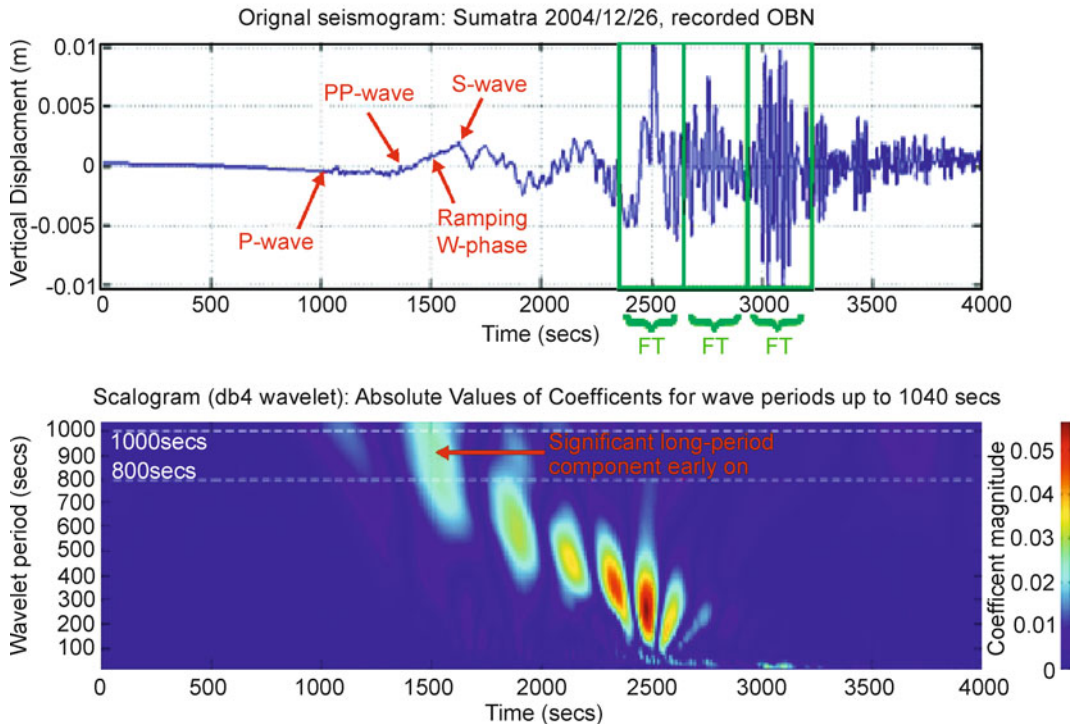
The method of Bormann and Wylegalla [13] and Bormann and Saul [14] calculate what they refer to as a cumulative body wave magnitude. They do this by summing up all of the peak velocity amplitudes for all pulses, which represent the rupturing of sub-faults, over the P-wave coda. For the December 2004 Sumatra earthquake they obtained an  $M_W$  of 9.3 in agreement with the estimate of Stein and Okal [71].

Since these methods are based on analyzing the P-waves in the P-S interval, they provide accurate moment magnitude estimates for great earthquakes within 20–25 min of the earthquake origin time. Therefore these estimates will come before estimates obtained by other methods like  $M_m$ .

The W-phase is a long period, up to 1000 s, wave that arrives before the S-wave (Fig. 12). It can be interpreted as a superposition of the fundamental, 1st, 2nd, and 3rd overtones of spheroidal modes or Rayleigh waves and has a group velocity of 8 km/s at 1000 s period, and 8.6 km/s at 100 s period [51,54]. Kanamori [54] has devised a magnitude scale based on the amplitude of the W-phase observed on deconvolved displacement records.

In addition to size and source duration, the warning centers are interested in more detailed properties of the source than can be obtained from the scalar measures we

## Wavelet Scalogram of the Sumatra-Andaman Earthquake



Earthquake Source Parameters, Rapid Estimates for Tsunami Warning, Figure 12

Top: (Fig. 2a and b from [58]). Displacement seismogram of the 2004 Sumatra-Andaman earthquake recorded at OBN. Bottom: Scalogram of top seismogram. A diagram which displays the wavelet scale as a function of time is called a "scalogram". Bottom figure shows the scalogram for the 2004 event. Color intensity at any point in the picture corresponds to the coefficient magnitude of a wavelet with a particular period at a particular point of the time series. The y-axis has been translated from wavelet scale into corresponding wavelet time period. The long-period component arrived at about 1500 s. The wavelet transform can simultaneously achieve: (1) Accurate frequency representation for low frequencies, and (2) Good time resolution for high frequencies [58]

have just discussed, such as direction of rupture and the distribution of slip along the fault. This information is important as it can be used by tsunami wave-height forecast models to better their predictions.

In the near future, the tsunami warning centers will incorporate the results of centroid moment tensors and finite fault modeling. Finite Fault modeling involves the inversion of seismic waveforms to recover more detailed information about the source process including the slip distribution, rupture propagation speed and moment release history [7,27,37,38,62,77]. Weinstein and Lundgren [80] explored the potential of a simple teleseismic P-wave inverse method for the rupture history of an earthquake for use in a tsunami warning center context. The calculations proceed quickly enough that a slip distribution may be available just a few minutes after a suitable set of P-waveforms are obtained. Hence finite fault modeling results can be used in tsunami wave height forecast

models to provide a timely initial estimate of tsunami wave heights.

The warning centers are also actively investigating the use of seismic arrays. Seismic arrays can be used to determine the direction along which the P-waves have propagated to the array. As the rupture propagates, this direction will change. By analyzing the seismic array data, this change in direction can be measured (by computing back azimuths) and the history of energy/moment release as well as the extent and direction of the rupture propagations can be determined [44,56].

### Acknowledgments

The authors greatly appreciate very thorough reviews by Kenji Satake, Anthony Lomax, Peter Bormann and our colleague Victor Sardina. Their comments greatly improved this manuscript. We also thank Paula Dunbar of

the National Geophysical Data Center for helping us obtain the data we needed and Nathan Becker's GMT artistry for Fig. 1. We also thank the PTWC for use of its facilities in preparing this manuscript.

## Bibliography

### Primary Literature

- Aki K (1966) Generation and propagation of G waves from the Niigata earthquake of June 16, 1964, Part 2. Estimation of earthquake moment, from the G wave spectrum. *Bull Earthquake Res Inst Tokyo Univ* 44:73–88
- Aki K (1967) Scaling law of seismic spectrum. *J Geophys Res* 72:1217–1231
- Allen RV (1978) Automatic earthquake recognition and timing from single traces. *Bull Seism Soc Am* 68:1521–1532
- Allen RV (1982) Automatic phase pickers: their present use and future prospects. *Bull Seism Soc Am* 72:225–242
- Ammon CJ, Ji C, Thio HK, Robinson D, Sidao N, Hjorleifsdottir V, Kanamori H, Lay T, Das S, Helmberger D, Ichinose G, Polet J, Wald D (2005) Rupture process of the 2004 Sumatra–Andaman Earthquake. *Science* 6:113–1139
- Bilek SL, Lay T (1999) Rigidity variations with depth along interplate megathrust faults in subduction zones. *Nature* 400:443–446
- Bilek SL, Ruff LJ (2002) Analysis of the 23 June 2001  $M_W = 8.4$  Peru underthrusting earthquake and its aftershocks. *Geophys Res Lett* 8:21–1–21–4
- Bilek SL, Lay T, Ruff LJ (2004) Radiated seismic energy and earthquake source duration variations from teleseismic source time functions for shallow subduction zone thrust earthquakes. *J Geophys Res* 109:B09308
- Boatwright J, Choy GL (1986) Teleseismic estimates of the energy radiated by shallow earthquakes. *J Geophys Res* 91:2095–2112
- Boatwright J, Choy GL, Seekins LC (2002) Regional estimates of radiated seismic energy. *Bull Seism Soc Am* 92:1241–1255
- Bormann P, Baumbach M, Bock G, Grosse H, Choy GL, Boatwright J (2002) Seismic sources and source parameters. In: Bormann P (ed) *IASPEI New Manual Seismological Observatory Practice*, vol 1, Chapter 3. GeoForschungsZentrum Potsdam, pp 1–94. (This can also go as in the text cited review paper)
- Bormann P, Wylegalla K (2005) Quick estimator of the size of great earthquakes. *Eos Trans AGU* 86(46):464
- Bormann P, Wylegalla K, Saul J (2006) Broadband body-wave magnitudes  $m_B$  and  $m_{BC}$  for quick reliable estimation of the size of great earthquakes. USGS Tsunami Sources Workshop 2006, poster, [http://spring.msi.umn.edu/USGS/Posters/Bormann\\_et\\_al\\_poster.pdf](http://spring.msi.umn.edu/USGS/Posters/Bormann_et_al_poster.pdf)
- Bormann P, Saul J (2008) Earthquake magnitude. In: Meyers A (ed) *Encyclopedia of complexity and systems science*. Springer, Heidelberg
- Bormann P, Saul J (2008) The new IASPEI standard broadband magnitude  $m_B$ . *Seism Res Lett* 79:698–705
- Brune J (1970) Tectonic stress and seismic shear waves from earthquakes. *J Geophys Res* 75:4997–5009
- Brune J (1971) Tectonic stress and seismic shear waves from earthquakes; Correction. *J Geophys Res* 76:5002
- Bryant E (2001) Distribution and fatalities. In: Bryant E (ed) *Tsunami: The underrated hazard*. School of Science and the Environment, Coventry University, Coventry. Cambridge University Press, Cambridge, pp 15–24
- Centroid Moment Tensor (2008) Catalog. <http://www.globalcmt.org>, accessed June 2008
- Choy GL, Boatwright J (2007) The energy radiated by the 26 December 2004 Sumatra–Andaman earthquake estimated from 10-minute P-wave windows. *Bull Seism Soc Am* 97: S18–S24
- Cummins PR (1997) Earthquake near field and W phase observations at teleseismic distances. *Geophys Res Lett* 24:2857–2860
- De Kool M, Jepsen D, Purss, Matthew (2007) Rapid moment estimation of large earthquakes using a variation of the Mwp method. In press
- Evans JR, Allen S (1983) A teleseism-specific detection algorithm for single short-period traces. *Bull Seism Soc Am* 73:1173–1186
- Fryer G, Hirshorn B, McCreery S, Cessaro RK, Weinstein S (2005) Tsunami warning in the near field: The approach in Hawaii. *EOS* 86:S44B-04
- Fukao Y (1979) Tsunami earthquakes and subduction processes near deep-sea trenches. *J Geophys Res* 84:2303–2314
- Geller RJ, Kanamori H (1977) Magnitudes of great shallow earthquakes from 1904 to 1952. *Bull Seismol Soc Am* 67: 587–598
- Giovanni MK, Beck SL, Wagner L (2002) The June 23, 2001 Peru earthquake and the southern Peru subduction zone. *Geophys Res Lett* 29:14-1–14-4
- Gower J (2005) Jason 1 detects the 26 December 2004 tsunami. *Eos Trans AGU* 86(4):37–38
- Gutenberg B (1945) Amplitudes of surface waves and magnitudes of shallow earthquakes. *Bull Seism Soc Am* 35:3–12
- Gutenberg B (1945) Amplitudes of P, PP, and S, and magnitudes of shallow earthquakes. *Bull Seism Soc Am* 35:57–69
- Gutenberg B (1945) Magnitude determinations of deep-focus earthquakes. *Bull Seism Soc Am* 35:117–130
- Gutenberg B, Richter CF (1956) Earthquake magnitude, intensity, energy and acceleration. *Bull Seism Soc Am* 46: 105–145
- Gutenberg B, Richter CF (1956) Magnitude and energy of earthquakes. *Annali di Geofisica* 9:1–15
- Hanks T, Kanamori H (1979) A moment magnitude scale. *J Geophys Res* 84:2348–2350
- Hara T (2007) Measurement of the duration of high-frequency radiation and its application to determination of the magnitudes of large shallow earthquakes. *Earth Planets Space* 59:227–231
- Hara T (2007) Magnitude determination using duration of high frequency energy radiation and displacement amplitude: application to tsunami earthquakes. *Earth Planets Space* 59: 561–565
- Hartzell S, Heaton T (1986) Rupture history of the 1984 Morgan Hill, California, Earthquake from the inversion of strong motion records. *Bull Seism Soc Am* 76:649–674
- Hartzell S, Mendoza C (1991) Application of an iterative least-squares waveform inversion of strong-motion and teleseismic records to the 1978 Tabas, Iran earthquake. *Bull Seism Soc Am* 81:1:305–331

39. Hirshorn B, Lindh A, Allen R (1987) Real Time Signal Duration Magnitudes from Low-gain Short Period Seismometers. USGS OFR 87:630
40. Hirshorn B, Lindh G, Allen RV, Johnson C (1993) Real time magnitude estimation for a prototype early warning system (EWS) from the P-wave, and for earthquake hazards monitoring from the coda envelope. *Seis Res Lett* 64:48
41. Hirshorn B (2004) Moment magnitudes from the initial P-wave for local tsunami warnings. *Seism Res Lett* 74:272–273
42. Hirshorn B (2007) The Pacific Tsunami Warning Center Response to the Mw6.7 Kiholo Bay Earthquake and Lessons for the Future. *Seism Res Lett* 78:299
43. Houston H, Kanamori H (1986) Source spectra of great earthquakes, teleseismic constraints on rupture process and strong motion. *Bull Seism Soc Am* 76:19–42
44. Ishii M, Shearer PM, Houston H, Vidale JE (2005) Extent, duration and speed of the 2004 Sumatra–Andaman earthquake imaged by the Hi-Net array. *Nature* 435:933–936
45. Johnson CE, Lindh A, Hirshorn B (1994) Robust regional phase association. *US Geol Surv Open-File Rept* 94:621
46. Johnson CE, Bittenbinder A, Bogaert B, Dietz L, Kohler W (1995) Earthworm: a flexible approach to seismic network processing. *IRIS Newslett XIV* 2:1–4
47. Kanamori H (1972) Mechanism of tsunami earthquakes. *Phys Earth Planet Inter* 6:246–259
48. Kanamori H (1977) The energy release in great earthquakes. *J Geophys Res* 82:2981–2987
49. Kanamori H (1978) Quantification of earthquakes. *Nature* 271:411–414
50. Kanamori H (1983) Magnitude scale and quantification of earthquakes. *Tectonophysics* 93:185–199
51. Kanamori H (1993) W Phase. *Geophys Res Lett* 20:1691–1694
52. Kanamori H, Kikuchi M (1993) The 1992 Nicaragua earthquake: a slow tsunami earthquake associated with subducted sediment. *Nature* 361:714–715
53. Kanamori H (2006) Seismological Aspects of the December 2004 great Sumatra Andaman Earthquake. *Earthquake Spectra* 22:S1–S12
54. Kanamori H, Rivera L (2008) Source inversion of W phase – Speeding up tsunami warning. *Geophys J Int* 175:222–238
55. Kikuchi M, Yamanaka Y (2001) EIC Seismological Note Number 105, [www.eic.eri-u-tokyo.ac.jp/EIC/EIC\\_news/105E.html](http://www.eic.eri-u-tokyo.ac.jp/EIC/EIC_news/105E.html)
56. Krüger F, Ohrnberger M (2005) Tracking the rupture of the Mw9.3 Sumatra earthquake over 1150 km at teleseismic distance. *Nature* 435:937–939
57. Lay T, Kanamori H, Ammon C et al (2005) The Great Sumatra – Andaman Earthquake of 26 December 2004. *Science* 308:1127–1133
58. Lockwood OG, Kanamori H (2006) Wavelet analysis of the seismograms of the 2004 Sumatra–Andaman earthquake and its application to tsunami early warning. *Geochem Geophys Geosyst* 7, Q09013, doi:10.1029/2006GC001272
59. Lomax A, Michelini A, Piatanesi A (2007) An energy-duration procedure for rapid determination of earthquake magnitude and tsunamigenic potential. *Geophys J Int* 170:1195–1209
60. Lomax A, Michelini A (2008) Mwpd: A duration-amplitude procedure for rapid determination of earthquake magnitude and tsunamigenic potential from P waveforms. *Geophys J Int* (accepted, in press)
61. Mendoza C (1996) Rapid derivation of rupture history for large earthquakes. *Seismol Res Lett* 67:19–26
62. Newman AV, Okal EA (1998) Teleseismic estimates of radiated seismic energy: The E/M0 discriminant for tsunami earthquakes. *J Geophys Res* 103:26885–26898
63. Okal EA, Talandier J (1989) Mm: a variable-period mantle magnitude. *J Geophys Res* 94:4169–4193
64. Okal EA (1992) Use of mantle magnitude  $M_m$  for reassessment of the moment of historical earthquakes-I: Shallow events. *PA-GEOPH* 139:17–57
65. Park J, Song TA, Tromp J, Okal E, Stein S, Roult G, Clevede E, Laske G, Kanamori H, Davis P, Berger J, Braitenberg C, Camp MV, Xiang'e L, Heping S, Houze X, Rosat S (2005) Earth's free oscillations excited by the 26 December 2004 Sumatra–Andaman earthquake. *Science* 20:1139–1144
66. Richter CF (1935) An instrumental earthquake magnitude scale. *Bull Seism Soc Am* 25:1–32
67. Satake K (1994) Mechanism of the 1992 Nicaragua tsunami earthquake. *Geophys Res Lett* 21:2519–2522
68. Savage JC (1972) Relation of corner frequency to fault dimensions. *J Geophys Res* 77:3788–3795
69. Scholz C (1982) Scaling laws for large earthquakes: Consequences for physical models. *Bull Seism Soc Am* 72:1–14
70. Seno T, Hirata K (2007) Did the 2004 Sumatra–Andaman earthquake involve a component of tsunami earthquakes? *Bull Seism Soc Am* 97:S296–S306
71. Stein S, Okal EA (2007) Ultralong period seismic study of the December 2004 Indian Ocean earthquake and implications for regional tectonics and the subduction process. *Bull Seism Soc Am* 97:279–295
72. Indian Ocean Earthquake and implications for regional tectonics and the subduction process. *Bull Seism Soc Am* 97: S279–S295
73. Tsuboi SK, Abe K, Takano K, Yamanaka Y (1995) Rapid determination of Mw from broadband P waveforms. *Bull Seism Soc Am* 83:606–613
74. Tsuboi S, Whitmore PM, Sokolowski TJ (1999) Application of Mwp to deep and teleseismic earthquakes. *Bull Seism Soc Am* 89:1345–1351
75. Vanek J, Zatopek A, Karnik V, Kondorskaya N, Riznichenko Y, Savarenski S, Solovev S, Shebalin N (1962) Standardization of magnitude scales. *Izv Acad Sci USSR Geophys Ser*, pp 108–111 (English translation)
76. Vassiliou MS, Kanamori H (1982) The energy release in earthquakes. *Bull Seism Soc Am* 72:371–387
77. Wald DJ, Helmlinger DV, Hartzell S (1990) Rupture process of the 1987 Superstition Hills earthquake from the inversion of strong-motion data. *Bull Seism Soc Am* 80:1079–1098
78. Weinstein SA, McCreery C, Hirshorn B, Whitmore P (2005) Comment on “A strategy to rapidly determine the magnitude of great earthquakes” by Menke W, Levin V. *Eos* 86:263
79. Weinstein SA, Okal EA (2005) The mantle magnitude  $M_m$  and the slowness parameter  $\theta$ : Five years of real-time use in the context of tsunami warning. *Bull Seism Soc Am* 95:779–799
80. Weinstein SA, Lundgren PL (2008) Finite fault modeling in a tsunami warning center context. In: Tiampo KF, Weatherley DK, Weinstein SA (eds) *Earthquakes: Simulations, sources and tsunamis*. Birkhauser, Basel
81. Whitmore PM, Sokolowski TJ (2002) Automatic earthquake processing developments at the US West Coast/Alaska

- tsunami warning center. Recent Research Developments in Seismology, 1–13. Kervala, India: Transworld Research Network. ISBN 81-7895, 072-3
82. Whitmore PM, Tsuboi S, Hirshorn B, Sokolowski TJ (2002) Magnitude-dependent correction for  $M_{wp}$ . *Sci Tsunami Hazards* J 20:187–192
  83. Widjo K et al (2006) Rapid survey on tsunami Java 17 July, 2006. [http://nctr.pmel.noaa.gov/java20060717/tsunami-java170706\\_e.pdf](http://nctr.pmel.noaa.gov/java20060717/tsunami-java170706_e.pdf). Accessed July 2008
  84. Withers M, Aster R, Young C, Beiriger J, Harris M, Trujillo J (1998) A comparison of select trigger algorithms for automated global seismic phase and event detection. *Bull Seism Soc Am* 88:95–106
- Books and Reviews**
- Abercrombie R, McGarr A, Di Toro G, Kanamori H (eds) Earthquakes: Radiated energy and the physics of faulting. In: *Geophysical Monographs* 170. American Geophysical Union, Washington DC
- Aki K, Richards PG (1980) *Quantitative Seismology Theory and Methods*, 2 vol. W.H. Freeman Co, San Francisco
- Aki K, Richards PG (2002) *Quantitative Seismology*, 2nd edn. University Science Books, Sausalito
- Båth M (1981) Earthquake magnitude – recent research and current trends. *Earth Sci Rev* 17:315–398
- Bormann P (ed)(2002) *IASPEI new manual of seismological observatory practice*, vol 1 and 2. GeoForschungsZentrum Potsdam, p 1250
- Duda S, Aki K (eds) (1983) Quantification of earthquakes. *Tectonophysics* 93, Special issue 3/4:183–356
- Kanamori H, Anderson DL (1975) Theoretical basis of some empirical relations in seismology. *Bull Seismol Soc Am* 65(5): 1073–1095
- Kanamori H (1994) Mechanics of Earthquakes. *Annu Rev Earth Planet Sci* 22:307–237
- Kanamori H, Rivera L (2006) Energy partitioning during an earthquake. In: Abercrombie R, McGarr A, Kanamori H, Di Toro G (eds) *Earthquakes: Radiated energy and the physics of faulting*. *Geophysical Monograph* 170, American Geophysical Union, Washington, DC, pp 3–13
- Kanamori H, Brodsky E (2004) The Physics of earthquakes. *Rep Prog Phys* 67:1429–1496
- Kanamori H, The diversity of the physics of earthquakes. *Proc Japan Acad Ser B* 80:297–316
- Lay T, Wallace TC (1995) *Modern global seismology*. Academic Press
- Okal EA (1992) A student's guide to teleseismic body wave amplitudes. *Seism Res Lett* 63(N2)169–180
- Richter CF (1958) *Elementary seismology*. WH Freeman Co, San Francisco
- Stein S, Wysession M (2003) *An introduction to seismology, earthquakes, and earth structure*. Blackwell Publishing

## Earthquakes, Dynamic Triggering of

STEPHANIE G. PREJEAN<sup>1</sup>, DAVID P. HILL<sup>2</sup>

<sup>1</sup> US Geological Survey, Alaska Science Center,  
Anchorage, USA

<sup>2</sup> US Geological Survey, Volcano Hazards Program,  
Menlo Park, USA

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Review of Dynamic Triggering Observations](#)

[Characteristics of Dynamic Triggering](#)

[Physical Models of Dynamic Triggering](#)

[Future Directions](#)

[Bibliography](#)

### Glossary

**Dynamic stress change** a transient, often oscillatory change in the Earth's stress field.

**Static stress change** a permanent, step-like change in the Earth's stress field.

### Definition of the Subject

Geoscientists have long sought understanding of how earthquakes interact. Can earthquakes trigger other earthquakes? The answer is clearly yes over short time and distance scales, as in the case of mainshock – aftershock sequences. Over increasing time and distance scales however, this question becomes more difficult to answer. The study of dynamically triggered earthquakes explores the most distant boundaries over which earthquakes trigger other earthquakes.

Dynamic triggering to temporary and oscillatory fluctuations in the stress/strain regime in a volume of the Earth's crust. Dynamic stress fluctuations are associated with ground shaking resulting from either anthropogenic activities or natural sources. Dynamic triggering occurs as seismic waves from an initial earthquake propagate through the Earth's crust, triggering secondary earthquakes. Once the seismic wave train has passed and ground shaking ends in a given locale, the crust returns to its previous stress state modified by the combined stress drops associated with any locally triggered earthquakes.

Dynamic triggering includes wide ranging phenomenon, both geographically and in its characteristics. It has been observed across the globe in a variety of geologic

and tectonic environments. It has been shown to occur at distances from the initial earthquake rupture varying from meters [21,27,52] to over 11,000 km [103]. In the most distant cases, earthquake triggering results from dynamic stress perturbations as low as 0.01 MPa. Earthquakes have also been shown to trigger other earthquakes at a variety of time scales. In many cases triggering of earthquakes occurs during or within minutes to hours following the responsible seismic waves (e. g. [26,39,74,103]). In other cases, earthquakes occurring weeks to months after the initial earthquake have been interpreted as a delayed response to dynamic triggering (e. g. [41,89,102]). Delayed triggered responses may reflect a more complex series of physical processes. For example, dynamic waves may trigger an aseismic process such as fault creep or changes in a volcanic system, which subsequently triggers earthquakes secondarily [2,4,37,50].

This field has been an area of extensive research in the past twenty five years. It offers a potentially important key to improving our understanding of earthquake nucleation in that, in principle, we can determine in-situ perturbations in the local stress field that lead to earthquake nucleation and rupture. In particular, the availability of broadband seismic data near sites of triggered seismicity allows us to calculate the time history of stress field fluctuations responsible for earthquake nucleation given adequate knowledge of the local seismic velocity structure [28,38,103].

The study of dynamically triggered earthquakes can also help better characterize the physical condition of the Earth's crust at seismogenic depths. Many researchers were surprised that earthquakes could be triggered by stress perturbations as small as 0.01 MPa. This observation indicates the Earth's crust is on the verge of failure in areas with triggered responses to distant earthquakes. This field of research may also provide clues to the hydrologic regime at depth. It has long been recognized that water tables change in response to earthquakes thousands of km distant [17]. The link between dynamically triggered earthquakes and dynamically triggered hydrological changes is an active area of research [8,19,78].

Within the context of complexity and system science, [82] suggest that remotely triggered seismicity may reflect large activation correlation lengths (ACL) in fault systems and stress fields that have reached a state of self-organized criticality. This statistical physics approach to earthquake occurrence focuses on the exploration of both analog and computational models that can mimic observed dynamical space-time patterns spanning a wide range of spatial-temporal scales. It is not concerned with inferred (or “non-observable”) physical models for the lo-



cal processes linking dynamic stresses and brittle failure (triggered earthquakes) on faults (see [82]). In this review, however, we focus on these physical models together with a description of documented patterns of remote dynamic triggering.

## Introduction

### Introduction to Stress Triggering of Earthquakes

Earthquake triggering refers to a process by which any change in fault properties or the processes acting on a fault leads to rupture initiation. More specifically, stress triggering occurs when a change in the stress field acting on a fault leads to rupture. Stress triggering of earthquakes can result from stresses applied over a variety of time scales and with a variety of frequencies, which generally fall into three partially overlapping categories, 1) static stress triggering, 2) quasi-static stress triggering, and 3) dynamic stress triggering. In the case of static stress changes, the state of stress acting across a fault is permanently perturbed. This form of stress triggering is important in the near field of an earthquake where fault displacement significantly alters the stress field in the surrounding crust. Static stress triggering is commonly regarded as the dominant factor controlling aftershock generation (e.g. [54]). Because static stress changes decay rapidly with distance from the earthquake rupture (as  $d^{-3}$ , where  $d$  is distance from the earthquake epicenter), they are generally thought to be significant only within two to three fault lengths of the earthquake rupture. The role of static stress changes in triggering aftershocks and other earthquakes has been a vigorous and productive area of research in the past two decades (for reviews see [34,53,92,93]). The relative importance of static vs. dynamic triggering of aftershocks in the near field, however, has recently become an actively debated topic (e.g. [21,71]).

Because static stress changes in the near field develop essentially simultaneously with earthquake rupture, simple static stress triggering must appeal to other mechanisms to explain the time delay associated with many aftershocks and subsequently triggered earthquakes. In contrast, quasi-static stress triggering results from viscoelastic relaxation of the crust after an earthquake. Because quasi-static stress changes decay as  $d^{-2}$  and because viscoelastic relaxation is a time dependent process, these stress changes may explain triggered earthquakes more distant from an initial earthquake and triggered earthquakes with delay times from years to decades [72].

Dynamic stress changes decay more slowly with distance than either static stress changes or quasi-static stress changes (as  $d^{-1.5}$  for surface waves). Thus dynamic

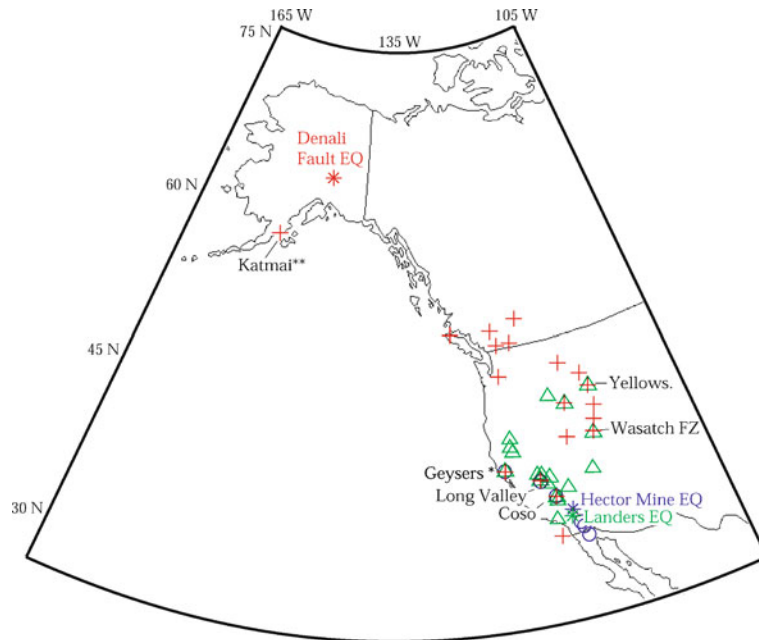
stress changes become increasingly dominant with increasing distance from the fault rupture. In this review we discuss dynamic triggering of earthquakes resulting from ground shaking due to the passing seismic wavetrain of other earthquakes. We focus on dynamic triggering due to remote earthquakes (greater than two fault lengths distance from the earthquake rupture), though we will briefly discuss the active research topic of dynamic triggering in an earthquake's aftershock zone as well. We also limit discussion to frequencies of ground shaking above  $\sim 0.01$  Hz (periods less than  $\sim 100$  s), though some work has been done on earthquakes triggered by longer wavelength fluctuations. For example, [16,95] recently found evidence that solid Earth tides can modulate background seismicity rates. Other reviews of dynamic stress triggering can be found in [23,37,92] and [38].

### Brief History of Dynamic Stress Triggering Research

The ability of earthquakes to trigger other earthquakes at great distances has been discussed in scientific literature throughout the latter half of the 20th century (see [38] for review). However, making a credible case for a causal link between two earthquakes remains a major challenge in this field. Beyond the realm of aftershock zones, it was difficult to justify statistically that one earthquake triggered another until the 1980s and 1990s. By then continuously recording telemetered seismic networks and automated processing of data became commonplace, providing reliable spatial and temporal records of earthquake occurrence at  $M \geq 1 - 2$  and the statistical leverage associated with large numbers of small earthquakes.

Dynamic triggering of earthquakes was widely accepted in the scientific community following the 1992 M 7.3 Landers earthquake in southern California. In the minutes to days following the Landers earthquake, earthquake rates increased dramatically across the western United States at distances well beyond the aftershock zone [39]. Earthquakes were triggered throughout California, Nevada, Utah, Wyoming, and Idaho at distances of up to 1250 km (Fig. 1). Although time delays of triggered events ranged from seconds to 33 hours after the arrival of the Landers earthquake wavetrain, the sudden increase in seismicity across the Western United States could not be ignored. This earthquake spawned a plethora of studies into the nature of earthquake triggering and remains one of the best studied triggering episodes to date.

The geophysical community had a unique research opportunity when the 1999 M 7.1 Hector Mine earthquake occurred. Because it was an earthquake with similar magnitude to the Landers earthquake occurring in a similar



Earthquakes, Dynamic Triggering of, Figure 1

Map showing sites of triggered seismicity in western North America from the Landers (green triangles), Hector Mine (blue circles), and Denali Fault (red crosses) earthquakes. Modified from [38], *Treatise on Geophysics*

location, it provided leverage to tune ideas about dynamically triggered seismicity and the underlying physical processes. Although the Hector Mine earthquake triggered seismicity at some of the same locations as the Landers earthquake, the Hector Mine earthquake had a much more limited triggered response (Fig. 1) [30]. The difference is likely due in part to differences in seismic radiation patterns between the two earthquakes [30,43]. The Landers earthquake ruptured unilaterally to the north, whereas the Hector Mine earthquake ruptured bilaterally, primarily to the south.

Following the Landers and Hector Mine earthquakes, the search for dynamic triggering began in earnest. Researchers across the globe began scanning earthquake catalogs and waveform data searching for dynamic triggering in a wide variety of environments (see “Sect. [Review of Dynamic Triggering Observations](#)”). At the Geysers, CA alone [91] identified 7 episodes of dynamic triggering between 1988 and 1994, making this among the most frequently triggered locations known.

Like the Landers earthquake, the 2002 M 7.9 Denali Fault earthquake triggered a widespread response across western Canada and the United States (Fig. 1) [27,33,45,64,68,74]. The increase in the number of broadband and high-dynamic range seismometers by 2002 allowed scientists to visually scan on-scale seismic data during the earth-

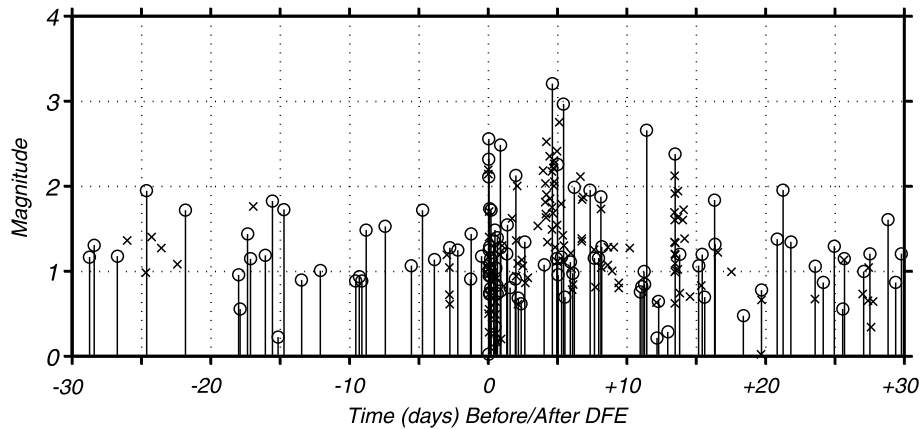
quake’s wavetrain. Thus, many triggered events were detected which were absent from earthquake catalogs. As an example of how instrumentation improvements increase our ability to detect dynamic triggering, [45] point out that the triggered response of the Yellowstone caldera to the Denali Fault earthquake could not have been detected at the time of the Hector Mine earthquake only three years earlier.

## Review of Dynamic Triggering Observations

### Detection of Dynamic Triggering

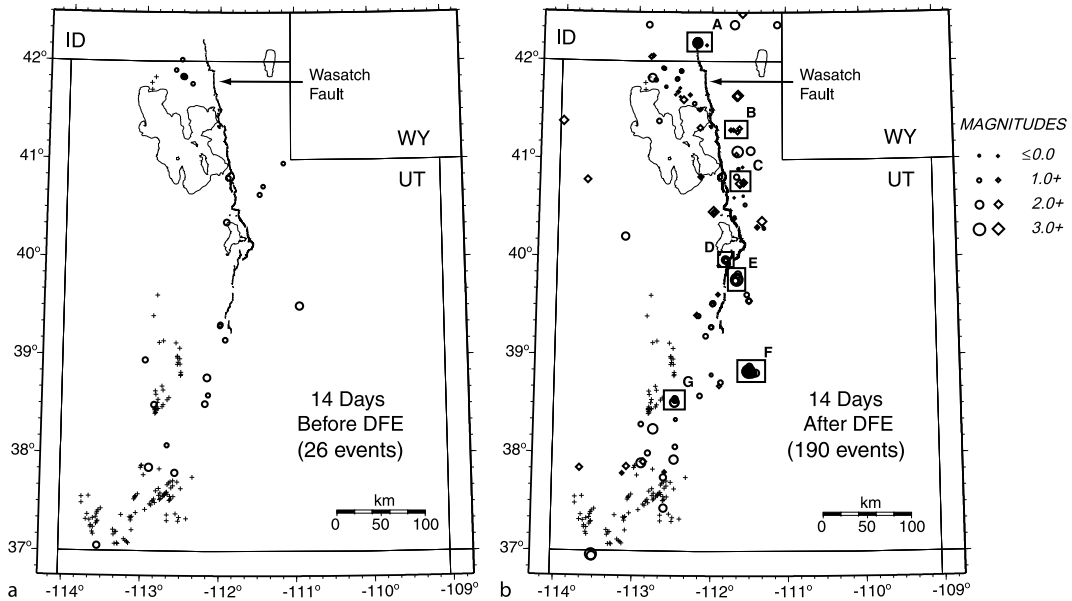
Dynamic triggering has been observed in a variety of locations around the globe. Some suggest that dynamic triggering of earthquakes is a ubiquitous process in the Earth’s crust (e.g. [26,41]). Others suggest that some areas are more likely to experience dynamic triggering of earthquakes than others (e.g. [39,64]). Observations of dynamic triggering are limited geographically due to uneven seismic network coverage and the effort applied to examining seismic data.

Following the 1992 M 7.3 Landers earthquake, dynamic triggering was recognized by the sudden increase in the number of earthquakes located through standard network processing across the western US in the days to weeks after the large earthquake. Searching earthquake



Earthquakes, Dynamic Triggering of, Figure 2

Plot of earthquake magnitude versus time in the area of the Wasatch Front, Utah, 30 days before and after the Denali Fault earthquake. Circles represent independent events. Crosses indicate secondary events determined by declustering the earthquake catalog. Figure reprinted from [68], BSSA

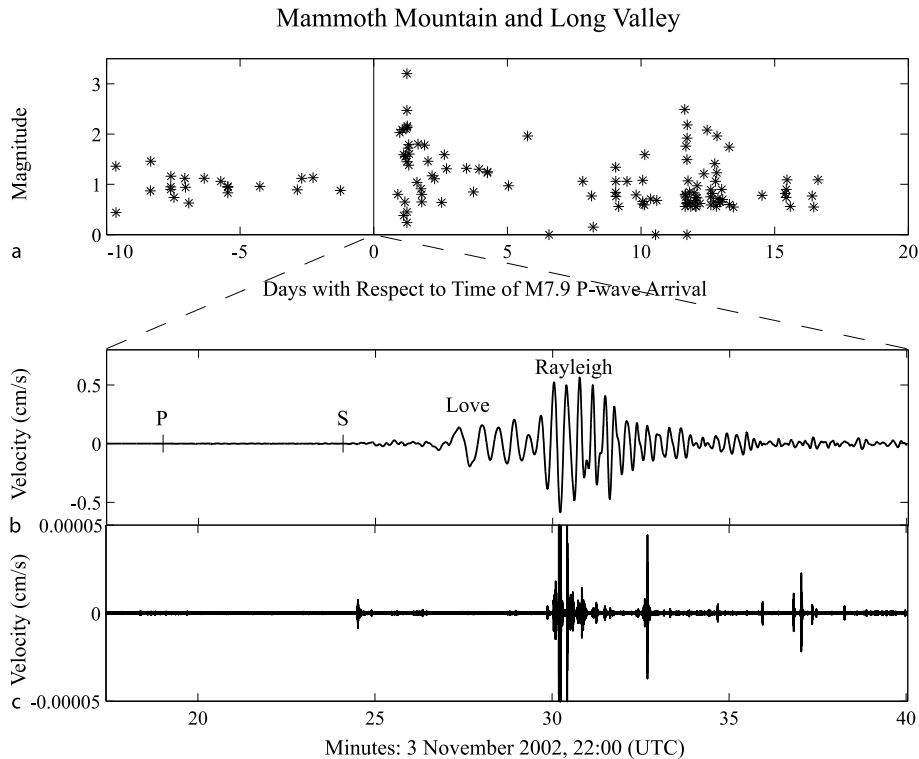


Earthquakes, Dynamic Triggering of, Figure 3

Seismicity in Utah 14 days before and after the Denali Fault earthquake. Diamonds in b show earthquakes occurring in the first 24 hours after the arrival of the wave train from the Denali Fault earthquake. Cross are locations of quaternary volcanic vents [3]. Figure reprinted from [68], BSSA

catalogs for sudden increases in seismicity after a large earthquake is one common method of identifying dynamic triggering (Figs. 2 and 3) (e.g. [6,30,39,41,68]). Identifying triggered seismicity using earthquake catalogs simplifies interpretation of triggered response with respect to background seismicity rates, as earthquake catalogs often provide stable long-term records of earthquake occurrence at a consistent threshold.

A second commonly used method of detecting dynamic triggering involves visually scanning continuous seismic data shortly before and after a large earthquake to identify a sudden increase in earthquakes too small to be detected and located through standard network processing (Fig. 4). This latter method is effective at identifying triggered seismicity in sparsely instrumented areas, identifying very small triggered earthquakes, and identifying



Earthquakes, Dynamic Triggering of, Figure 4

Seismicity triggered at Mammoth Mt. **b-c** and within the Long Valley caldera, California, **a** following the Denali Fault earthquake. **a** Catalog from NCEDC showing two swarms following the Denali fault earthquake in the caldera's south moat. The two lower panels show data from this very small earthquake swarm recorded on the broadband UNR/USGS station OMM from rotated to transverse direction, showing Denali earthquake wavetrain at Long Valley. Major arrivals are labeled. **c** Record from **b** high pass-filtered, showing small local earthquakes occurring during Denali wavetrain. Modified from [74], BSSA

earthquakes that occurred during the wavetrain from the initial earthquake (e. g. [26,47,64,74,103]). This method of detecting triggering has become more common with increasing availability of continuously recorded high-dynamic-range seismic data.

Possible instances of dynamic triggering have also been proposed based on historical accounts [40,42,44,59]. Because these studies rely on felt reports, they are generally limited to moderate to large sized triggered earthquakes that are separated in time by days to months.

With any method of detecting dynamic triggering, it must be shown that one earthquake is likely causally linked to the dynamic waves radiating from a previous earthquake, rather than by coincidence. Earthquakes near each other in time are more likely to be related than earthquakes separated in time. Additionally, earthquakes unlikely to occur randomly (e. g. large events in seismically quiet areas) are more likely to be related than earthquakes occurring commonly as background seismicity (e. g. small earthquakes in a seismically active area). To calculate the

probability that two earthquakes are related, one must first calculate the probability of each occurring randomly. This is usually done using patterns of earthquake occurrence based on local earthquake catalogs. The most commonly used statistical test to identify whether an increase in number of earthquakes is statistically significant is the Beta statistic [58]. Pankow et al. [68] also employ a binomial distribution analysis to this end. These techniques have potential pitfalls however, as they rely on assumptions about earthquake distributions and compare snapshots of seismicity in time in regions where seismicity rates fluctuate regularly [58]. Objectively determining whether one earthquake is genetically related to another remains a challenge.

Because spatial-temporal clusters of earthquakes are less common than isolated events, clusters of earthquakes temporally coincident with dynamic stresses are more easily identified as being triggered than isolated earthquakes. In the case of earthquake clusters, however, it may be difficult to discriminate between earthquakes directly trig-

gered by dynamic stresses from a remote earthquake and secondary aftershocks to directly triggered events [9]. To address this, earthquake catalogs are frequently ‘declustered’ (e. g. [45,68]). This process involves decomposing a catalog into primary and secondary earthquakes based on statistical patterns of aftershock sequences (e. g. [76]). [68] and [9] show that in some cases triggered seismicity is modeled well as an aftershock sequence. In other cases, however, triggered seismicity swarms cannot be dismissed as secondary aftershock sequences (e. g. [74,103]). In such cases it is likely that most events in a swarm were triggered directly by the dynamic waves radiated from a distant earthquake or perhaps as a secondary response to some aseismic process (e. g. fluid flow, or local deformation associated with fault creep) triggered by the dynamic waves.

### Dynamic Triggering in Volcanic and Geothermal Regimes

Although dynamic triggering has been observed in a variety of environments, many of these observations are from areas with active volcanic and hydrothermal systems (Table 1) [6,73,74]. These areas typically have high background seismicity rates indicating that the crust habitually hovers near failure and thus is particularly susceptible to dynamic triggering. Furthermore, because these areas tend to be well instrumented, dynamically triggered earthquakes may be unusually easy to detect. Here we summarize observations of triggered seismicity in volcanic and hydrothermal areas.

The Geysers geothermal field in northern California is among the most frequently triggered sites known with 9 cases of dynamic triggering documented in the past 20 years [28,74,91]. Earthquakes that have caused triggering at the Geysers range in magnitude from 6.9 to 7.9 and in distance from 212 km to 3120 km. The Coso geothermal field in southern California has also experienced repeated episodes of dynamic triggering following the M 7.3 Landers, the M 7.1 Hector Mine, and the M 7.9 Denali Fault earthquakes (Fig. 1) [39,74]. Following the Hector Mine earthquake, dynamically triggered earthquakes and ground deformation were observed near a third geothermal field – Cerro Prieto, Baja California [24,30].

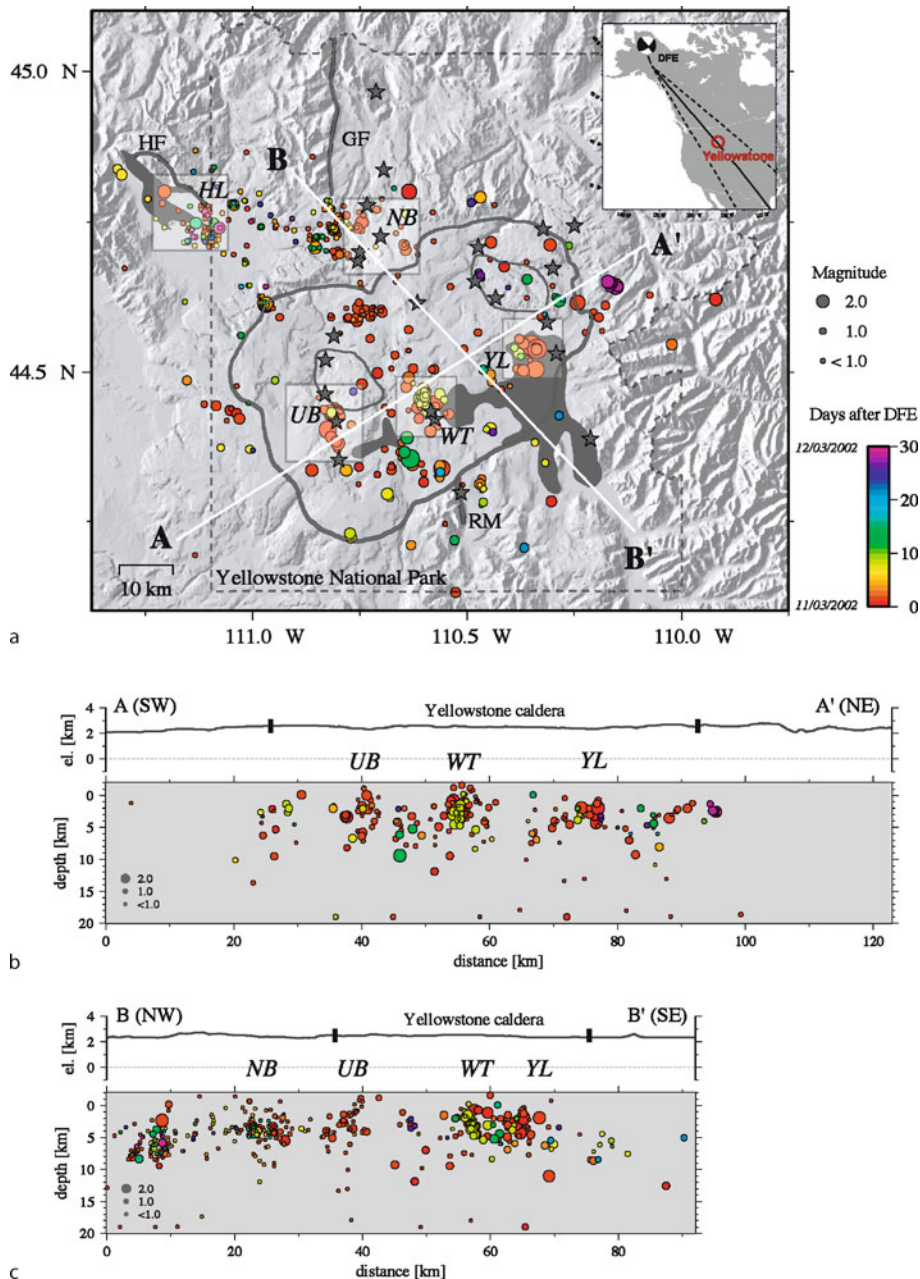
Yellowstone, Wyoming is a larger and more complicated system than the geothermal fields mentioned above, as it is a caldera system characterized by ongoing magmatic and tectonic activity, in addition to hydrothermal activity. Yellowstone had a triggered response to the M 7.3 Landers earthquake [39] and the M 7.9 Denali Fault earthquake [45,47]. The triggered response to the Denali Fault

earthquake was particularly dramatic (Figs. 5 and 6). Seismicity increased immediately following the arrival of surface waves from the Denali Fault earthquake and remained unusually high for 30 days with magnitudes ranging from  $< 0.0$  to M 3.2 [45]. The time scale of the triggered response was spatially variable (Figs. 5 and 6). Because the Denali Fault earthquake led to immediate triggering in the area of geysers and affected periodicity of geysers at Yellowstone, it is likely that changes in the hydrothermal regime induced a triggered response in some areas ([45,46]. In other areas however, the development of triggered earthquake sequences was delayed and similar to commonly observed tectonic activity [45].

Like the Yellowstone caldera, the Long Valley caldera experiences dynamic triggering with complex characteristics. The area responded to the Landers earthquake [39], the Hector Mine earthquake [50], and the Denali Fault earthquake [50,74] both seismically and geodetically, although each response varied in location and intensity (Fig. 7). The Landers earthquake produced the largest triggered response with 340 earthquakes in seven days up to M 3.4 throughout the south moat of the caldera [39]. The seismic response to the Hector Mine earthquake was comparatively short lived and limited to the region of Mammoth Mountain. After the Denali fault earthquake, the caldera area experienced two phases of triggered seismicity. A burst of  $\sim 60M \leq 0.8$  earthquakes occurred beneath Mammoth Mountain during and shortly after the arrival of the surface waves from the Denali Fault earthquake [74]. Twenty-four hours later a larger swarm of earthquakes of  $M \leq 3.4$  occurred in the Long Valley caldera’s south moat [74]. All three episodes of dynamic triggering in the Long Valley caldera were accompanied by deformation transients with geodetic moments an order of magnitude larger than the cumulative seismic moment of the triggered seismicity [36,50], though the time history and magnitude of each deformation response varied.

Iwo Jima, Japan, a volcanic island hosting a Holocene eruption, geothermal activity, and historic phreatic (steam) eruptions is a third complex caldera system which has experienced dynamic triggering of local earthquakes. [98] examined continuous waveform data of  $21M > 7$  earthquakes  $< 3000$  km distance from Iwo Jima, and identified 4 cases of resulting increased local seismicity. In all cases earthquakes were triggered locally during surface wave arrivals and persisted for 6–15 minutes.

Dynamic triggering of earthquakes has been observed at shallow depths in volcanic edifices at a variety of locales (Table 1). In the Pacific Northwest, Mt. Rainier experienced 6–8  $M < 0$  earthquakes during the wavetrain of the Denali Fault earthquake and 8  $M \leq 0.9$  earthquakes in



Earthquakes, Dynamic Triggering of, Figure 5

Seismicity within one month of the Denali Fault earthquake at Yellowstone caldera, *color* coded with time: **a** earthquake locations, **b** cross sections along AA', **c** cross section along BB'. Specific areas labeled: HL, Hebgen Lake area; NB, Norris geyser basin; UB, Upper geyser basin; WT, West Thumb geyser basin; YL, northern end of Yellowstone Lake. Large normal faults are represented with *thick black lines*: RM, Red Mountain fault zone; GF, Gallatin fault; HF, Hebgen and Red Canyon faults. *Inset* shows location of Denali Fault earthquake and Yellowstone. *Solid and dashed lines in inset* show the great circle path  $\pm 10$  degrees along the strike of the Denali Fault earthquake. Figure reprinted from [45], BSSA

Earthquakes, Dynamic Triggering of, Table 1  
Published occurrences of discrete remotely triggered earthquakes

Locations	Citation	Distance (km)	Triggering Earthquake	M	Onset	M <sub>max</sub>	Reg	Env
Aso, Japan	[61]	1400	Chi-Chi, 1999	7.7	During P waves	–	E	V
British Columbia	[26]	1800–2200	Denali Fault, 2002	7.9	During surface and coda waves	–	–	N
Burney, CA	[39]	900	Landers, 1992	7.3	23 hour	2.8	E	N
Central and Southern, CA	[41]	variable	15 Central and Southern CA earthquakes, 1988–2004	5.3–7.1	Within 1 month	–	–	–
Cerro Prieto, Mexico	[24]	260	Hector Mine, 1999	7.1	–	4.1	E	V,G
Coso, CA	[39]	165–205	Landers, 1992	7.3	~ 3 hour	4.4	E	G
Coso, CA	[74]		Hector Mine, 1999	7.1	–	–	E	G
Coso, CA	[74]	3,660	Denali Fault, 2002	7.9	15 min	2.3	E	G
Geysers, CA	[28]	2,500	Gulf of Alaska, 1988	7.6	–	0.2–2.5	E	G
Geysers, CA	[28]	212	Loma Prieta, 1989	7.1	–	0.2–2.5	E	G
Geysers, CA	[28]	443	Off Oregon Coast, 1991	6.9	–	0.2–2.5	E	G
Geysers, CA	[28]	390	Gorda Plate, CA, 1991	7.1	–	0.2–2.5	E	G
Geysers, CA	[28]	202	Petrolia, CA, 1992	7.0	–	0.2–2.5	E	G
Geysers, CA	[39]	740	Landers, 1992	7.3	3 min	1.6	E	G
Geysers, CA	[91]	635	Northridge, 1994	6.6	–	–	E	G
Geysers, CA	[28]	308	Cape Mendoceno, CA, 1994	6.9	–	0.2–2.5	E	G
Geysers, CA	[74]	3,120	Denali Fault, 2002	7.9	12 min	2.5	E	G
Greece	[6]	400–1000	Izmit, 1999	7.4	After surface waves	3.5	E	G
Iceland	[2]	64–78	South Iceland Seismic Zone, 2000	6.5	< 5 min	5	E	V, G
Idaho, Cascade	[39]	1100	Landers, 1992	7.3	33 hour	1.7	E	G
Idaho, Cascade	[26,47]	2300	Denali Fault, 2002	7.9	During Rayleigh waves	4.6	E	G
Iwo Jima, Japan	[98]	≤ 2009 km	4 earthquakes, 1983–1993	7.1–8.0	During surface waves	< 2	–	V,G
Katmai volcanoes	[73]	115	1999	7.0	< 3 min	2.3	–	V, G
Katmai volcanoes	[64]	122	2000	6.8	–	0.9	–	V, G
Katmai volcanoes	[64]	161	2001	7.0	< 2 min	1.5	–	V, G
Katmai volcanoes	[64]	161	2001	6.8	–	–	–	V, G
Katmai volcanoes	[64]	740	Denali Fault, 2002	7.9	3.9 min	2.0	–	V, G
Katmai volcanoes	This paper	3620	Kurile, 2007	8.2	During surface waves	–	–	V, G
Lassen, CA	[39]	840	Landers, 1992	7.3	12 min	2.8	E	V, G
Little Skull Mt., NV	[39]	240	Landers, 1992	7.3	1.5 hour	5.6	E	N
Long Valley, CA	[39]	415	Landers, 1992	7.3	9 min.	3.4	E	V, G
Long Valley, CA	[74]	3,454	Denali Fault, 2002	7.9	23.5 hour	3.0	E	V, G
Mammoth, CA	[50]	450	Hector Mine, 1999	7.1	20 min.	–	E	V, G
Mammoth, CA	[74]	3,454	Denali Fault, 2002	7.9	17 min.	0.8	E	V, G
Mono Basin, CA	[39]	450	Landers, 1992	7.3	19 hour	3.1	E	N
Mt. Rainier, WA	[74]	3,108	Denali Fault, 2002	7.9	12 min.	0.0	E	V
Mt. Rainier, WA	[74]	3,108	Denali Fault, 2002	7.9	2.5 hour	0.9	E	V
Nanki Trough, Japan	[60]	900–1400	Tokachi-oki, 2003	8.1	After surface waves	–	S	–

the following days. In response to the Landers earthquake, Mt. Lassen in northern California hosted 14 earthquakes of  $M \leq 2.8$ . Volcanoes in the Katmai Volcanic Clus-

ter, Alaska, have experienced triggered seismicity on at least seven occasions since 1999 ([64,65,73], this chapter). The largest of these triggered responses included 17 earth-

Earthquakes, Dynamic Triggering of, Table 1  
(continued)

Locations	Citation	Distance (km)	Triggering Earthquake	M	Onset	M <sub>max</sub>	Reg	Env
The Netherlands Roer Valley	[12]	40	Roermond, 1992	5.4	–	3.7	E	N
New Madrid, MO	[41]	1000	1811–1812 New Madrid	~ 7.8	–	–	C	N
Offshore Southern CA	[74]	4,003	Denali Fault, 2002	7.9	Mainshock coda	2.5	E	N
Salton Sea, CA	[43]	120–150	Hector Mine, 1999	7.1	–	4.7	E	V,G
Syria – Lebanon border	[62]	500	Gulf of Aqaba, 1995	7.3	2 hr 47 min	3.7	C	N
Taiwan	[102]	variable	9 earthquakes, 1978–1994	6.5–7.1	≤ 15 days	≥ 4.0	–	V,G
Tonga Trench	[96]	290–313	Tonga Region, 2002	7.6	2 min.	7.7	S	–
Utah, Cedar City	[39]	490	Landers, 1993	7.3	39 min.	4.1	E	G
Utah, Wasatch Front	[68]	3000–3500	Denali Fault, 2002	7.9	During surface waves	3.2	E	G
Valley of Mexico	[89]	303–588	7 earthquakes	7.6–8.0	–	~ 4.0	E	V,G
Western Nevada	[1,39]	450–650	Landers, 1993	7.3	9 min.	4.0	E	G
White Mts., CA	[39]	380–420	Landers, 1993	7.3	11.6 hour	3.7	E	N
Mt. Wrangell, AK	[45]	11,000	Denali Fault, 2002	7.9	During Rayleigh waves	1.0	–	V, G
Yellowstone	[39]	1250	Landers, 1993	7.3	1.8 hour	2.1	E	V, G
Yellowstone	[45]	120	Hector Mine, 1999	7.1	During surface waves	–	E	V,G
Yellowstone	[45]	3150	Denali Fault, 2002	7.9	During Love waves	3.2	E	V, G

Location is location of triggered seismicity. Distance is distance from location of triggering to triggering earthquake (mainshock) epicenter. M is magnitude of mainshock. Onset is onset time of triggered activity with respect to arrival of waves from mainshock. M<sub>max</sub> is magnitude of the largest triggered earthquake. Reg describes stress regime: E-extensional or transtensional, C-compressional or transpressional, S-subduction zone. Env describes if the area is volcanic (V), geothermally active (G), or neither (N). – indicates data not available or inconclusive.

quakes of  $M \leq 2.3$ . During the wavetrain of the 2006  $M$  8.7 Sumatra–Andaman Islands earthquake, Mt. Wrangell, Alaska had triggered 14 earthquakes [103]. With the exception of the Mt. Lassen response following the Landers earthquake and the delayed events at Mt. Rainier following the Denali Fault earthquake, these earthquakes triggered in volcanic edifices were too small to be detected and located by automatic processing systems.

The Valley of Mexico is a large volcanically and geothermally active area located in the Trans Mexican Volcanic Belt. [89] searched for dynamically triggered earthquakes in the Valley of Mexico following 18  $M \geq 7.0$  Mexican earthquakes between 1920 and 1998. In seven cases, they found evidence for dynamic triggering of earthquakes within 2 days of a large earthquake. In four additional cases, seismicity increased after a large earthquake, but was delayed by up to one month. Because this study used only one station however, the potentially triggered events can only be located to within some ill-defined region surrounding the station.

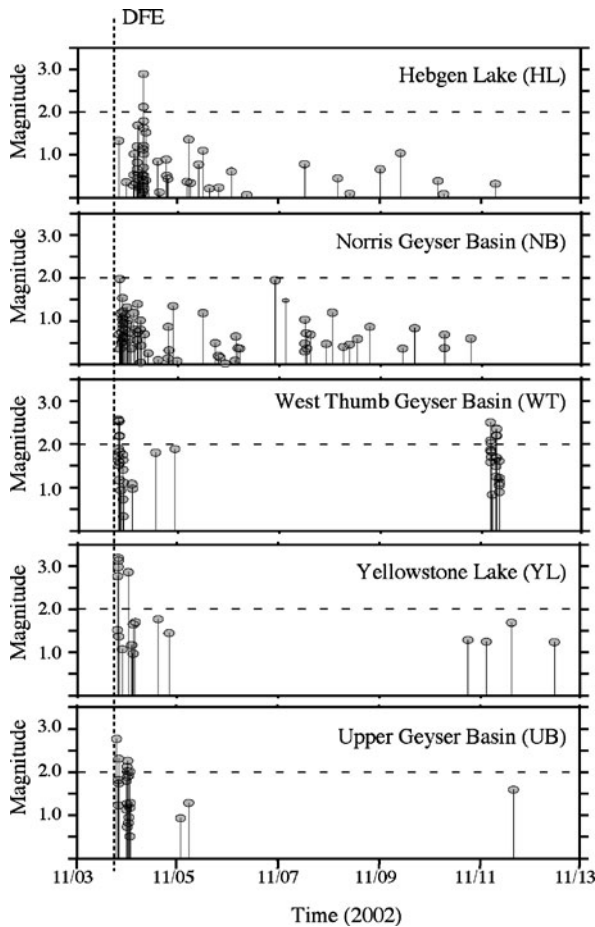
The South Iceland Seismic Zone is a transform zone in a volcanically and geothermally active area. In 2000, a  $M_w = 6.5$  earthquake in the South Iceland Seismic Zone trig-

gered widespread seismicity, including three  $M_w \sim 5.0$  earthquakes within 5 minutes of its occurrence. Coulomb failure stress calculations indicate that the two  $M > 5$  earthquakes located  $\sim 100$  km to the west on the Reykjanes Peninsula are beyond the range where static stress changes are significant [2], and thus appear to have been dynamically triggered. Furthermore, one of these  $M > 5$  earthquakes had a geodetic moment significantly larger than its seismic moment suggesting that deformation associated with aseismic fault creep may have indirectly triggered many of the smaller earthquakes in the area [2].

### Dynamic Triggering in Regimes with Limited Volcanic and Geothermal Activity

**Extensional and Transtensional Environments** The majority of occurrences of triggered seismicity documented to date have been in extensional or transtensional tectonic regimes (Table 1). In the western United States dynamic triggering following the  $M$  7.3 Landers earthquake occurred exclusively in transtensional tectonic regimes, many of which were also volcanically or geother-





Earthquakes, Dynamic Triggering of, Figure 6  
 Plot of earthquake magnitude versus time of seismicity for selected areas in Yellowstone caldera. See Fig. 5 for locations of these areas. Dashed line DFE as the origin time of the Denali Fault Earthquake. Figure reprinted from [45], BSSA

mally active, to distances of up to 1250 km [1,39]. These locations included Little Skull Mountain, Nevada; western Nevada, White Mountains, California, Mono Basin California, Cedar City Utah, the Wastach Front in central Utah, Burney, California, and Cascade, Idaho. The onset of triggering ranged from during the passage of the Landers wavetrain to 33 hours after the Landers earthquake. The largest of these earthquakes was a  $M = 5.6$  earthquake triggered beneath Little Skull Mountain, Nevada. Otherwise, triggered earthquakes had  $M \leq 3.0$ . The most vigorous responses containing tens to hundreds of triggered earthquakes occurred near Cedar City Utah, Western Nevada, and Cascade Idaho.

The  $M_w = 7.1$  Hector Mine earthquake also led to an impressive display of triggered earthquakes in exclusively extensional, transtensional, and geothermal envi-

ronments in the western United States. Triggered earthquakes began during the passage of the wavetrain near the Salton Trough in Indio and at the southern end of the Salton Sea [30,43]. In general, the triggered response to Hector Mine was less extensive and energetic than that of the Landers earthquake [30].

Following the Denali Fault earthquake seismicity was triggered in several extensional and transtensional areas in the western United States. [47] detected a  $M 4.6$  earthquake triggered during the Denali Fault earthquake wavetrain in Cascade, Idaho. Seismicity remained elevated for 25 days along a 500 km stretch of the Intermountain seismic belt in Utah, on the border of the Basin and Range province (Figs. 2 and 3) [68].

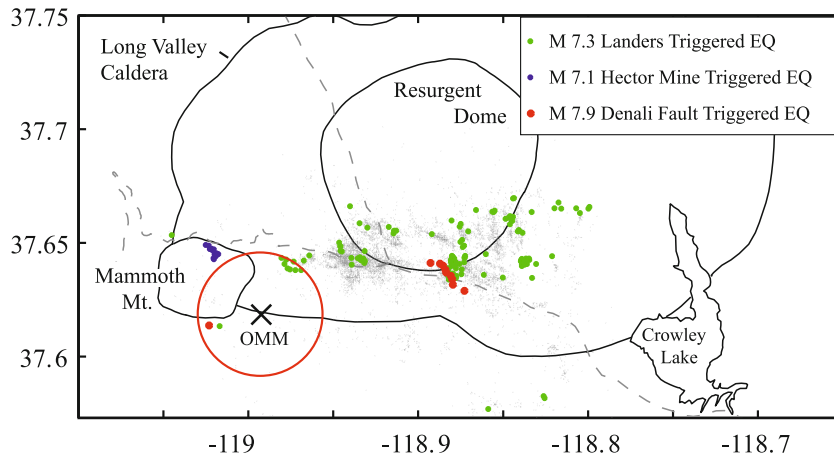
Through examining historical documents [59] identify several earthquakes in extensional/transtensional environments that may have been dynamically triggered by the  $M_w = 7.8$  1906 San Francisco earthquake, including a  $M 3.5$  and  $M 4.5$  earthquake in western Nevada and a  $M 6.1$  earthquake in the Brawley Seismic Zone near the Salton Sea in Southern California. These events are within 400–700 km from the fault rupture, thus beyond the aftershock zone of the San Francisco earthquake.

In the day following the arrival of surface waves from the  $M_w = 7.4$  Izmit, Turkey earthquake, catalog seismicity rates throughout continental Greece, 400–1000 km from the epicenter, increased significantly [6]. Greece is an area of active extension and hosts significant hydrothermal activity. Although [6] did not address a possible correlation to hydrothermal activity systematically, at least some clusters of dynamically triggered seismicity occurred in areas with active hot springs.

A second report of dynamically triggered seismicity in Europe comes the Roer Valley, the Netherlands. This area is an actively extending northern branch of the Rhine Graben System. Following a  $M_w = 5.4$  earthquake in 1992, [12] determine that a large cluster of aftershocks occurred at distance of 40 km from the mainshock. They conclude that these events are dynamically triggered because they are located beyond the zone where static stress changes are significant.

### Transpressional and Compressional Environments

Although dynamic triggering is not commonly observed in compressional environments, several studies suggest it does occur. Less than three hours after a  $M_s = 7.3$  earthquake in the Gulf of Aqaba 1995, an earthquake swarm began 500 km distant from the mainshock epicenter in a restraining bend of the Dead Sea transform fault on the Syria–Lebanon border [62]. The swarm consisted of 21 earthquakes of  $M_d \leq 3.7$ .



Earthquakes, Dynamic Triggering of, Figure 7

Map of triggered seismicity beneath Long Valley caldera and Mammoth Mountain, California, for the Landers (green), Hector Mine (blue), and Denali Fault (red) earthquakes. Gray dots show background seismicity from 1997–1998. The red circle centered on station OMM indicates area within which the earthquakes triggered by the Denali Fault earthquake must be located based on S-P phase arrival times. The single red dot was large enough to be located [74]. Modified from [38], *Treatise on Geophysics*

The central United States is a transpressional environment with low strain rates. Dynamic triggering in the central US has not yet been observed instrumentally. However, [40,44,66] suggest that dynamic triggering occurred during the 1886 Charleston, South Carolina earthquake and 1811–1812 New Madrid earthquakes based on examination of historical felt reports. Similarly, [42] describe historical evidence for dynamic triggering of a  $M \sim 7$  earthquake following the 1905 Kangra earthquake in India.

The stress state in Taiwan is variable, but generally transpressional [104]. [102] searched for dynamically triggered seismicity in the Taiwan region following 12 regional  $M 6.5+$  earthquakes occurring between 1973 and 1994. They identify 9–10 cases of increased seismicity following a large event, although the increase is small in all cases, with 1–7  $M 4$ – $4.5$  earthquakes more in the 15 days following the large earthquake than in the 15 days before.

### Dynamic Triggering in Subcrustal Environments

The occurrence of dynamically triggered earthquakes in subcrustal environments has been investigated in subduction zones in South America and Japan. [96] found that a  $M 7.6$  earthquake at 598 km depth in the Tonga trench in 2002 was followed by  $M 5.9$  and  $M 7.7$  earthquakes at 647 and 664 km depth within 2 and 7 minutes of the initial earthquake, respectively. By investigating the rupture history and Coulomb stress change resulting from the initial event, they conclude that the secondary events were triggered dynamically. They highlight 4 additional earth-

quakes of  $> 450$  km depth which have similar large aftershocks that may be dynamically triggered. During the surface waves of the  $M 8.1$  Tokachi-oki earthquake, [60] identified deep low frequency earthquakes triggered in the Nankai subduction zone through analyzing Hi-Net borehole seismic data and use of the Beta statistic. The triggering occurred during a slow slip event in a region of the subduction zone which was active with deep low frequency tremor.

### Triggered Tremor

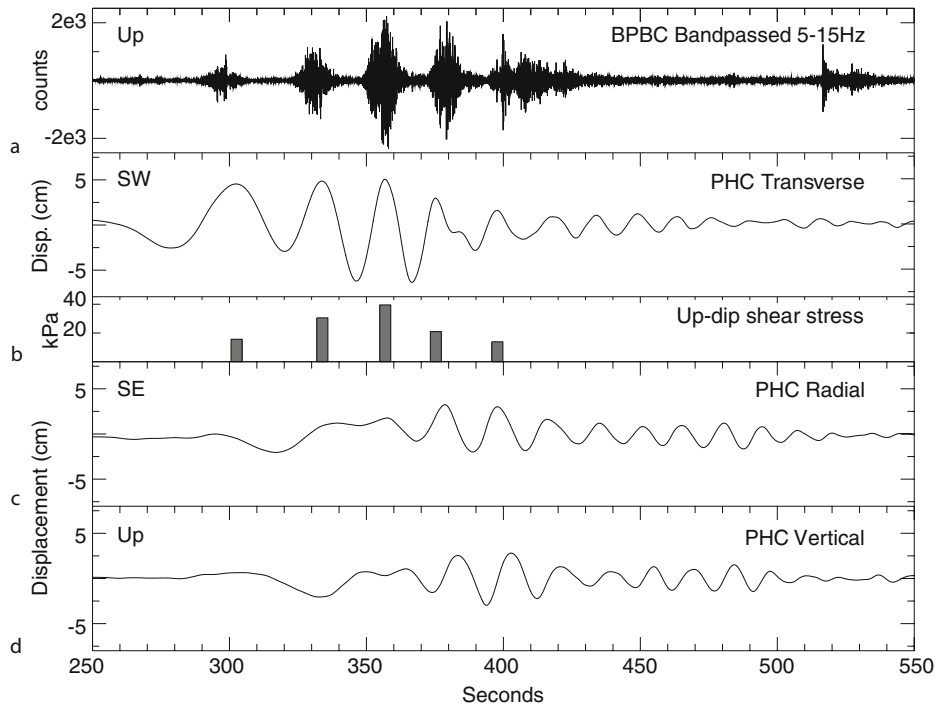
With the exception of triggered subduction zone seismicity described above, the majority of dynamically triggered earthquakes described in this review are typical brittle failure earthquakes. For example, [43] show that the earthquakes triggered by the Hector Mine earthquake near the Salton Sea had typical spectra and stress drops, consistent with standard brittle failure source mechanism. In the last few years however researchers have demonstrated that volcanic tremor and deep non-volcanic tremor respond to dynamic waves from regional and teleseismic earthquakes as well as typical crustal earthquakes (Table 2) ([32,60,61,81]). These findings emphasize that dynamic triggering can occur in a wide variety of environments and affect multiple seismic processes in addition to brittle failure of crustal rock. They provide an intriguing new perspective on the triggering processes.

At Aso volcano, Japan, [61] identify dynamically triggered earthquakes and volcanic tremor following the 1999

Earthquakes, Dynamic Triggering of, Table 2  
Published occurrences of dynamically triggered tremor

Site	Citation	Triggering Earthquake	M**	Type of tremor	Responsible phase
Aso volcano, Japan	[61]	Chi-Chi, 1999	7.7	Shallow volcanic	P waves
Cascadia subduction zone, Canada	[81]	Denali Fault, 2002	7.9	Non-volcanic subduction zone	Love waves
7 sites throughout California	[32]	Denali Fault, 2002	7.9	Non-volcanic	Surface waves

\*\* M is the magnitude of the triggering earthquake.



Earthquakes, Dynamic Triggering of, Figure 8

Time series showing tremor triggered by Love waves from the Denali Fault Earthquake in the Cascadia subduction zone: **a** Tremor at station BPBC, time adjust to correct for travel time from source to seismometer, **b–d** Displacement seismograms for transverse, radial, and vertical components at station PCH, the closest 3 component broadband station to the tremor, time adjust to correct for travel time from source to seismometer. Tremor occurs when the Love wave displacement is to the SW. Figure reprinted from [81], Nature

M 7.7 Chi-Chi earthquake. To test the uniqueness of these observations, they searched for triggered tremor at Aso following 20 other  $M_w \geq 7$  earthquakes occurring within 3000 km distance between 1995 and 2002. Five of these earthquakes triggered tremor following P wave arrivals at Aso. All occurred between 1998 and 1999, a time with usually high heat supply to the volcano's crater. As yet, this is the only documented episode of dynamically triggered volcanic tremor.

On the other side of the Pacific Ocean and a different tectonic environment, [81] identified episodes of deep non-volcanic tremor in the Cascadia subduction zone,

Canada, which were triggered by the Love waves of the M 7.9 Denali Fault earthquake. In this case tremor amplitude modulates perfectly with strain amplitude from the incident Love waves (Fig. 8).

More recently, [32] identified triggered non-volcanic tremor in seven locations in California following the Denali Fault earthquake. In all cases tremor amplitude modulates with strain amplitude from incident surface waves. Five of these are strike-slip faulting regimes. These observations are the first reported cases of non-volcanic tremor beyond subduction zones (e. g. [80]) and the San Andreas fault [67].

## Lack of Triggering Observations

Interestingly, some areas of high ambient seismicity show a notable lack of dynamically triggered seismicity. For example, the San Andreas fault near Parkfield, California showed no triggered response to the Landers earthquake [90]. Japan boasts high rates of shallow background seismicity, frequent large earthquakes from the subduction zone, high seismic network density, and a variety of crustal stress environments and volcanic and geothermal regions. However, through examining both earthquake catalogs and waveform data from individual seismic stations before and after nine large remote events, [33], show that dynamic triggering in Japan is not common, as it is in extensional regimes of the Western United States.

Similarly, Alaska abounds with crustal and subduction zone seismicity and active volcanic and geothermal systems, although network density is far lower than in Japan. Though the Katmai Volcanic cluster appears to be particularly susceptible to triggering [64,73] and dynamic triggering has been observed at Mt. Wrangell [103], dynamic triggering is rare compared to the western United States. [64] suggest that this results from unknown differences in the magmatic and hydrothermal systems of the volcanoes. [83] document a decrease in seismicity at Mt. Wrangell and Veniaminof volcanoes following the M 7.9 Denali Fault earthquake. To date these are the only documented examples of seismicity repression from a large distant earthquake.

## Characteristics of Dynamic Triggering

### Environmental Controls on Dynamic Triggering

Extensional and transtensional tectonic regimes with high levels of background seismicity are highly susceptible to dynamic triggering [38]. This may reflect the ease with which fluids can migrate upwards in these stress environments [32,38]. Because such fluids are often hot with high concentrations of dissolved solids, rapid precipitation may form high pressure compartments over rapid time scales, further enhancing a tendency toward failure. Faults in extensional stress regimes are also inherently weak compared to those in compressional environments [43,88]. [26,41] suggest that dynamic triggering is a ubiquitous process in the crust which is detected more commonly in certain areas due to high instrumentation and scrutiny levels. Only one study to date has carefully addressed this question. By comparing seismicity rates on the San Andreas fault in California and the Western United States Basin and Range Province, [90] show that the San Andreas fault is less likely to experience dynamic triggering

than similarly instrumented areas with similar levels of background seismicity in the Western United States Basin and Range province. More studies like [90] are necessary to resolve whether triggering is truly ubiquitous or favored in specific tectonic environments.

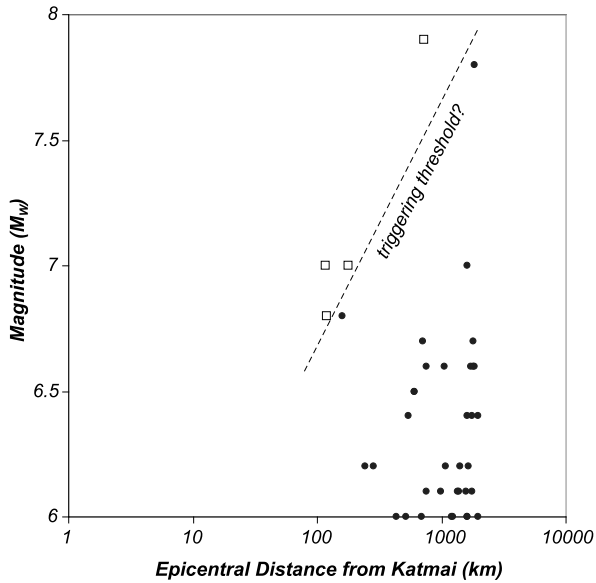
### Triggering Thresholds and Recharge Times

In most reports of remote dynamic triggering, seismicity is triggered by earthquakes of M 6.5 or greater (Table 1). Dynamic triggering responses are strongest in areas that experience strong directivity [26,30,39]. These first order observations suggest that strength of triggered response is a function of ground shaking amplitude. Although amplitude-based triggering thresholds have been suggested for some areas [28,29,31,64], a consistent triggering threshold that applies throughout the crust has not been established [38]. Large earthquakes regularly occur without dynamically triggering seismicity beyond their aftershock zones.

Lack of triggering reports below M 6.5 may reflect subtle triggered responses. [41] uses the Beta statistic to give evidence of small seismicity increases at distances of 70–110 km in the month following 14 moderate (M 5.5–7) earthquakes in California. Because this distance corresponds with where a large SMS phase should arrive, [41] suggests that the SMS phase is responsible for the triggered response in these cases.

If a simple amplitude-of-shaking threshold is required to dynamically trigger earthquakes, we would expect that even moderate earthquakes trigger seismicity near their epicenters. [21,27,52,71] give strong evidence that dynamic triggering occurs in the near field. Because it is difficult to distinguish the influence of static and dynamic stress changes in the near field, many studies of dynamic triggering have limited their investigation to the realm beyond the aftershock zone.

Because many aftershocks in the near field are likely dynamically triggered, [31] include aftershocks in a search for an amplitude-based triggering threshold. They find that peak dynamic stress distributions correlate well with aftershock and remotely triggered seismicity distributions, except in the Long Valley caldera, CA. The result of [31] is consistent with failure thresholds found in laboratory studies [49] and independent of frequency of shaking. [64] also find evidence for a ground shaking amplitude-based triggering threshold at the Katmai Volcanic Cluster, Alaska by comparing magnitude and distance of mainshock with triggered response (Fig. 9). Their magnitude-distance relationship is similar to that proposed by [28] for the Geysers, CA. However, the triggering



Earthquakes, Dynamic Triggering of, Figure 9  
 Plot of magnitude vs. distance from Mageik volcano in the Katmai Volcanic Cluster for all  $M_w > 6$  earthquakes between 1996 and 2003 located within 2000 km of Katmai. *Hollow squares* triggered seismicity in the KVC. *Solid circles* did not. *Dashed line* represents possible triggering threshold. Figure reprinted from [64], BSSA

threshold at Katmai appears to be higher than that suggested for the Geysers.

In other cases, a simple amplitude-of-shaking threshold is not consistent with data, and large amplitude ground shaking is neither a necessary nor sufficient condition to cause dynamically triggered earthquakes. [31] show that the Long Valley caldera appears to be more susceptible to triggering than other areas they studied. Because their study was based on catalog seismicity, it did not include triggered earthquakes that were too small to appear in earthquake catalogs, such as those at the Coso geothermal field in response to the Denali Fault earthquake. These events were triggered by dynamic stresses of  $< 0.01$  MPa [74] and, like Long Valley, would not fit the thresholds proposed by [29] and [31].

By comparing spectra of all earthquakes with high amplitude ground shaking in the Long Valley caldera [7] find that in this area, high-amplitude low-frequency shaking is more likely to trigger seismicity than high-amplitude high-frequency shaking. [1] come to the same conclusion after examining strong ground shaking spectra of earthquakes which did and did not trigger seismicity in the Western Great Basin. Longer wave lengths associated with low frequency ground shaking favor triggering by larger earthquakes in at least some locales. Whether remote dy-

amic triggering in both the near and far field results from the same physical process or processes remains an open question.

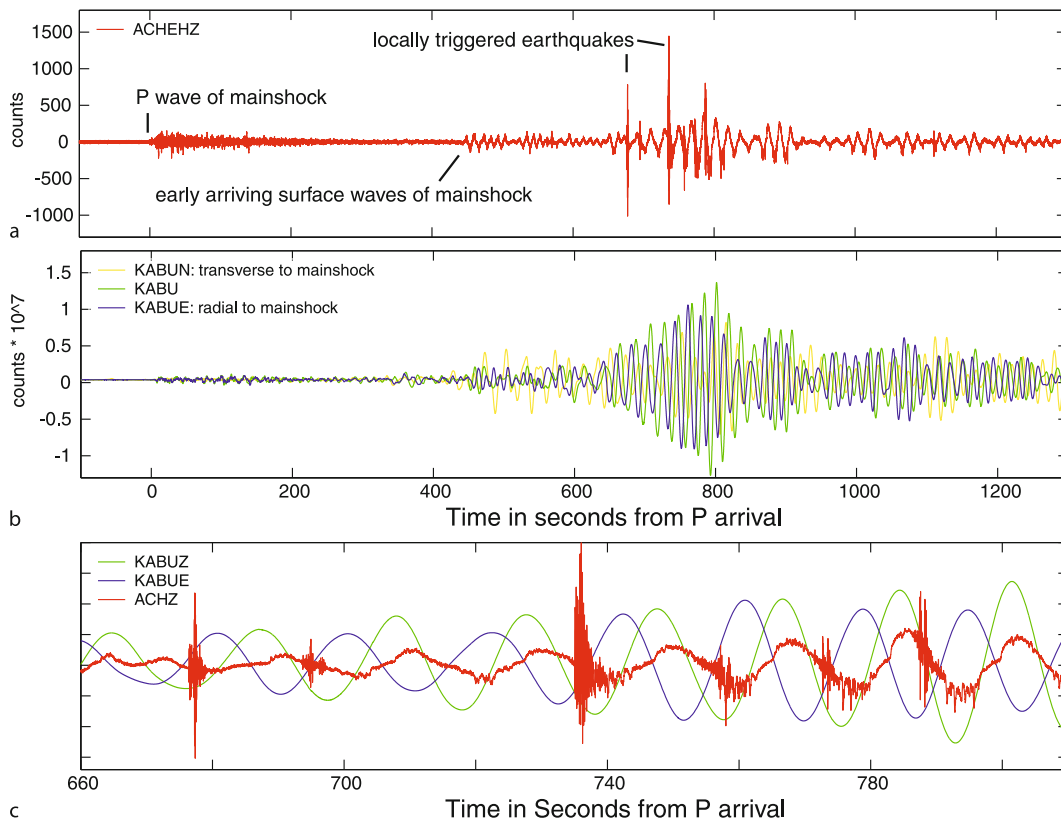
One parameter that may complicate the search for a triggering threshold in amplitude and/or frequency is recharge time. Because the occurrence of earthquakes releases stored strain energy, it may take time for an area to re-accumulate strain energy sufficiently to be primed for failure again following local earthquake activity or previous episodes of remotely triggered seismicity [38]. However some areas, such as the Geysers geothermal field require little to no time to recharge, as triggered seismicity episodes have been separated by time intervals of months or less [28]. Recharge times are dependent on many parameters including earthquake history, regional tectonic strain rates, and mass and heat advection rates in areas of hydrothermal and volcanic activity.

### Time Scales of Dynamic Triggering and Responsible Phases

Remote dynamic triggering of earthquakes occurs over a variety of time scales following the onset of dynamic stressing. At Aso Volcano, Japan [61] triggered tremor begins with the P-wave arrival from distant large earthquake. In the case of discrete earthquakes however, it is most common for triggering to begin during surface wave arrivals (Table 1), leading many to suggest that the specific low-frequency large-amplitude ground motions associated with surface waves initiate the failure process [1,7,38,103]. Although the onset of dynamic triggering at remote locations is most commonly observed during Rayleigh wave arrivals [38], clear cases of remote triggering of tremor on the Love wave exist as well [81].

In some cases, dynamic triggering begins hours to days after an initial stress perturbation (e.g. [39,89,102]), hinting that the physical process responsible for initiating earthquake failure evolve with time. For example the largest triggered event following the M 7.3 Landers earthquake, a M 5.6 at the Little Skull Mountain, Nevada, occurred 33 hours after the mainshock [39]. In the case of Long Valley caldera's south moat and Mt. Rainier after the Denali Fault earthquake, delayed earthquake swarms began 24 hours and 2 hours respectively after the passage of the dynamic waves from the mainshock (Fig. 4) [74]. Both of these areas also had much smaller triggered swarms during the mainshock's wavetrain.

Determining the duration and decay time of triggered swarms is more difficult than detecting their onsets, particularly in areas of high ambient seismicity. Many triggered earthquakes may be triggered secondarily as aftershocks



Earthquakes, Dynamic Triggering of, Figure 10

Phase modulated dynamically triggered earthquakes in the Katmai Volcanic Cluster following the 2007 M 8.2 Kurile earthquake: **a** short period record from station ACH showing both wavetrains for the Kurile earthquake and the larger amplitude, locally triggered earthquakes, **b** broadband record from station KABU showing wave motion of the Kurile earthquake, **c** time series from ACH and KABU zoomed in to show how local earthquakes seen clearly in red are occurring on a specific phase of the wavetrain from the Kurile earthquake

to earlier triggered earthquakes [9]. The Yellowstone response to the Denali Fault earthquake and the Long Valley caldera response to Landers are fit well with an Omori-type law decay [45]. In some cases however, triggered swarms end abruptly after the dynamic stress perturbation stops (e. g. [103]). Although our understanding of decay rates of triggered swarms is incomplete emphasizing that the subject deserves further investigation, decay rates give strong constraints on physical processes responsible for triggering.

### Phase Modulated Triggering

Recent findings show that earthquakes can be triggered during specific phases of the wavetrain. At Mt. Wrangell, Alaska, triggered earthquakes occurred preferentially during phases of the largest positive vertical ground displacement from the 2004 M 9.0 Sumatra earthquake [103]. Sim-

ilarly at Katmai Volcanic Cluster, Alaska, triggered earthquakes occurred only during specific phases of Rayleigh waves from the 2007 M 8.2 Kurile earthquake (Fig. 10). Such observations will allow us to resolve the precise dynamic stress field perturbations at the moment of earthquake nucleation on specific failure planes (e. g. [38]).

### Physical Models of Dynamic Triggering

The wide variations in the characteristics of dynamic triggering and the limited data for individual response instances admit a spectrum of competing models for the physical processes linking dynamic stresses from a large, distant earthquake to the locally triggered response. Broadly considered, published models fall into three partially overlapping categories: 1) those involving some form of stress-driven brittle failure across local fractures, 2) those involving the activation of hydrous or mag-

matic fluids, and 3) those involving some form of localized aseismic relaxation (deformation). The brittle failure models are generally consistent with the onset of locally triggered seismicity during dynamic stressing (rapid-onset triggering), including the possibility that seismicity may persist as aftershocks for some time after the dynamic stressing has stopped [21]. Under the latter two categories, the onset of local seismicity represents a second-order phenomenon driven by a first order response to dynamic stressing in the form of fluid activation or transient deformation. In principle, models under these two categories admit a significant delay in the onset of the triggered seismicity with respect to the dynamic stresses generated by a distant earthquake. Because the dynamic stress amplitudes that trigger a response at remote distances are typically an order of magnitude or more below background tectonic stress levels, all models carry the implicit assumption that a crustal volume susceptible to dynamic stress triggering must be in a near-critical stress state prior to a triggered response.

### Brittle Failure

Brittle failure models are based on the premise that the dynamic stresses propagating with the seismic waves from a distant earthquake are sufficient to nudge the local stress acting on a pre-existing dislocation beyond the threshold for the particular failure mode. This threshold may be the Griffith criteria for the tensile strength of a partially healed crack or the Coulomb criteria for frictional strength of a fault [86,87]. Crustal fluids play an important passive role in all brittle failure models by counteracting the rock matrix stress acting on a fracture through pore pressure,  $p$ , according to

$$\sigma' = \sigma - Ip$$

where  $\sigma'$  and  $\sigma$  are the effective and rock matrix stress tensors, respectively, and  $I$  is the identity tensor. Thus, pore pressure reduces the effective normal stress,  $\sigma'_n$ , acting on a fracture by opposing the rock matrix normal stress as  $\sigma'_n = \sigma_n - p$ . Alternatively, for pressure-sensitive friction models the role of pore pressure can be expressed in terms of an effective coefficient of friction as  $\mu' = \mu(1 - \lambda_p)$ , where  $\lambda_p = p/\sigma_n$ . Elevated pore pressures lower the effectively strength by moving the background stress state closer to extensional or shear failure thresholds thereby increasing vulnerability for failure by imposition of small dynamic stress perturbations.

In the simplest frictional failure model, a triggered earthquake occurs when the stress acting on a fault exceeds the Coulomb threshold for static friction, or  $CFF(t) = 0$ ,

and friction abruptly drops from static to dynamic values with  $\mu_s > \mu_d$ , respectively. Here,  $CFF(t)$  is the Coulomb Failure Function defined as

$$\begin{aligned} CFF(t) &= |\tau(t)| - \mu_s \sigma'_n(t) - C, \quad \text{or its equivalent} \\ &= |\tau(t)| - \mu'_s \sigma_n(t) - C \end{aligned}$$

where  $\sigma_n$ ,  $\sigma'_n$ ,  $\mu_s$ ,  $\mu'_s$  are defined in the preceding paragraph,  $\tau$  is the shear, and  $C$  is the cohesive strength ([34], and references therein). This simple case implies rapid-onset triggering with the triggered seismicity beginning promptly when  $CFF(t)$  first becomes positive for a fault optimally oriented for failure in the background stress field. The combination of dynamic stress components  $\Delta\tau$  and  $\Delta\sigma_n$  for which  $CFF > 0$  will depend on the wave type (e.g. Love or Rayleigh wave) and its incidence angle on the optimally oriented fault [35]. Although details vary, Love waves will generally have a greater triggering potential than Rayleigh waves when incident on vertical, strike-slip faults while the opposite is the case for incidence on inclined, dip-slip faults.

The Coulomb failure criterion applies to more elaborate non-linear friction models as well (see [18,25,31,70,100]). Because the behavior of non-linear models depend on factors such as slip history and slip rate, however, the failure threshold for static friction may vary with time, and the triggered earthquake may be delayed with respect to the time the failure criterion was first exceeded (e.g. [69]). Susceptibility to dynamic triggering may result when a dynamic stress is imposed on quasi-static loading under a conditionally stable regime (e.g. [84]). Based on their analysis of the dynamic triggering observed at Long Valley caldera, [7] conclude that this mechanism requires near-lithostatic pore pressures to be effective.

Models based on the non-linear response of granular media to dynamic stresses may apply to dynamic triggering of mature faults with a well-developed core of fault gouge. [49] document an abrupt decrease in the modulus of fault gouge under low effective normal stress ( $\sigma'_n$  0.1 MPa) when excited by dynamic strains  $> 10^{-6}$  in the laboratory. Thus, this model also requires near-lithostatic pore pressures to be effective.

Sub-critical crack growth, or stress corrosion, is another non-linear form of brittle failure that has a potential role in dynamic triggering. Under this model, a sudden increase in differential stress or an oscillatory stress applied to a pre-existing crack can lead to crack growth due to weakening of the crack tip by chemical corrosion. This can shorten the time to earthquake rupture. This process will be enhanced in an environment with fluids at elevated temperatures. It turns out that the equations governing sub-critical crack growth have the same mathemat-

ical form as rate-state friction equations above [51]. Thus near-lithostatic pore pressure appears to be a requirement for each of these non-linear brittle-failure models, at least as they apply to dynamic triggering at remote distances.

### Fluid Activation Models

In addition to their passive role in reducing the effective strength of a rock volume through ambient pore pressure, fluids may play an active role in the dynamic triggering process. Fluid activation models generally appeal to either 1) pore-pressure re-distribution associated with changes in permeability and fluid transport, or 2) state changes induced in multi-phase fluids.

Dynamic stressing may be capable of physically disrupting permeability barriers separating volumes of differing pore pressure. For example, dynamic stress may shake accumulated detritus from clogged fractures or opening partially healed fractures by extensional failure. In either case, fluid diffusion down the pressure gradient will result in a re-distribution of pore pressure with the potential for triggering seismicity in previously under-pressured volumes in a near-critically stressed state. The evolution of triggered seismicity in this case will be governed by the diffusion length for a given permeability and the proximity of the pre-existing stress state to brittle failure. [8] proposed the clogged fracture model as an explanation for the hydrologic response of water wells in southern Oregon to surface waves from  $M > 7$  earthquakes at distances of 300 km and 3850 km.

Geothermal areas may be particularly susceptible to dynamic triggering through pore pressure re-distribution. In these areas fractures are rapidly sealed by precipitation from circulating, solute-rich geothermal fluids and plastic deformation of quartz-rich rocks under elevated temperatures tend to isolate pockets of elevated pore pressure. Most active geothermal systems are located in areas of extensional tectonism. In these areas normal stresses induced by Rayleigh waves on vertical planes may open vertical fractures, allowing high-pore-pressure fluids access to shallower crustal volumes with lower pore pressure [35]. The hydraulic surge model described by [22] for volcanic and geothermal systems is a version of this process in which the brittle-plastic transition at the base of the seismogenic crust serves as a low-permeability barrier separating near-lithostatic pore pressures in the plastic regime from a hydrostatic regime in the overlying seismogenic crust. Rupturing the permeability seal by dynamic stresses would release near-lithostatic pore pressures into the brittle, seismogenic crust thereby inducing a surge in triggered seismicity.

Models for bubble excitation by dynamic stresses in a two-phase fluid (multi-phase in a partially crystallized magma) offers interesting possibilities for remotely triggered responses in geothermal and volcanic systems. This is a particularly intriguing concept for remote triggering in volcanic systems because of the importance of bubbles in eruption dynamics [57] and the source mechanisms of long-period volcanic earthquakes [13], ► **Volcanoes, Non-linear Processes in**. Advective overpressure and rectified diffusion were the first bubble models proposed as explanations for remotely triggered seismicity [10,55,94]; although subsequent work has shown that both hold less promise as viable explanations than initially thought [56].

Under the advective overpressure model, the pressure in a gas-saturated, incompressible fluid confined in a rigid container increases as  $\rho g \Delta h$  as a pre-existing bubble adhering to the wall of the container is shaken loose by passing seismic waves. The bubble ascends buoyantly a distance  $\Delta h$  through a fluid of density  $\rho$  where  $g$  is the acceleration of gravity [55]. The resulting pressure increase in the container (magma body) deforms the surrounding rock inducing small earthquakes. This model is criticized on the basis that assumptions of a ridged container and an incompressible fluid seriously violate realistic conditions in the earth [75].

Under rectified diffusion, pressure oscillations imposed on a gas-saturated fluid with pre-existing bubbles pump gas into the bubbles over multiple cycles. Gas evolves from the fluid into the bubble during the dilatational phase, when bubble surface area is maximal, and out of the bubble back into solution during compressional phase, when the bubble surface area is minimal [94]. The implied pressure gain integrated over multiple cycles is then transmitted to the surrounding rock inducing small earthquakes. [48] point out, however, that the effectiveness of this model is limited by reasonable gas diffusion rates in hydrous fluids or magma with respect to the frequencies of seismic waves driving the pressure oscillations.

More promising bubble models appeal to the strong sensitivity of bubble nucleation rate to the supersaturation pressure [56] and the results of numerical models by ► **Volcanoes, Non-linear Processes in** and [14,85], indicating that a small pressure drop imposed on a densely packed matrix of tiny bubbles can lead to rapid, diffusion-driven bubble growth. The implications of these models, however, have yet to be more fully explored in the context of dynamic triggering.

Two more speculative models involve magma instabilities triggered by dynamic stresses. In one, a loosely held crystal mush accumulated on the walls of a crystal-



lizing magma body may be dislodged by dynamic shaking. The sinking crystal mush would induce a convective plume as it displaced hotter, less dense magma. In the case of volatile-rich magma, buoyant convection would be enhanced by bubble nucleation and growth as confining pressure drops with decreasing depth [37]. Under suitable conditions, the resulting pressure increase within the magma body could evolve over days [56]. If the magma chamber was already in a near critical state, the culmination could be magma intrusion into the overlying crust or the onset of an eruption. Whether this process culminates in a simple pressure increase, an intrusion, or an eruption, the sensible onset of locally triggered seismicity and deformation might be delayed by hours to perhaps days with respect to the passing seismic waves from the distant earthquake. A second, even more speculative model appeals to dynamic stresses disrupting the solid matrix of a partially crystallized magma body thereby releasing any differential tectonic stress sustained by the solid matrix [36,37]. As the magma body relaxes with a time constant governed by the effective viscosity of the disrupted crystal mush, stress would be transferred to the surround crust inducing deformation and local seismicity. In essence, this model corresponds to the relaxation of an Eshelby inclusion in an elastic medium [20].

### Aseismic Deformation

The relaxing magma body of the previous paragraph is one example of aseismic deformation with the potential of triggering local deformation and the onset of secondary seismicity. A less speculative example involves aseismic creep on faults triggered by dynamic stressing. Deformation associated with fault creep transfers stress to the adjacent crust, which in turn triggers local seismicity, as in the example involving seismicity triggered on the Reykjanes Peninsula following the  $M = 6.5$  earthquake in the South Iceland Seismic Zone in 2000 [2]. [4] document aseismic fault slip (creep) on faults in the Salton Trough of southern California triggered by the three  $M > 6$  earthquakes in the Landers, California sequence of 1992 (the  $M = 6.1$  Joshua Tree,  $M = 7.3$  Landers, and the  $M = 6.2$  Big Bear earthquakes). In this case, all instances of triggered slip were on faults within 150 km of the  $M > 6$  earthquakes. In these examples and observations from triggering in Long Valley caldera and Sierra Prieto geothermal field in Baja California [38,50] the geodetic moment for triggered aseismic deformation exceeds the cumulative seismic moment for the triggered earthquakes by a factor of two or more. This emphasizes the importance of high-resolution deformation monitoring in ar-

reas susceptible to dynamic triggering for resolving the role of aseismic deformation in the dynamic triggering process.

### Future Directions

In the last 25 years, in the wake of the Landers earthquake, the study of dynamically triggered seismicity has given us new insight into earthquake initiation and the failure regime in the Earth's crust. Some argue that the state of stress in the crust is highly spatially variable [77]. Given this, the likelihood of triggering seismicity would also be spatially variable. [97] and [105], however, argue that the Earth's crust is critically stressed and on the verge of failure nearly everywhere. If this were the case, one might expect triggering due to small dynamic stress perturbations to be a ubiquitous phenomenon. In either case, the study of remotely triggered seismicity provides clues to spatial distribution of critically stressed crustal volumes.

Unfortunately, although we can measure stress field perturbations from dynamic waves from earthquakes, we rarely have a detailed understanding of the background stress field these perturbations are modulating. In addition, dynamically triggered earthquakes are often too small or occur in too sparsely instrumented areas to resolve reliable focal mechanisms. Because of these limitations, our understanding of how dynamic stresses from remote earthquake wavetrains induce a given crustal volume to respond with triggered seismicity remains incomplete. Advances will require more cases of dynamically triggered seismicity captured by both spatially dense seismic networks and continuous, high-resolution deformation monitoring networks.

Recent observations of phase modulated dynamic triggering offer powerful datasets of the precise time history of dynamic stress triggering. Because similarities exist between phase modulated dynamic triggering of seismicity in the shallow crust of a volcano's edifice [103] and deep in a subduction zone [81], the emerging study of non-volcanic tremor may provide new leverage on understanding how dynamic stresses influence seismic slip.

As new observations of dynamically triggered seismicity are reported, one conclusion is becoming increasingly evident: multiple causative processes exist. The wide variety in time scales over which triggering occurs and the spatial and temporal characteristics of triggered seismicity sequences and associated deformation responses cannot be fit with any one model yet proposed. Rather, different models are consistent with different episodes of triggering. For example, fluid activation and stress corrosion models

are most applicable in volcanically and geothermally active environments. In some cases, such as the complex triggered response of Yellowstone, Mt. Rainier, and the Long Valley caldera areas to the Denali Fault earthquake, multiple processes may be occurring simultaneously in the same locale, yet on different time scales. Of the physical models described above, seismicity triggered instantaneously or within seconds of the dynamic stress perturbation is consistent with models based on simple brittle failure, brittle failure with nonlinear friction effects, stress corrosion, unclogging of fractures, or rectified diffusion, whereas triggered seismicity delayed by hours to days is more consistent with models involving aseismic deformation, advective overpressure, sinking crystal plumes, or a relaxing magma body.

In the few cases where hydrologic and high-sample rate strain data are available, dynamically triggered seismicity is accompanied by changes in water levels in wells [79] and significant deformation signals [2,25,36,50]. A complete understanding of dynamic triggering will require research approaches that integrate seismic, deformation, and hydrologic datastreams. To this end, we challenge Earth scientists to broaden their thinking and tap these observations to better understand the initiation of earthquake failure.

## Bibliography

### Primary Literature

- Anderson JG, Brune JN, Louie JN, Zeng Y, Savage M, Yu G, Chen Q, de Polo D (1994) Seismicity in the western Great Basin apparently triggered by the Landers, California, earthquake, 28 June 1992. *Bull Seismol Soc Am* 84:863–891
- Arnadottir T, Geirsson H, Einarsson P (2004) Coseismic stress changes and crustal deformation on the Reykjanes Peninsula due to triggered earthquakes on 17 June 2000. *J Geophys Res.* doi:10.1029/2004JB003130
- Blackett RE, Wakefield S (2002) Geothermal resources of Utah, Utah geological survey open file report 397, ISBN 1-55791-677-2
- Bodin P, Bilham R, Behr J, Gomberg J, Hudnut K (1994) Slip triggered on southern California faults by the 1992 Joshua Tree, Landers, and Big Bear earthquakes. *Bull Seismol Soc Am* 84:806–816
- Bodin P, Gomberg J (1995) Triggered seismicity and deformation between the Landers, California and Little Skull Mountain Nevada earthquakes. *Bull Seismol Soc Am* 84:835–843
- Brodsky EE, Karakostas V, Kanamori H (2000) A new observation of dynamically triggered regional seismicity: earthquakes in Greece following the August, 1999, Izmit, Turkey earthquake. *Geophys Res Lett* 27:2741–2744
- Brodsky EE, Prejean SG (2005) New constraints on mechanisms of remotely triggered seismicity at Long Valley Caldera. *J Geophys Res.* doi:10.1029/2004JB003211
- Brodsky EE, Roeloffs E, Woodcock D, Gall I, Manga M (2003) A mechanism for sustained groundwater pressure changes induced by distant earthquakes. *J Geophys Res.* doi:10.1029/2002JB002321
- Brodsky EE (2006) [http://www.pmc.ucsc.edu/~brodsky/reprints/Sus5\\_merged.pdf](http://www.pmc.ucsc.edu/~brodsky/reprints/Sus5_merged.pdf). Long-range triggered earthquakes that continue after the wavetrain passes. *Geophys Res Lett* 33:L15313
- Brodsky EE, Sturtevant B, Kanamori H (1998) Earthquakes, volcanoes, and rectified diffusion. *J Geophys Res* 103:23827–23838
- Brune J (1970) Tectonic stress and the spectra of seismic shear waves from earthquakes. *J Geophys Res* 75:4997–5009
- Camelbeek T, van Eck T, Pelzing R, Ahorner L, Loohuis J, Haak HW, Hoang-Trong P, Hollnack D (1994) The 1992 Roermond earthquake, the Netherlands, and its aftershocks. *Geologie en Mijnbouw* 73:181–197
- Chouet B (1992) A seismic model for the source of long-period events and harmonic tremor. In: Gasparini P, Scarpa R, Aki K (eds) *Volcanic seismology*. IAVCEI Proceedings in Volcanology. Springer, Berlin, pp 133–156
- Chouet B, Dawson P, Nakano M (2006) Dynamics of diffusive bubble growth and pressure recovery in a bubbly rhyolitic melt embedded in an elastic solid. *J Geophys Res.* doi:10.1029/2005JB004174
- Cocco M, Rice JR (2002) Pore pressure and poroelasticity effects in Coulomb stress analysis of earthquake interactions. *J Geophys Res* doi:10.1029/2002JB002319
- Cochran ES, Vidale JE, Tanaka S (2004) Earth tides can trigger shallow thrust fault earthquakes. *Science* 306:1164–1166
- Cooper HH, Bredehoeft JD, Papadopoulos S, Bennett RR (1965) The response of well-aquifer systems to seismic waves. *J Geophys Res* 70:3915–3926
- Dieterich JH (1979) Modeling of rock friction 1, Experimental results and constitutive equations. *J Geophys Res* 84:2161–2168
- Elkhoury JE, Brodsky EE, Agnew DC (2006) Seismic waves increase permeability. *Nature* 441:1135–1138
- Eshelby JD (1957) The determination of the elastic field of an ellipsoidal inclusion, and related problems. *Proceedings of the Royal Society of London A* 241:376–396
- Felzer KR, Brodsky EE (2006) <http://www.nature.com/nature/journal/v441/n7094/full/nature04799.html>. Decay of after-shock density with distance indicates triggering by dynamic stress. *Nature* 441:735–738
- Fournier RO (1999) Hydrothermal processes related to movement of fluid from plastic to brittle rock in the magmatic-epithermal environment. *Economic Geology* 94:1193–1211
- Freed AM (2005) Earthquake triggering by static, dynamic, and postseismic stress transfer. *Annual Rev Earth and Planet Sci* 33:1255–1256
- Glowacka E, Nava AF, Cossio DD, Wong V, Farfan F (2002) Fault slip, seismicity, and deformation in the Mexicali Valley, Baja California, Mexico, after the M 7.1 Hector Mine earthquake. *Bull Seismol Soc Am* 92:1290–1299
- Gomberg J, Blanpied ML, Beeler NM (1997) Transient triggering of near and distant earthquakes. *Bull Seismol Soc Am* 87:294–309
- Gomberg J, Bodin P, Larson K, Dragert H (2004) Earthquakes nucleated by transient deformations caused by the M=7.9 Denali, Alaska, earthquake. *Nature* 427:621–624

27. Gomberg J, Bodin P, Reasenber PA (2003) Observing earthquakes triggered in the near field by dynamic deformations. *Bul Seismo Soc Am* 93:118–138
28. Gomberg J, Davis S (1996) Stress/strain changes and triggered seismicity at The Geysers, California. *J Geophys Res* 101:733–749
29. Gomberg J, Johnson P (2005) Dynamic triggering of earthquakes. *Nature* 437:830
30. Gomberg J, Reasenber PA, Bodin P, Harris R (2001) Earthquakes triggering by seismic waves following the Landera and Hector Mine earthquakes. *Nature* 411:462–465
31. Gomberg J, Reasenber PA, Cocco M, Belardinelli ME (2005) A frictional population model of seismicity rate change. *J Geophys Res*. doi:10.1029/2004JB003404
32. Gomberg J, Rubinstein JL, Peng Z, Creager KC, Vidale JE, Bodin P (in press) Widespread triggering of non-volcanic tremor in California. *Science* 319:117
33. Harrington RM, Brodsky EE (2006) The absence of remotely triggered seismicity in Japan. *Bul Seismo Soc Am* 96:871–878
34. Harris RA (1998) Introduction to a special section: Stress triggers, stress shadows, and implications for seismic hazards. *J Geophys Res* 103:24347–24358
35. Hill DP (2008) Dynamic stresses, Coulomb failure, and remote triggering. *Bul Seismo Soc Am* 98:66–92
36. Hill DP, Johnston MJS, Langbein JO (1995) Response of Long Valley caldera to the Mw = 7.3 Landers, California, earthquake. *J Geophys Res*. doi:10.12985-13005
37. Hill DP, Pollitz F, Newhall C (2002) Earthquake-volcano interactions. *Physics Today* 55:41–47
38. Hill DP, Prejean SG (2007) Dynamic triggering. In: Kanamori H (ed) *Geophysical treatise, earthquake seismology*. Elsevier, Amsterdam
39. Hill DP, Reasenber PA, Michael A, Arabaz WJ, Beroza G, Brumbaugh D, Brune JN, Castro R, Davis S, dePolo D, Ellsworth WL, Gomberg J, Harmsen S, House L, Jackson SM, Johnston MJS, Jones L, Keller R, Malone S, Munguia L, Nava S, Pechmann JC, Sanford A, Simpson RW, Smith RB, Stark M, Stickney M, Vidal A, Walter A, Wong A, Zollweg J (1993) Seismicity remotely triggered by the magnitude 7.3 Landers, California, earthquake. *Science* 260:1617–1622
40. Hough SE (2001) Triggered earthquakes and the 1811–1812 New Madrid, central United States, earthquake sequence. *Bul Seismo Soc Am* 91:1547–1581
41. Hough SE (2005) Remotely triggered earthquakes following moderate mainshocks (or why California is not falling into the ocean). *Seismological Research Letters* 76:58–66
42. Hough SE, Billham R, Ambraseys N, Field N (2005) Revisiting the 1897 Shillong and 1905 Kangra earthquakes in northern India: site response, Moho reflections and a triggered earthquake. *Current Science* 88:1632–1638
43. Hough SE, Kanamori H (2002) Source properties of earthquakes near the Salton Sea triggered by the 16 October 1999 Mw 7.1 Hector Mine, California, earthquake. *Bul Seismo Soc Am* 92:1281–1289
44. Hough SE, Seeber L, Armbruster JG (2003) Intraplate triggered earthquakes: observations and interpretation. *Bul Seismo Soc Am* 93:2212–2221
45. Husen S, Taylor R, Smith RB, Healsler H (2004) Changes in geyser eruption behavior and remotely triggered seismicity in Yellowstone National Park produced by the 2002 M 7.9 Denali fault earthquake, Alaska. *Geology* 32:537–540
46. Husen S, Wiemer S, Smith RB (2004) Remotely triggered seismicity in the Yellowstone National Park region by the 2002 Mw 7.9 Denali Fault earthquake, Alaska. *Bul Seismo Soc Am* 94:5317–5331
47. Husker AL, Brodsky EE (2004) Seismicity in Idaho and Montana triggered by the Denali Fault earthquake: a window into the geologic context for seismic triggering. *Bul Seismo Soc Am* 94:5310–5316
48. Ichihara M, Brodsky EE (2006) <http://www.pmc.ucsc.edu/~brodsky/reprints/2005GL024753.pdf>. A limit on the effect of rectified diffusion in volcanic systems. *Geophys Res Lett*. doi:10.1029/2005GL024753
49. Johnson P, Jia X (2005) Nonlinear dynamic, granular media and dynamic earthquake triggering. *Nature* 437:871–874
50. Johnston MJS, Prejean SG, Hill DP (2004) Triggered deformation and seismic activity under Mammoth Mountain in Long Valley caldera by the 3 November 2002 Mw 7.9 Denali Fault earthquake. *Bul Seismo Soc Am* 94:5360–5369
51. Kanamori H, Brodsky EE (2004) [http://www.pmc.ucsc.edu/~brodsky/reprints/rpp4\\_8\\_R03.pdf](http://www.pmc.ucsc.edu/~brodsky/reprints/rpp4_8_R03.pdf). The physics of earthquakes, Reports on Progress in Physics 67:1429–1496
52. Kilb D, Gomberg J, Bodin P (2000) Triggering of earthquake aftershocks by dynamic stresses. *Nature* 408:570–574
53. King GCP, Cocco M (2001) Fault interactions by elastic stress changes: new clues from earthquake sequences. *Advances in Geophysics* 44:1–38
54. King GCP, Stein RS, Lin J (1994) Static stress changes and the triggering of earthquakes. *Bul Seismol Soc Am* 84:935–953
55. Linde AT, Sacks IS, Johnston MJS, Hill DP, Billham RG (1994) Increased pressure from rising bubbles as a mechanism for remotely triggered seismicity. *Nature* 371:408–410
56. Manga M, Brodsky EE (2006) Seismic triggering of eruptions in the far field: volcanoes and geysers. *Annual Rev Earth and Planet Sci* 34:263–291
57. Mangan M, Sisson T (2000) Delayed, disequilibrium degassing in rhyolite magma: decompression experiments and implications for explosive volcanism. *Earth and Planetary Science Letters* 183:441–455
58. Matthews MV, Reasenber PA (1988) Statistical methods for investigating quiescence and other temporal seismicity patterns. *Pure Appl Geophys* 126:357–372
59. Meltzner AJ, Wald DJ (2003) Aftershocks and triggered events of the great 1906 California earthquake. *Bul Seismo Soc Am* 93:2160–2186
60. Miyazawa M, Mori J (2005) Detection of triggered deep low-frequency events from the 2003 Takachi-oki earthquake. *Geophys Res Lett* 32:L10307
61. Miyazawa M, Nakanishi I, Sudo Y, Ohkura T (2005) Dynamic response of frequent tremors at Aso volcano to teleseismic waves from the 1999 Chi-Chi, Taiwan earthquake. *J Vol Geotherm Res* 147:173–186
62. Mohamad RA, Darkal N, Seber D, Sandoval E, Gomez F, Barazangi M (2000) Remote earthquake triggering along the Dead Sea Fault in Syria following the 1995 Gulf of Aqaba earthquake ( $M_s = 7.3$ ). *Seismol Res Lett* 71:47–52
63. Moran SC (2003) Multiple seismogenic processes for high-frequency earthquakes at Katmai National Park, Alaska: evidence from stress tensor inversions of fault plane solutions. *Bul Seismo Soc Am* 93:94–108
64. Moran SC, Power JA, Stihler SD, Sanchez JJ, Caplin-Auerbach J (2004) Earthquake triggering at Alaskan volcanoes follow-

- ing the 3 November 2002 Denali Fault earthquake. *Bul Seismo Soc Am* 94:5300–5309
65. Moran SC, Zimbelman DR, Malone SD (2003) A model for the magmatic-hydrothermal system at Mount Rainier, Washington, from seismic and geochemical observations. *Bull Volcanol* 61:425–436
  66. Mueller K, Hough SE, Bilham R (2004) Analysing the 1811–1812 New Madrid earthquakes with recent instrumentally recorded aftershocks. *Nature* 429:284–288
  67. Nadeau R, Dolenc D (2005) Nonvolcanic tremors deep beneath the San Andreas Fault. *Science* 300:1942–1943
  68. Pankow KL, Arabasz WJ, Pechmann JC, Nava SJ (2004) Triggered seismicity in Utah from the 3 November 2002 Denali Fault earthquake. *Bul Seismo Soc Am* 94:5332–5347
  69. Parsons T (2005) A hypothesis for delayed dynamic earthquake triggering. *Geophys Res Lett* 32:L04302
  70. Perfettini HJ, Schmittbuhl J, Cochard A (2003) Shear and normal load perturbations on a two-dimensional continuous fault: 2. dynamic triggering. *J Geophys Res.* doi:10.1029/2002JB001805
  71. Pollitz FF, Johnston MJS (2006) Direct test of static-stress versus dynamic-stress triggering of aftershocks. *Geophys Res Lett* 33:L15318
  72. Pollitz FF, Sacks IS (2002) Stress triggering of the 1999 Hector Mine earthquake by transient deformation following the 1992 Landers earthquake. *Bul Seismo Soc Am* 92:1487–1496
  73. Power JA, Moran SC, McNutt SR, Stihler SD, Sanchez JJ (2001) Seismic response of the Katmai volcanoes to the 6 December 1999 magnitude 7.0 Karluk Lake earthquake, Alaska. *Bul Seismo Soc Am* 91:57–63
  74. Prejean SG, Hill DP, Brodsky EE, Hough SE, Johnston MJS, Malone SD, Oppenheimer DH, Pitt AM, Richards-Dinger KB (2004) Remotely triggered seismicity on the United States west coast following the Mw 7.9 Denali Fault earthquake. *Bul Seismo Soc Am* 94:5348–5359
  75. Pyle DM, Pyle DL (1995) Bubble migration and the initiation of volcanic eruptions. *J Vol Geotherm Res* 67:227–232
  76. Reasenberg P (1985) Second-order moment of central California seismicity, 1969–1982. *J Geophys Res* 90:5479–5495
  77. Rivera L, Kanamori H (2002) Spatial heterogeneity of tectonic stress and friction in the crust. *Geophys Res Lett* 29:10.1029/2001GL013803
  78. Roeloffs E (1998) Poroelastic techniques in the study of earthquake-related hydrologic phenomena. *Adv Geophys* 37:135–195
  79. Roeloffs E, Sneed M, Galloway DL, Sorey ML, Farrar CD, Howle JF, Hughes J (2003) Water level changes induced by local and distant earthquakes at Long Valley Caldera, California. *J Vol Geotherm Res* 127:269–303
  80. Rogers G, Dragert H (2003) Episodic tremor and slip on the Cascadia subduction zone: The chatter of silent slip. *Science* 296:1679–1681
  81. Rubinstein JL, Vidale JE, Gomberg J, Bodin P, Creager KC, Malone S (2007) <http://www.nature.com/nature/journal/v448/n7153/full/nature06017.html>. Non-Volcanic Tremor Driven by Large Transient Shear Stresses. *Nature* 448:579–582
  82. Rundle JB, Turcotte D, Shcherbakov R, Klein W, Sammis C (2003) Statistical physics approach to understanding the multiscale dynamics of earthquake fault systems. *Rev Geophys* 41:4/1019
  83. Sanchez JJ, McNutt SR (2004) Intermediate-term declines in seismicity at Mt. Wrangell and Mt. Veniaminof volcanoes, Alaska, following the 3 November 2002 Mw 7.9 Denali Fault earthquake. *Bul Seismo Soc Am* 94:5370–5383
  84. Scholz CH (1998) Earthquakes and friction laws. *Nature* 391:37–42
  85. Shimomura Y, Nishimura T, Sato H (2006) Bubble growth processes in magma surrounded by an elastic medium. *J Vol Geotherm Res* 155:307–322
  86. Sibson R (2000a) A brittle failure mode plot defining conditions for high-flux flow. *Economic Geology* 95:41–48
  87. Sibson R (2000b) Fluid involvement in normal faulting. *Geodynamics* 29:469–499
  88. Sibson RH (1982) Fault zone models, heat flow, and the depth distribution of earthquakes in the continental crust of the United States. *Bul Seismo Soc Am* 72:151–163
  89. Singh SK, Anderson JG, Rodriguez M (1998) Triggered seismicity in the Valley of Mexico from major Mexican earthquakes. *Geofiscia International* 37:3–15
  90. Spudich P, Steck LK, Hellweg M, Fletcher JB, Baker LM (1995) Transient stresses at Parkfield, California, produced by the M 7.4 Landers earthquake of June 28, 1992: observations from the UPSAR dense seismograph array. *J Geophys Res* 100:675–690
  91. Stark MA, Davis SD (1996) Remotely triggered microearthquakes at The Geysers geothermal field, California. *Geophys Res Lett* 23:945–948
  92. Steacy S, Gomberg J, Cocco M (2005) Introduction to special section: Stress transfer, earthquake triggering, and time-dependent seismic hazard. *J Geophys Res.* doi:10.1029/2005JB003692
  93. Stein RS (1999) The role of stress transfer in earthquake occurrence. *Nature* 402:605–609
  94. Sturtevant B, Kanamori H, Brodsky E (1996) Seismic triggering by rectified diffusion in geothermal systems. *J Geophys Res* 101:25269–25282
  95. Tanaka S, Ohtake M, Sato H (2003) Tidal triggering of earthquakes in Japan related to the regional tectonic stress. *Earth Planets and Space* 56:511–515
  96. Tibi R, Wiens DA, Inoue H (2003) Remote triggering of deep earthquakes in the 2002 Tonga sequence. *Nature* 424:921–925
  97. Townend J, Zoback MD (2000) How faulting keeps the crust strong. *Geology* 28:399–402
  98. Ukawa M, Fujita E, Kumagai T (2002) Remote triggering of microearthquakes at the Iwo-Jima volcano. *J Geography* 111:277–286
  99. Unruh JR, Hauksson E, Monastero FC, Twiss RJ, Lewis JC (2002) Seismotectonics of the Coso Range – Indian Wells Valley region, California: Transtensional deformation along the southeastern margin of the Sierran microplate. *Geol Soc Am Mem* 195:277–294
  100. Voisin C (2002) Dynamic triggering of earthquakes: the nonlinear slip-dependent friction case. *J Geophys Res* 107(B12):10.1–10.11
  101. Weaver CS, Hill DP (1978/79) Earthquake swarms and local crustal spreading along major strike-slip faults in California. *Pageoph* 117:51–64
  102. Wen KL, Beresnev IA, Cheng S (1996) Moderate-magnitude seismicity remotely triggered in the Taiwan Region by large earthquakes around the Philippine Sea Plate. *Bul Seismo Soc Am* 86:843–847

103. West M, Sanchez JJ, McNutt SR (2005) Periodically triggered seismicity at Mount Wrangell, Alaska, after the Sumatra earthquake. *Science* 308:1144–1146
104. Yabe Y, Song S, Wang C (2005) Stress state around Chelungpu Fault, Taiwan, Estimated from boring core samples. *EOS Trans* 86:T51A–1316
105. Zoback MD, Zoback ML (2002) State of stress in the Earth's lithosphere In: Lee WH, Kanamori H, Jennings PC, Kisslinger C (eds) *International handbook of earthquake and engineering seismology*, Part A. Academic Press, Amsterdam, pp 559–568

### Books and Reviews

- Freed AM (2005) Earthquake triggering by static, dynamic, and postseismic stress transfer. *Annual Rev Earth and Planet Sci* 33:1255–1256
- Harris RA (1998) Introduction to a special section: Stress triggers, stress shadows, and implications for seismic hazards. *J Geophys Res* 103:24347–24358
- Hill DP, Pollitz F, Newhall C (2002) Earthquake-volcano interactions. *Physics Today* 55:41–47
- Hill DP, Prejean SG (2007) Dynamic triggering. In: Kanamori H (ed) *Geophysical treatise, earthquake seismology*. Elsevier, Amsterdam
- Manga M, Brodsky EE (2005) Seismic triggering of eruptions in the far field: volcanoes and geysers. *Annual Rev Earth and Planet Sci* 34:263–291
- Steady S, Gombert J, Cocco M (2005) Introduction to special section: Stress transfer, earthquake triggering, and time-dependent seismic hazard. *J Geophys Res*. doi:10.1029/2005JB003692

## Earthquakes, Electromagnetic Signals of

SEIYA UYEDA<sup>1</sup>, MASASHI KAMOGAWA<sup>2</sup>,  
TOSHIYASU NAGAO<sup>1</sup>

<sup>1</sup> Earthquake Prediction Research Center,  
Tokai University, Shizuoka, Japan

<sup>2</sup> Department of Physics, Tokyo Gakugei University,  
Koganei-shi, Japan

### Article Outline

Glossary

Definition of the Subject

Introduction

Telluric Current Anomalies and Natural Time

Ultra Low Frequency (ULF) Anomalies

Higher Frequency Electromagnetic Emission  
and Earthquake Light

Lithosphere-Atmosphere-Ionosphere (LAI) Coupling

Mechanism of Pre-Seismic EM Phenomena

Future Directions

Bibliography

### Glossary

**Earthquake prediction** Place of epicenter, time of occurrence, and magnitude are the three main items of earthquake prediction. Occurrence time is the most difficult to predict. Depending on the concerned time scales, prediction is usually classified as long term ( $\sim$  tens of years), intermediate term ( $\sim$  a few years), and short term (months to days) predictions. Electromagnetic signals of earthquakes are mainly concerned with the short term prediction.

**Piezo-electric effect** Piezo-electricity is the electric polarization produced in certain crystals and ceramics by the application of mechanical stress. Among rock-forming minerals, quartz is most strongly piezo-electric, but its effect is much reduced because quartz crystals are usually randomly oriented. Moreover, stress-induced piezo-electric polarization in rocks is kept canceled by compensating charges. At rapid stress drop, bulk polarization appears as the compensating charge cannot disappear instantly and decays with a time constant  $\tau = \epsilon/\sigma$ , where  $\epsilon$  is dielectric constant and  $\sigma$  electric conductivity.

**Electro-kinetic effect** Electro-kinetic effect, also called streaming potential, is caused by the presence of the solid-liquid interface. The double layer consists of ions (anions in most cases of rock-water system) that are

firmly anchored to the solid phase and ions of the opposite sign (cations) in the liquid phase attracted to them near the boundary. The liquid phase is in surplus of cations so that when the liquid flows due to a pressure gradient, an electric potential gradient is formed. It is expressed as  $\text{grad } V = -(\epsilon\zeta/\eta\sigma)\text{grad } P$ , where  $\epsilon$ ,  $\sigma$  and  $\eta$  are the dielectric constant, electric conductivity, and viscosity of the fluid, whereas  $\zeta$  is a constant called zeta potential. Thus, the streaming potential is small for high conductive and viscous liquid.

**Telluric current** Electric current flowing in the surface layer of the earth's crust is called telluric current. Mainly it consists of the current induced by extra-terrestrial geomagnetic field variations (called magneto-telluric or MT current) and the current as a part of the global circuit between ionosphere and ground. MT current carries information on the electrical structure of the earth's interior: higher (lower) frequency for shallower (deeper) structure. Telluric current can also be of man-made origin leaking from such electric sources as factories and trains. Telluric current is measured by dipoles of electrodes inserted into the ground at separate points. It has been postulated that transient anomalous telluric currents are observed before earthquakes.

### Frequency bands of electromagnetic waves

Electromagnetic waves are classified by frequency bands as follows:

ULF ( $<$  a few Hz), ELF (a few Hz  $\sim$  3 kHz), VLF (3–30 kHz), LF (30–300 kHz), MF (300–3000 kHz), HF (3–30 MHz), VHF (30–300 MHz), UHF (300–3000 MHz), SHF (3–30 GHz). Not only ULF to VHF bands, but also infrared ( $\sim 10^{13}$  Hz) and visible ( $\sim 10^{14}$  Hz) bands are considered to be involved in earthquake-related electromagnetic waves.

**Skin effect** The intensity of electromagnetic wave decreases exponentially with distance in a conductive medium. In a simple case, the distance where the intensity becomes  $1/e$ , called the skin depth  $\delta$ , is expressed as  $\delta = \sqrt{2/\mu\sigma\omega}$ , where  $\mu$  and  $\sigma$  are magnetic permeability, and electric conductivity of the medium and  $\omega$  is the angular frequency of the wave.

**Ionosphere** The upper atmosphere, where electrons are stripped off from oxygen and nitrogen atoms by solar radiation, is called the ionosphere. It consists of a D-layer (60–90 km), E-layer (90–130 km),  $F_1$ -layer (130–210 km), and  $F_2$ -layer (210–1000 km). Electron density is highest in the  $F_2$ -layer. The electron density of the ionospheric lower layer can be measured by ground-based ionosonde, whereas total electron con-

tent (TEC) of the whole ionosphere is estimated by global position system (GPS). Electric currents in the ionosphere produce transient variations of geomagnetic field. The suggestion has been made that the ionosphere is affected before earthquakes.

### Definition of the Subject

Throughout most of human history, electromagnetic phenomena associated with earthquakes have been repeatedly told. A typical one is earthquake light. Until rather recently, however, most records were in the realm of folklore [31,71]. Since earthquakes are understood as a catastrophic event to occur when slowly increasing tectonic stress in the earth's crust reaches a critical level, it may well be expected that the same stress may give rise to some electric, magnetic, or electromagnetic phenomena (EM phenomena hereafter) and some persistent research on them was initiated more or less simultaneously in varied parts of the world in the 1980s in two main streams. One was monitoring of possible emissions from focal regions in a wide range of frequency from DC to VHF, whereas the other was to monitor the anomalous transmission of man-made EM waves of varied frequencies over focal regions. Theoretical and experimental studies on the mechanism of EM phenomena have also been made. This relatively new branch of science is now called Seismo-Electromagnetics.

These EM phenomena attract high attention for their possible usefulness in earthquake prediction, which is of immense societal importance and considered as one of the last frontiers in earth sciences. Because many of the EM phenomena are observed prior to earthquakes, they may serve as their precursors, which have been difficult to find by usual seismological and geodetic methods. The extremely interdisciplinary nature of the subject matter is the distinct feature of Seismo-Electromagnetics and the backgrounds of many research fore-runners are neither seismology nor geodesy, but other fields, e. g., general geophysics, solid state, statistical, and ionospheric physics, radio, space, and even biological sciences. This situation in turn tends to make their accomplishments difficult to be understood and accepted by the conventional earthquake community. Of course, EM phenomena do not cause earthquakes. Both EM phenomena and earthquakes are considered to be caused by regional or local tectonic stresses, but some EM phenomena seem to appear shortly before the occurrence of earthquakes. That EM phenomena do not cause earthquakes may be another reason why few seismologists are interested in them.

### Introduction

Earthquake (EQ) related EM signals may be classified into two major groups, each covering wide frequency ranges. One is EM signals supposedly emitted from the focal zone and the other is anomalous transmission of EM waves over the epicentral region.

The emission type signals are reported for geo-electrical potential (telluric current) and geomagnetic field and for EM waves. The best known example of the former is the Seismic Electric Signals (SES) in the VAN-method which has been developed in Greece since the early 1980s and applied also in Japan since the 1990s [76,78,82]. SES are transient DC geo-electrical potential variations observed before EQs by dipoles of buried electrodes. Experimentally and theoretically, the VAN method is by far the best equipped method in this category and has survived debates [22,41]. Along with the current views that earthquakes are catastrophic events at a critical state of complex systems, a new time domain called Natural Time has been introduced to integrate SES with seismicity for short-term EQ prediction (e. g., [82]). Late in the 1980s, pre-seismic magnetic signals began to be reported. They were ultra low frequency (ULF) anomalous changes observed before *M*7.1 Loma–Prieta EQ, 1988 [16] and *M*7.1 Spitak EQ, 1989 [39] followed by *M*8.0 Guam EQ, 1993 [29]. Among them the case of Loma–Prieta is considered as most convincing. For the higher frequency range, there have been reports of VLF signals received by aerial antennas [3,23,95]. There have been reports from Greece that even VHF signals have also been received [9]. For high frequency signals, considering the electric conductivity of the crust, their emergence to the surface from the source regions at depth presents problems to be solved.

Pre-seismic appearance of EM signals makes them useful for EQ prediction. It has often been questioned, however, why they appear only pre-seismically and not co-seismically. This point makes the scientific community dubious about EQ-related EM signals in general. Actually, co-seismic signals are routinely observed but not reported each time for the obvious reason that they are not useful for EQ prediction. However, all the co-seismic signals observed so far were found to occur at the time of the arrival of seismic waves. They are, therefore, co-seismic wave signals and not “true” co-seismic. The fact that no true co-seismic signals are observed may be an important clue in exploring the physical mechanism of signal generation.

The generation mechanism of EM signal emission may be different for different frequencies. They may involve the electro-kinetic effects and pressure-stimulated polarization effects [79] for DC to low frequency signals, and

piezo-electric effects and exo-electron emission [10] for higher frequency ones. For the optical range like EQ light, a very different mechanism such as de-excitation of air molecules may be involved. Apart from these, a mechanism involving so-called positive holes (p-holes) in rock forming minerals under stress has also been proposed recently [18].

The second class of signals, i. e., the anomalous transmission of EM waves, began to be actively discussed in the late 1980s [24]. One of the best documented early cases may be at the 1995 Kobe earthquake in which received VLF radio waves for navigation purposes showed anomalies in both phase and amplitude a few days before the main shock [52]. Moreover, at the same Kobe EQ, it was first found that FM radio waves from stations beyond the line of sight can be received before main shocks [40].

The anomalous transmission of EM waves means that there are some anomalies in the path, i. e., ionosphere or atmosphere, over the epicentral region, which may be verified independently. Investigation to detect such changes has been vigorously conducted both by ionosonde [45] and by topside observation from satellites [63]. On the basis of these observations, the concept of Lithosphere-Atmosphere-Ionosphere (LAI) coupling through which pre-seismic changes in the earth's crust may be transferred to the upper atmosphere became one of the central issues of Seismo-EM studies.

This article does not deal with the aspects of EM studies of the earth devoted for elucidating the subterranean electrical structures. Although their EQ related structural time changes, if observed, would be of great interest, there has practically been no significant reported progress.

## Telluric Current Anomalies and Natural Time

### The VAN Method

The best example of modern research on DC electric signals is that of the VAN method [18,80,82], named after the initials of the founding Greek scientists, P. Varotsos, K. Alexopoulos, and K. Nomikos. The VAN group has been making actual short-term predictions of  $M \geq 5$  Greek EQs during well over a couple of decades. The criteria for successful prediction imposed by themselves are: < a few weeks in time, < 0.7 units in magnitude ( $M$ , hereafter), and < 100 km in epicentral distance. The length of time window depends on the type of signals.

The changes in the geo-electrical potential differences between buried electrodes, called Seismic Electric Signals (SES), are continuously monitored at many stations (Fig. 1a). At each station, several short (50–200 m) dipoles in both EW and NS directions and a few long dipoles (2–

20 km) in appropriate directions are installed. Compared with all earlier works using only one or two dipoles, adoption of the multiple dipole system was a distinct progress in noise rejection.

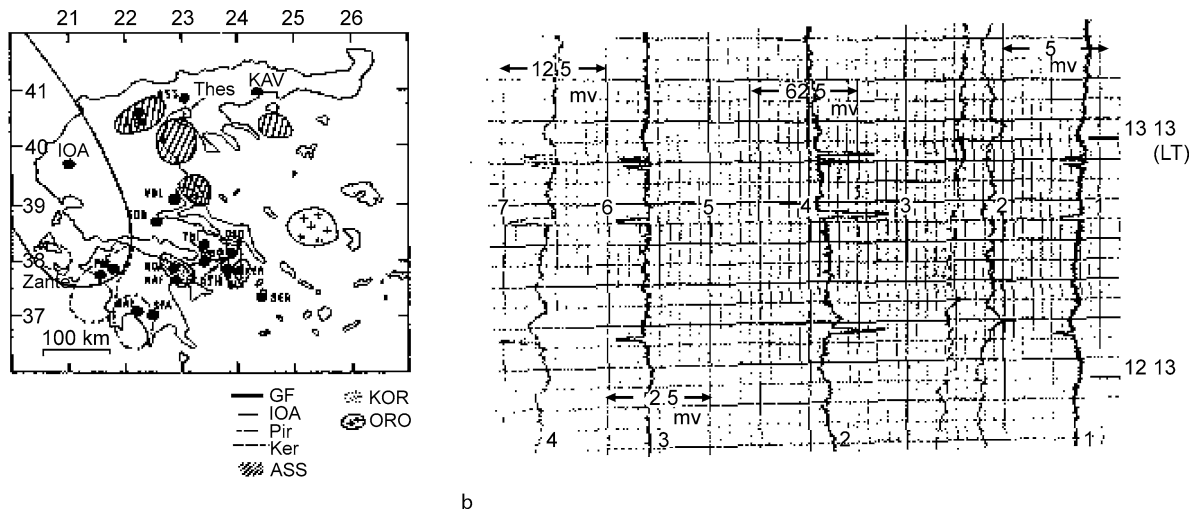
Amplitude of SES is of the order of 1 mV/100 m. There are four types of signals, i. e., single SES, SES Activity, Gradual Variation of Electric Field (GVEF), and short duration pulse. Single SES, having duration 1/2 min ~ several hours, precedes single EQ, whereas SES Activity, which consists of a number of SES in a short-time, is followed by a series of EQs before the main shock (Fig. 1b). As will be explained later, SES Activity has been playing a major role in the recent VAN work related to Natural Time analysis. GVEF has amplitude an order of magnitude larger than usual SES, but is only rarely observed for large EQ. The last type, i. e., short duration pulses appear shortly (some minutes) before EQs. These pulses, with amplitude sometimes amounting orders stronger than SES, have received rather little attention mainly because their lead time of minutes has been considered too short for useful EQ prediction.

In the VAN type of observation, noise discrimination is critically important. To eliminate noise, they have developed a set of rules as follows:

1. Changes with magneto-telluric origin can be eliminated because they appear at all the stations simultaneously.
2. SES must appear simultaneously on all of short and long dipoles, but only at the concerned station.
3. SES must satisfy the  $\Delta V/L = \text{constant}$  relation for short parallel dipoles, where  $\Delta V$  is the amplitude of SES and  $L$  the dipole length.
4. The polarity and amplitude of SES of short and long dipoles must be compatible with the assumption that the source is distant compared with the dipole lengths.

The VAN group made two major discoveries. One is the so-called "Selectivity" and the other is the so-called "VAN-relation". The Selectivity has two aspects. (1) There are only selected sites which are sensitive to SES (sensitive sites). They were found only through testing at many sites: Almost 90% of sites were insensitive. This fact gives another strong reason why earlier efforts to catch precursory electric signals failed. (2) A sensitive site is sensitive only to SES from some specific focal area(s), which are not always in close proximity. A map identifying those focal area(s), SES from which are sensed by a site, is called the "Selectivity map" of that site (Fig. 1a), which provides information on the epicentral location of the impending EQ when a SES is observed at the site. The Selectivity is considered to originate from the inhomogeneity of the subterranean electrical structures, i. e., SES goes only through conduc-





Earthquakes, Electromagnetic Signals of, Figure 1

**a** Distribution of VAN stations and “Selectivity map” of several stations as of 1996 (after [74]). For Zante, see text. **b** An example of SES Activity recorded on three short dipoles (labeled 1, 3, and 4) and long dipole (2) at Ioannina (IOA in **a**) station on August 31, 1988. Note that intensity scales in mV are different for different dipoles. The Killini–Varthelomio EQs were predicted based on these data [82]

tive channels. The VAN group has presented many model studies of channels [82]. So far, however, the real existence of such subterranean channels has not been verified by usual MT or other electric exploration techniques, possibly because the scales of the proposed channel structures are too small for the presently available resolving power.

The other discovery from the VAN research, i. e., the “VAN relation”, is the following relationship among the focal distance,  $r$ , EQ magnitude,  $M$  and the observed intensity of SES,  $\Delta V/L$ .

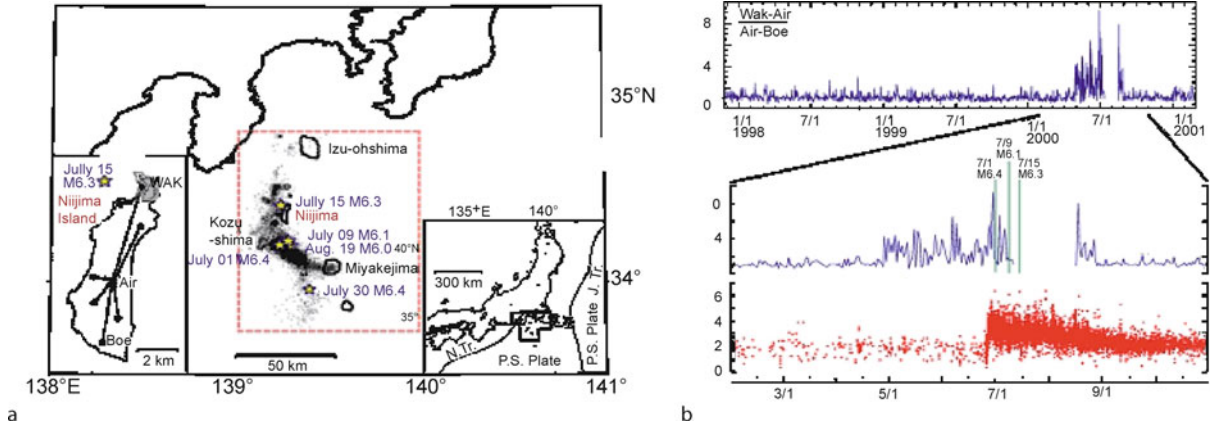
$$\log(\Delta V/L \times r) = aM + b, \quad (1)$$

where  $a$  is a constant 0.34–0.37 and  $b$  is a site-dependent constant. Once the epicentral location is estimated from the Selectivity map mentioned above,  $M$  of the impending EQ can be assessed since both  $\Delta V/L$  and  $r$  are known.

The VAN method has been a contentious subject (e. g., [22,41]). It is difficult, in principle, to prove the causal relationship between SES and EQ occurring at separate times and the only reasonable way to make the relationship credible would be to accumulate as many case studies as possible on one hand and to build plausible physical models on the other. Both endeavors have been published in papers too many to quote here and they are summarized in [82]. According to an independent evaluation (Uyeda et al. [75] and later additional check), out of 16 mb  $\geq 5.5$  EQs which occurred in the Greek region during Jan. 1, 1984–Jan. 1, 2004, 13 were successfully predicted. After the predictions of three large EQs in 1995,

there were no large EQs until another three mb  $\geq 5.5$  occurred in late 1997. It is remarkable that during the 2.5-year of quiescence no prediction was issued for the area and two out of the three 1997 events were predicted remarkably well. The score of the VAN method has been assessed by many authors. Most of the evaluations were in favor, but some were not. The low scores often resulted either when the assessors did not follow what the VAN group designated on the items such as allowable lead times for different type of SES, or the magnitude scales to use. Mulargia and Gasperini [53] claimed that “the apparent success of VAN predictions can be confidently ascribed to chance” and ignited heated debates (e. g., [22,41]). In the present authors’ view, however, VAN has well survived them. In Greece, VAN-type SES has been observed by other groups also.

In Japan, VAN-type monitoring was initiated on a trial basis in the late 1980s and was expanded in 1996 [76]. Despite the serious problems caused by the high level of artificial noise, in particular from DC-driven electric trains, the existence of the VAN-type SES has been confirmed for  $M > 5$  EQs occurring within  $\sim 20$  km or so of a station. Moreover, phenomena attesting to “Selectivity” were discovered. In the year 2000, a two-month long seismic swarm, with  $\sim 7,000$   $M \geq 3$  shocks and five  $M \geq 6$  shocks, occurred in Izu Island region: See Fig. 2a. For this swarm activity, significant pre-seismic electric disturbances were observed [77]. From about 2 months before the swarm onset on June 26, innumerable clear, un-



Earthquakes, Electromagnetic Signals of, Figure 2

**a** Seismic swarm activity in 2000 in Izu island region. *Inset in the left* shows the dipole configuration in Niijima Island. Each end of the long dipole had short dipoles. Only Wak-Air dipole showed the pre-swarm signals shown in **b**. The *bottom panel* in **b** shows seismicity (modified from [77])

usual geo-electrical potential changes started on Niijima Island (Fig. 2). These anomalous changes appeared only in the northern part of the island, possibly reflecting the extremely heterogeneous underground structure of the volcanic region.

Co-seismic signals have been observed for many EQs. However, they always started with the arrival of seismic waves and not at the origin time of EQs. The changes are probably local effects of passing seismic waves. There may be many reasons why no true co-seismic signals are observed. One is that, as laboratory experiments show, signals generated at ruptures are in much higher frequency range, so that they cannot be registered by usual high-cut measurement (0.1–1 Hz sampling) and the second is that, even when a higher sampling rate is employed, the high frequency signals attenuate before reaching the receiver. In fact, the pre-seismic stress accumulating process giving rise to SES and the instantaneous stress releasing event are physically very different processes and there seems to be no compelling reason why they generate similar signals.

### Natural Time

Seismicity as a critical phenomenon has been actively discussed by many authors (e. g., [4,38,65,68,73]). It has been shown that SES and EQs reveal dynamic evolution characteristics to the critical stage when their time series is analyzed in the framework of natural time  $\chi$ , which was introduced by the Varotsos' group (e. g. [82,83]). The symbol  $\chi$  stands for the ancient Greek word  $\chi\rho\omicron\nu\omicron\sigma$ , which means "time". The possible usefulness of natural time analysis in predicting catastrophic events has been demonstrated

not only for the subjects of our immediate concern, but also for other critical phenomena, including sudden cardiac death [84,85].

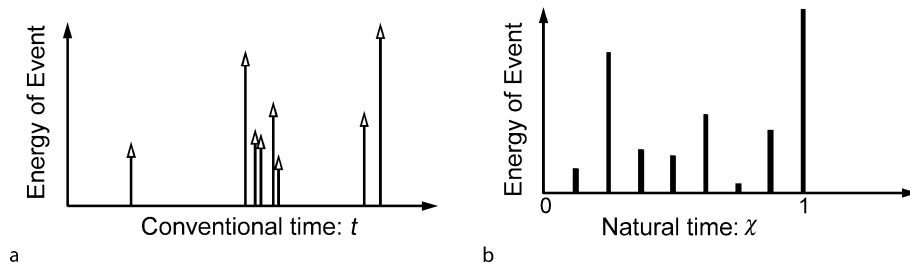
In a time series comprised of  $N$  events, the natural time  $\chi_k = k/N$  serves as the index for the occurrence of the  $k$ th event. In natural time analysis, the evolution of the pair of two quantities ( $\chi_k, Q_k$ ) is investigated where  $Q_k$  denotes the quantity proportional to the energy of the  $k$ th event. The time series of events as shown in Fig. 3a is expressed in natural time as in Fig. 3b. For the purpose of analysis, the following function  $\Phi(\omega)$  was introduced.

$$\Phi(\omega) = \sum_{k=1}^N p_k e^{i\omega \frac{k}{N}}, \quad (2)$$

where  $p_k = Q_k / \sum_{n=1}^N Q_n$  and  $\omega = 2\pi\phi$ ,  $\phi$  standing for the frequency in natural time (natural frequency). This  $\Phi(\omega)$  should not be confused with the discrete Fourier Transform because  $\omega$  is a continuous variable. If we regard  $p_k$  as the probability density function of  $\chi$ , in analogy with probability theory, its Fourier transform  $\Phi(\omega)$  may be regarded as the characteristic function of  $\omega$ . The power spectrum of  $\Phi(\omega)$ ,  $\Pi(\omega) = |\Phi(\omega)|^2$ , for the dynamical system approaching critical state with infinitely long-range temporal correlation was calculated to be as follows (see p.259–260 in [82]):

$$\Pi(\omega) = |\Phi(\omega)|^2 = \frac{18}{5\omega^2} - \frac{6 \cos \omega}{2\omega^2} - \frac{12 \sin \omega}{5\omega^3}. \quad (3)$$

Taylor expansion of Eq. (3) gives, for small values of  $\omega$  ( $\omega \rightarrow 0$ ),



Earthquakes, Electromagnetic Signals of, Figure 3  
Time series of events, **a** in conventional time  $t$ , and **b** in the natural time  $\chi$

$$\begin{aligned} \Pi(\omega) &= 1 - \kappa_1\omega^2 + \kappa_2\omega^4 + \kappa_3\omega^6 + \kappa_4\omega^8 + \dots \\ &= 1 - 0.07\omega^2 + \dots \end{aligned} \quad (4)$$

Thus, a time series should show  $\kappa_1 \approx 0.07$  when approaching the critical stage. The reason why the natural time domain is useful when information on intervals between events is lost while retaining only information on the order and relative importance of events is an intriguing question. As to this point, Abe et al. [1] have shown that this time domain in fact is optimal for enhancing the signals in time-frequency space.

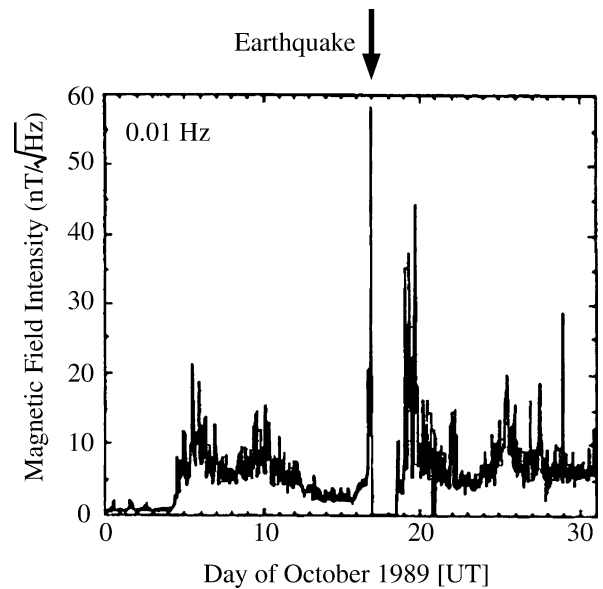
In Greece,  $\kappa_1 \approx 0.07$  was experimentally ascertained first for SES activities preceding four large EQs; 1995  $M6.6$  Kozani–Grevena EQ, 1995  $M6.5$  Eratini–Egio EQ, 1997  $M6.4$  Strofades EQ, 2001  $M6.6$  Aegean Sea EQ and later also for other major EQs, supporting this view [82,86]. Infinitely long-range temporal correlations of Greek SES were independently confirmed by Weron et al. [89].

In the case of seismicity, to investigate its time evolution, the power spectrum  $\Pi(\omega)$  of the seismicity in natural time subsequent to associated SES activity was calculated as each consecutive EQ occurred. It was, then, shown that, for major Greek EQs,  $\Pi(\omega)$  approached that of the critical state ( $\kappa_1 = 0.07$ ) a few days before the main shocks [82]. This indicated that the seismicity approached the critical state at that time. This unexpected discovery may shed new light on the EQ generation mechanism itself. At the same time, this suggests the possibility of narrowing the time window of predicting EQs to a few days, when SES data are available. It may be added, albeit different from the Natural Time defined here, that an attempt was made to identify seismic quiescence with the viewpoint that the seismic process proceeds with its internal clock called “events time scale” [66].

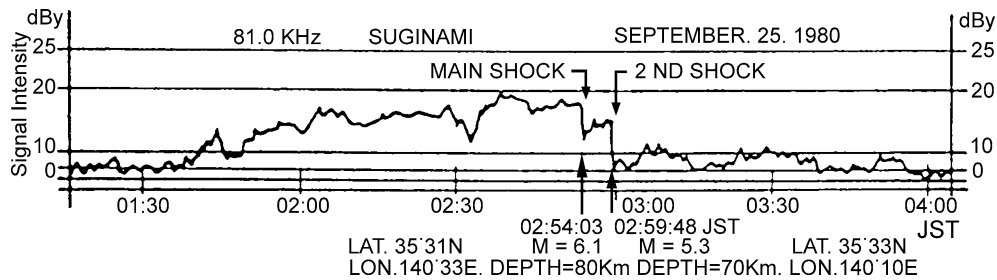
### Ultra Low Frequency (ULF) Anomalies

ULF generally means lower than several Hz. Research in this frequency range was started late in the 1980s. ULF

signals are advantageous over those in higher frequencies because of their large skin depth. The best-known example is the case of the  $M7.1$  Loma-Prieta (California) EQ in 1989 [16]. Observation was made at a site which happened to be at 7 km from the epicenter. The amplitude of the horizontal component started anomalous enhancement at about 2 weeks prior to and a sharp increase a few hours before the EQ. Figure 4 shows the records at 0.01 Hz band. The disturbance lasted for about 3 months after the EQ. These anomalous changes were not of solar terrestrial origins because they were not observed at other distant stations. Moreover, these have never been observed at any other time during the whole period of observation of more than 15 years. It was, thus, concluded that the anomalies were related to the EQ. Reports of observing



Earthquakes, Electromagnetic Signals of, Figure 4  
The amplitude of the geomagnetic horizontal component at 0.01 Hz band [16]



Earthquakes, Electromagnetic Signals of, Figure 5  
Change of 81 kHz electromagnetic wave observed in Tokyo [23]

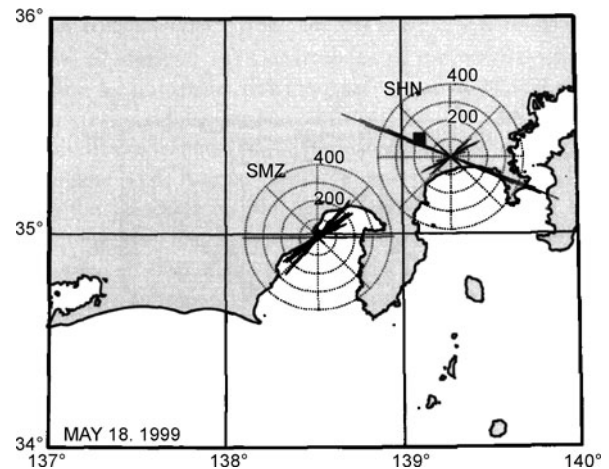
pre-seismic ULF geomagnetic anomalies have been made also for  $M6.9$  Spitak (Armenia) EQ in 1988 [39] and  $M8.0$  Guam (Marianas) EQ in 1993 [29]. Further efforts in Japan and elsewhere have been summarized by Hattori [28]. It seems, however, that a more rigorous approach is needed to make the ULF studies sufficiently credible to the scientific community.

### Higher Frequency Electromagnetic Emission and Earthquake Light

Pre-seismic electromagnetic wave emission in the VLF–LF-range has been reported since the 1980s. Gokhberg et al. [23] reported pioneering observations as shown in Fig. 5. Emissions at 81 kHz increased one or two hours before  $M6.1$  and  $M5.3$  earthquakes took place and decreased after the second shock.

Asada and his group started investigation of EQ-related VLF emissions in the early 1990s [3]. They monitored the wave forms of two horizontal magnetic components of VLF waves, through which the apparent incoming direction of VLF pulses was determined. They found that, before  $M5$  class land EQs within 100 km of their stations, some pulses with a fixed incoming direction appeared and the EQs actually occurred in that direction, whereas the sources of overwhelmingly numerous and stronger noises were moving along with lightning sources (Fig. 6). Moreover, there is a well-documented report of undeniable noise in commercial MF radio bands, experienced by an automobile driver approaching Kobe, some minutes before the Kobe EQ of 1995 (see [55]).

Enomoto et al. [12] recorded anomalous pulses of geo-electrical current (HF-band) at Erimo station, Hokkaido, Japan, from February 2000 to March 2001 and from August to September 2003. The former anomalies occurred before and during the volcanic activity of Mt. Usu (200 km away), while the latter started one month before the 2003 September 26  $M8.0$  Tokachi-Oki EQ (80 km away). These



Earthquakes, Electromagnetic Signals of, Figure 6  
Rose diagram of incoming VLF signals observed on May 18, 1999 at two sites.  $M4.1$  EQ occurred at black square point on May 22 [3]

were the only anomalies during their 10-year observation period.

For the Kobe EQ, while measuring sporadic Jovian decametric emissions with a radio interferometer at an observatory at about 80 km from the epicenter, unusual pulsed emissions at 22.2 MHz were detected tens of minutes both before and after the main shock [47]. Such unusual pulses have never been observed at other times and the possible source direction was estimated to be that of the main surface exposure of the EQ fault. There was no clear co-seismic radiation. Warwick et al. [88] reported a similar observation related to the 1960 Great  $M9.5$  Chilean EQ.

Also in the high frequency range, Eftaxias et al. [9] have reported results obtained in Greece. Since 1994, they have been running a station on Zante Island in the Ionian Sea (see Fig. 1a), where installation was performed of (i) six loop antennas in EW, NS and vertical magnetic field at both 3 kHz and 10 kHz; (ii)  $\lambda/2$  electric dipoles for 41,

54, and 135 MHz and (iii) two Short Thin Wire Antennas for ULF ( $< 1$  Hz) anomalies. Sampling rate is 1 Hz. They report that MHz–kHz EM anomalies have been detected during a few days to a few hours prior to near-surface land EQ with  $M > 6$ , i. e., the 1995 Kozani-Grevena EQ and 1999 Athens EQ. The MHz radiation appeared earlier than the kHz. They interpret the phenomena as due to small-scale cracking and assume that the more grown-up cracks generate the lower frequency anomalies. They also look to the EQ as a critical phenomenon and suggest that the shift from MHz to kHz activity corresponds to an anti-persistence to persistence shift. In their observations, there was no co-seismic anomaly.

As mentioned earlier, however, transmission of EM waves in the conducting earth beyond the skin depth distance is an important unresolved problem common to all the topics reported in this section.

### EQ Light

Earthquake light (mostly co-seismic) has been reported all over the world from ancient Greek, Roman, and Chinese times. There is no doubt that the phenomena exist. Light emanates from the whole sky, or locally from the ground. The shape reported is like aurora, pole, flash, ball lightning, and so on, while the color widely ranged from blue and blue-white to red-yellow and orange. We, however, note that all reports may not be on natural phenomena but on some artificial effects such as sparks at power lines.

Galli [21] collected 148 eyewitness reports in late 19th Century Europe (see [71]). In Japan, Musha [54] collected about 2,000 eyewitness reports for 65 EQs, while Terada [69] discussed the theoretical aspects and suggested that the electro-kinetic effect may be a possible cause. From 1965 to 1967, there was a large EQ swarm at Matsushiro area in central Japan and numerous luminous phenomena were seen and photographs were taken as shown in Fig. 7 [91].

For the 1995,  $M7.3$  Kobe EQ, Enomoto and Zheng [11] examined the trace of gas emission in the Awaji fault where the rupture started. They suggested that the gas plasma emission might have emitted the light. Kamogawa et al. [37] reported some independent witnesses that a luminous object moved a long distance a few seconds before the main shock in the direction of the rupture. Ikeya and Takaki [32] numerically showed that the screening charges neutralizing the polarized piezoelectrical rock may generate a strong co-seismic electric field, and the de-excitation of nitrogen molecules excited by collision of electrons accelerated by the electric field produce the blue EQ-light.



Earthquakes, Electromagnetic Signals of, Figure 7  
Photograph of EQ light at Matsushiro seismic swarm taken by Kuribayashi (after [91])

### Lithosphere-Atmosphere-Ionosphere (LAI) Coupling

Pre-seismic atmospheric-ionospheric anomalies before EQs have been reported since the 1970s [2,20,24,27] and the concept of pre-seismic lithosphere-atmosphere-ionosphere coupling arose. Historical reviews and important works that are not introduced in this article may be tracked from references of Pulinet and Boyarchuk [61] and Kamogawa [33].

Liu et al. [43] found in Taiwan that the ionosonde measured critical plasma frequency, foF2, corresponding to the electron density of the ionospheric F2 layer, significantly decreased during afternoons within a few days before  $M \geq 6$  EQs. For example, such ionospheric anomalies appeared 3 and 4 days before the 1999  $M7.7$  Chi-Chi EQ. Similar EQ-related electron density depression occurring above Taiwan Island was observed in the GPS total electron density (TEC) [44]. From such observations, Liu et al. [45] demonstrated that the appearance of the anomalies within 5 days was statistically significant at 5% level for the  $M \geq 5.4$  EQs occurring within 150 km.

Sub-ionospheric anomalies before large EQs were reported by Gokhberg et al. [24] and Gufeld et al. [27]. They used VLF ship-navigation waves (10–20 kHz) and observed pre-seismic anomalies between the transmitter and the receiver during mid-night. Marenko [49] statistically supported the results of Gokhberg et al. [24], while Michael [50] obtained a less optimistic conclusion. Meanwhile, the studies have been further developed mainly in Russia, Japan, and Italy. For example, pre-seismic variations of terminator-times, i. e. the sunrise and sunset for VLF waves, were demonstrated [52]. Clilverd et al. [6], on the other hand, did not obtain similar positive results when they applied the terminator-time method to their 5-year data of reception at Faraday, Antarctica (receiver) of VLF waves transmitted from the northern United States. Maekawa et al. [48], measuring LF waves,

statistically investigated the correlation between sub-ionospheric anomalies and  $M \geq 6$  EQs in Japan and found that the amplitude and dispersion of received signals significantly decreased 2–6 days before the EQs. Thus, this issue is still controversial [34,64].

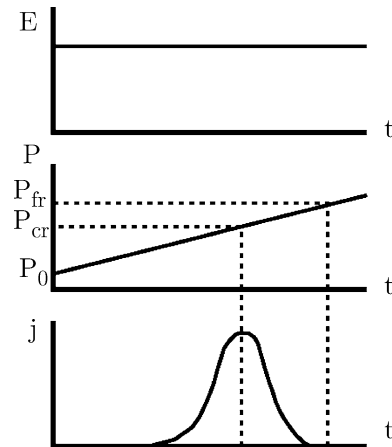
In the VHF range, Kushida and Kushida [40], while monitoring meteorites plunging into the high atmosphere by reflection of FM radio waves, detected anomalous reception, just a few days before the Kobe EQ, of the FM radio waves from distant (beyond the line-of-sight) stations. This was a new discovery and these authors consequently began extensive measurements on other EQs. With regard to the anomalous reception of VHF waves from transmitters beyond the line-of-sight, Fujiwara et al. [20] statistically showed significant enhancement of atmospheric anomalies lasting for a few minutes–several hours within 5 days before  $M \geq 4.8$  EQs.

## Mechanism of Pre-Seismic EM Phenomena

### Generation Mechanism of EM Signals

**Electro-Kinetic Effect** The electro-kinetic effect can be a plausible source for SES (DC) and ULF emission. Mizutani et al. [51] first proposed a model in which, during the dilatancy stage, pore pressure in the dilatant region decreases and water flows into this region from the surrounding area, generating electric and magnetic precursors of EQs. Since then, many models have been proposed (e. g., [15,93]). Fedorov et al. [13], however, suggested that the expected magnitude of seismo-EM signals in the ULF-VAN range from an electro-kinetic source may reach the detection level only for a favorable set of crustal parameters.

**Models Related to Defects in Solids** A SES-generation model by pressure-stimulated currents (PSC) was proposed by Varotsos and Alexopoulos [79]. Their model is based on the physics of the point defects in solids. The impurities and vacancies have excessive and opposite-sign effective charges and form local electric dipoles. The directions of the local electric dipoles usually distribute randomly. Under an electric field, dipoles will align in its direction. The alignment is an activation process in which the time constant is an Arrhenius-type function of stress as well as temperature. Therefore, an avalanche of alignment takes place when stress approaches a critical level (Fig. 8). It has been later suggested that, instead of electric field, inhomogeneous deformation mentioned in the next paragraph may work to align the dipoles in the direction of the stress gradient (see [14]). This model is unique among other models in that SES is generated spontaneously dur-

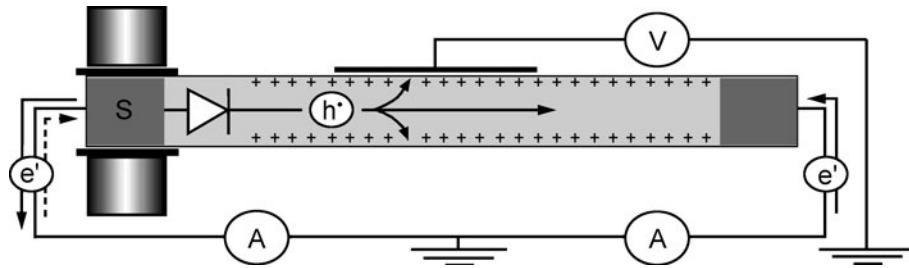


Earthquakes, Electromagnetic Signals of, Figure 8  
 Pressure-stimulated current  $j$  occurs at a critical pressure  $P_{cr}$  under the external electric field  $E$ .  $P_{fr}$  is the fracture pressure (after [79])

ing gradual increase of stress without requiring any sudden change of stress such as micro-fracturing. For the SES to work as a precursor, it is assumed that the critical level of stress for SES generation is lower than that of mechanical failure causing EQ. A thorough verification of the PSC model by laboratory pressure experiments is fatally lacking up to this stage.

In relation to SES generation, deformation-induced charged flow is an interesting possibility [56]. This flow was observed to take place as a result of inhomogeneous plastic deformation of ionic crystals, such as NaCl, in the direction of the stress gradient without applying electric field. It was interpreted that charge carriers are charged dislocations. Some experiments were conducted on rocks with similar results (see [82]). Independent of these, Freund and his colleagues have recently been proposing a unique mechanism for ULF electric signals ([18] and Ref. therein). They have discovered in the laboratory that when a block of igneous rock is put under stress locally, the rock turns into a battery without any external electric field (Fig. 9).

This striking phenomenon is interpreted as follows: A fraction of the oxygen anions in the rock-forming silicate minerals is not in their usual 2-valence state ( $O^{2-}$ ) but in the 1-valence state ( $O^{1-}$ ), which represent defect electrons, i. e., positive holes ( $p$ -holes). They are unstable and form more stable positive hole pairs (PHP), chemically equivalent to peroxy links,  $O_3X/\text{ox}\backslash XO_3$ , which are electrically inactive. These dormant PHPs, however, are awoken by deviatoric stress, and make the insulating host material a  $p$ -type semi-conductor. The  $p$ -holes flow out of



Earthquakes, Electromagnetic Signals of, Figure 9

Conceptual diagram of the battery current, carried by electrons which flow out of the stressed portion *S* (left) through the outer circuit and by *p*-holes which close the circuit by flowing through the unstressed portion and meeting the electrons at the far end (right), flowing out from the stressed portion *S*, (after [18])

the stressed volume because of mutual electrostatic repulsion. If scaling to earthquake size is allowed, the current thus produced may attain  $10^3$ – $10^5$  A/km<sup>3</sup>.

**Other Models** For generation of high frequency signals, models related to micro-cracking have been proposed from laboratory experiments. They are (1) Discharge of screening charge of piezo-electric polarization [31,94], (2) electrification of fresh crack surfaces [90], (3) exo-electron [10]. As to the occurrence of pre-main shock micro-cracking, there have been only a few reliable field reports. Furthermore, it might be pointed out that in these models much stronger co-seismic signals would be expected. Some ad hoc mechanism, therefore, would be needed to explain that no co-seismic signals have been observed so far in the field.

### Transmission Mechanism of EM Signals

Even if EM signals are generated around a seismic focal region, signals except in the ULF range cannot be transmitted long distance in the conductive crust due to the decay caused by the skin effect, as long as the displacement current component is negligible. Even for DC signals, geometric decay would prohibit their long distance reception in a homogeneous or simple layered earth [5]. To overcome this difficulty, Varotsos et al. [81] proposed a conductive channel model, in which electric signals are transmitted through the conductive channel to a surface point close to the upper end of the channel. Freund [17] reported that, in the laboratory experiment, mobile positively charged holes (*p*-holes) appeared on the rock surface when a stress gradient was given to the rock sample. His results implied the possibility of appearance of a positively charged area at long distance on the surface before EQs. Kamogawa and Ohtsuki [35] proposed a model to explain how the higher frequency EM waves can be observed be-

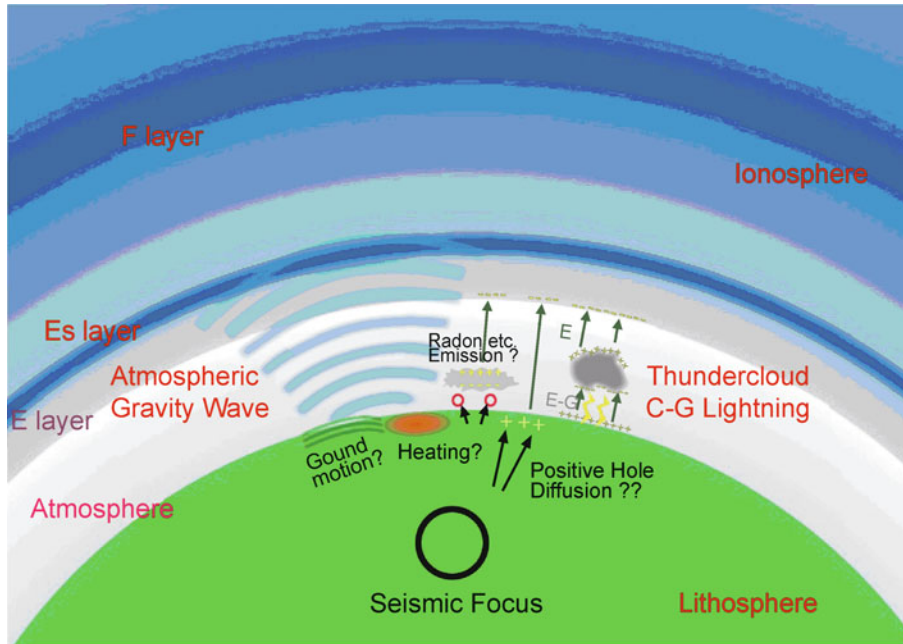
fore EQs, i. e., longitudinal plasma waves excited by exo-electrons [10] may be transformed into EM waves by the surface roughness.

### LAI Coupling Mechanism

If the pre-seismic atmospheric-ionospheric anomalies are real, some causative factors may be detected on the ground surface. Possible mechanisms of pre-seismic lithosphere-atmosphere-ionosphere coupling have been proposed by many researchers. They may be categorized in two groups as shown in Fig. 10.

First, some atmospheric electric field  $\vec{E}$  is generated on/near the ground surface during the pre-seismic period and it will cause the ionospheric anomalies [25,26,62]. Pulinets et al. [62] proposed that such an atmospheric electric field is caused by radon emission (see [30,87,92]). Alternatively, it is proposed that positively charged holes diffused from the seismic focal area to the ground surface generate the electric field [17]. However, such an electric field on the ground has not yet been observed even when pre-seismic ionospheric anomalies were detected [36].

Second, some researchers proposed that atmospheric gravity waves (AGW) propagate into the ionosphere, and disturb it before EQs [46,52,59]. The proposed source of AGW is the long-period ground oscillation or appearance of thermal anomalies on the ground. The former was inferred from some observations that co-seismic ground vibration actually excited AGW which propagated into the ionosphere (e.g. [8]). However, there is no report that long-period ground oscillations were detected at the pre-seismic stage even by high-sensitive superconducting gravimeter observations so far. The latter proposed source of AGW is the pre-seismic temperature rise, “thermal anomalies”, reaching 2–4°C or higher in a wide area around impending EQs, based on satellite observation of enhanced infrared (IR) emission from the ground sur-



Earthquakes, Electromagnetic Signals of, Figure 10  
Concept of LAI coupling (modified from [33])

face [7,70,72]. Many models have been put forward to explain the origin of the “thermal anomalies”, including latent heat release at condensing water vapor due to enhanced radon emission. Pulinets [60] develops a scenario where the “thermal anomalies” give rise to ionospheric anomalies. Freund et al. [19] cast another interpretation on the enhanced IR emission based on their *p*-hole model mentioned above. When *p*-holes appear on the surface of the unstressed area, they form a positive charge layer and recombine to form the more stable  $O^--O^-$  bond, emitting IR as de-excitation energy.

### Future Directions

It seems that, despite much circumstantial evidence, earthquake related electromagnetic signals, in particular those at the pre-seismic stage, have not yet been completely accepted as real physical quantities. Putting the common indifference and prejudice of the conventional scientific community against new science aside, it seems appropriate at this stage to recognize that there are legitimate reasons for the critical views. In fact, most of the problems of fundamental importance in seismo-electromagnetics are still unresolved.

To name a few, propagation of high-frequency EM signals in conductive earth has been proven unequivocally enough neither empirically nor theoretically. Techniques

of direction finding of EM signals at various frequency ranges and atmospheric-ionospheric anomalies have not shown sufficiently credible results yet. Solving these problems will be important issues in near future investigation.

The mechanisms of signal generation are still far from established. The majority of proposed mechanisms attribute pre-seismic signals to effects such as piezo-electric, electro-kinetic, charged dislocations, *p*-holes, and exoelectrons, all induced by stress release at micro-fracturing in the last stage of EQ preparation. In such cases, critics demand that by far the largest signals should be observed at the instance of the main shocks when the largest stress drop takes place. However, as described above, all the “co-seismically” observed electric or ULF signals are associated with the arrivals of seismic waves (to be called co-seismic wave) and are not co-seismic in the true sense. For higher frequency signals, even co-seismic wave signals have not been confirmed. This fact is a popular basis for negating the EQ-related signals in general. However, this very fact, i. e., the non-observation of true co-seismic signals in any frequency range may present some important clues with regard to the mechanism of both the signal generation and earthquakes as follows.

Numerous lab-experiments show strong co-fracturing signals in the form of high frequency EM waves. They are very different from the low frequency signals observed during pre-seismic stages. Thus, one explanation for non-



observation of co-seismic signals in the ULF range is (1) field observation uses a low-pass recording system to avoid high frequency noise, and (2) high frequency signals are attenuated in short distance in the earth. However, this explanation seems to suffer from a weak point as follows: Since it takes some seconds for the fault motion of a large EQ to terminate, the overall signals should contain a low frequency component powerful enough to be caught by the low-pass recording system. Moreover, even higher frequency wave monitoring systems have not captured any co-seismic signals. All these seem to speak for non-generation of co-seismic signals of any frequency in the field and researchers had to devise some ad hoc scenarios as to how to reconcile with lab results, often invoking over-growth of micro-faults by the time of main shocks to produce signals. For pre-seismic signal generation, these suggested mechanisms assume pre-seismic micro-fractures, which in fact, are micro-EQs that may be observed by high sensitivity seismic networks. Although depending on the required size, there has been no such observation, which constitutes another objection to pre-seismic EM signals.

The pressure-stimulated currents cited above regards the SES generation as a critical phenomenon. SES is supposed to be spontaneously generated when the gradually increasing tectonic stress level reaches a critical value. This seems to be the only proposed mechanism which needs no stress release by micro-fracturing or any special events, although it makes the causal relationship between SES and EQ less apparent. For the same reason, this mechanism does not need to generate any strong signals at EQ itself. Some of the other mechanisms, such as electro-kinetic, deformation-induced charged flow or *p*-holes flow mentioned earlier might be modified to fit the observation by incorporating the concept of critical state since they only need a development of stress gradient for signal generation.

Experimental verification of these mechanisms is urgently needed as it has been decisively inadequate. In any case, it should be kept in mind that the EQ preparation and EQ itself are different physical processes, the former being a gradual stress increasing process, whereas the latter is an instantaneous stress drop.

No true co-seismic signal in contrast to lab fracture experiments presents a question if the EQ is a fracture or not. It is now well-known that the EQ is a sudden sliding of pre-existing faults. However, according to Yoshida et al. [94], even stick-slip experiments on dry granite revealed strong signals at the time of slip which were understood as due to the piezo-electric effect. Non-generation of co-seismic EM signals still remains an important problem requiring further investigation.

Important unsolved questions are by no means confined to pre-seismic signal emissions. On the contrary, even more fundamental unsolved problems lie in LAI coupling. Here, the very origin on the ground to cause any of the suggested elementary agents, such as anomalous atmospheric electric field, atmospheric gravity wave, and thermal anomaly, is unknown. So far, observations on these features have been carried out by various researchers independently, so that the integration of fragmentary results for constructing a unified physical scenario of the whole process has been difficult. Since very active multi-national as well as multi-disciplinary cooperative research has been underway recently, involving GPS-TEC estimation and even topside measurement of the ionosphere by satellites of several nations, substantial progress in the upper end of LAI coupling is expected in the near future. Lately, active pursuit of EQ-related ionospheric anomalies has been made through topside and in-situ observation by satellites such as the French micro-satellite DEMETER [58]. However, the lower initiating side of the LAI coupling appears much more difficult to elucidate. It would require long sustained pre-seismic ground-based network observations on such phenomena as long-period ground motion and radon emission in as many earthquake prone areas as possible. But these tedious efforts should be enhanced on a global scale at all cost. Finally, it may be added that Kamogawa [33] pointed out that reported atmospheric-ionospheric anomalies might be caused by some EM phenomena which also trigger seismicity. For instance, suggestions have been made that geomagnetic storms [67] and cloud-to-ground lightning [42,57] may trigger EQs. It may be worthwhile to keep such a possibility of difference in cause and effect in mind in the future studies.

## Bibliography

1. Abe S, Sarlis NV, Skordas ES, Tanaka HK, Varotsos PA (2005) Origin of the usefulness of the natural-time representation of complex time series. *Phys Rev Lett* 94:170601
2. Antsevich MG (1971) The influence of Tashkent earthquake on the Earth's magnetic field and the ionosphere. In: Tashkent earthquake 26 April 1966. FAN, Tashkent, pp 187–188 (in Russian)
3. Asada T, Baba H, Kawazoe K, Sugiura M (2001) An attempt to delineate very low frequency electromagnetic signals associated with earthquakes. *Earth Planets Space* 53:55–62
4. Bak P, Tang C (1989) Earthquakes as a self-organized critical phenomenon. *J Geophys Res* 94(15):15637–15639
5. Bernard P (1992) Plausibility of long distance electrotelluric precursors to earthquakes. *J Geophys Res* 97:17531–17536
6. Clilverd MA, Rodger CJ, Thomson NR (1999) Investigating seismo-ionospheric effects on a long subionospheric path. *J Geophys Res* 104(A12):28171–28179

7. Dey S, Singh RP (2003) Surface latent heat flux as an earthquake precursor. *Nat Haz Earth Syst Sci* 3:749–755
8. Ducic V, Artru J, Lognonne P (2003) Ionospheric remote sensing of the Denali Earthquake Rayleigh surface waves. *Geophys Res Lett* 30(18):1951. doi:10.1029/2003GL017812
9. Eftaxias K, Kapiris P, Polygiannakis J, Peratzakis A, Kopanas J, Antonopoulos G, Rigas D (2002) Experience of short term earthquake precursors with VLF-VHF electromagnetic emissions. *Nat Hazards Earth Syst Sci* 20:1–12
10. Enomoto Y, Hashimoto H (1990) Emission of charged particles from indentation fracture of rocks. *Nature* 346:641–643
11. Enomoto Y, Zheng Z (1998) Possible evidences of earthquake lightning accompanying the 1995 Kobe earthquake inferred from the Nojima fault gouge. *Geophys Res Lett* 25:2721–2724
12. Enomoto Y, Hashimoto H, Shirai N, Murakami Y, Mogi T, Takada M, Kasahara M (2006) Anomalous geoelectric signals possibly related to the 2000 Mt. Usu eruption and 2003 Tokachi-oki earthquake. *Phys Chem Earth* 31:319–324
13. Fedorov E, Pilipenko V, Uyeda S (2001) Electric and magnetic fields generated by electrokinetic processes in a conductive crust. *Phys Chem Earth C* 26:793–799
14. Fischbach DB, Nowick AS (1958) Some transient electrical effects of plastic deformation in NaCl crystals. *J Phys Chem Solids* 5:302–315
15. Fitterman DV (1978) Electrokinetic and magnetic anomalies associated with dilatant regions in a layered earth. *J Geophys Res* 83:5923–5928
16. Fraser-Smith AC, Bernardi A, McGill PR, Ladd ME, Helliwell RA, Villard OG Jr (1990) Low-frequency magnetic field measurements near the epicenter of the Ms 7.1 Loma Prieta earthquake. *Geophys Res Lett* 17:1465–1468
17. Freund F (2000) Time-resolved study of charge generation and propagation in igneous rocks. *J Geophys Res* 105:11001–11019
18. Freund F, Takeuchi A, Lau BES (2006) Electric currents streaming out of stressed igneous rocks – a step towards understanding pre-earthquake low frequency EM emissions. *Phys Chem Earth* 31:389–396
19. Freund FT, Takeuchi A, Lau BWS, Al-Manaseer A, Fu CC, Bryant NA, Ouzounov D (2007) Stimulated infrared emission from rocks: Assessing a stress indicator. *eEarth* 2:7–16
20. Fujiwara H, Kamogawa M, Ikeda M, Liu JY, Sakata H, Chen YI, Ofuruton H, Muramatsu S, Chuo YJ, Ohtsuki YH (2004) Atmospheric anomalies observed during earthquake occurrences. *Geophys Res Lett* 31:L17110. doi:10.1029/2004GL019865
21. Galli I (1910) Raccolta e classificazione de fenomeni luminosi osservati nei terremoti. *Bull Soc Sis Ital* 14:221–447 (in Italian)
22. Geller R (ed) (1996) Debate on “VAN”. *Geophys Res Lett* 23:1291–1452
23. Gokhberg MB, Morgounov VA, Yoshino T, Tomizawa I (1982) Experimental measurement of electromagnetic emissions possibly related to earthquakes in Japan. *J Geophys Res* 87(B9):7824–7828
24. Gokhberg MB, Gufeld IL, Rozhnoy AA, Marenko VF, Yampolsky VS, Ponomarev EA (1989) Study of seismic influence on the ionosphere by super long-wave probing of the Earth ionosphere wave-guide. *Phys Earth Planet Inter* 57:64–67
25. Gokhberg MB, Morgounov VA, Pokhotelov OA (1995) Earthquake prediction, seismo-electromagnetic phenomena. Gordon and Breach, Reading, p 289
26. Grimalsky VV, Hayakawa M, Ivchenko VN, Rapoport YG, Zadorozhnyi VI (2003) Penetration of an electrostatic field from the lithosphere into the ionosphere and its effect on the D-region before earthquakes. *J Atmos Solar-Terr Phys* 65:391–407
27. Gufeld IL, Rozhnoi AA, Tyumensev SN, Sherstuk SV, Yampolsky VS (1992) Radiowave disturbances in period to Rudber and Rachinsk earthquakes. *Phys Solid Earth* 28:267–270
28. Hattori K (2004) ULF geomagnetic changes associated with large earthquakes. *Terr Atmos Ocean Sci* 15:329–360
29. Hayakawa M, Kawate R, Molchanov OA, Yumoto K (1996) Results of ultra-low frequency magnetic field measurements during Guam earthquake of 8 August 1993. *Geophys Res Lett* 23:241–244
30. Igarashi G, Saeki S, Takahata N, Sumikawa K, Tasaka S, Sasaki Y, Takahashi M, Sano Y (1995) Ground-water radon anomaly before the Kobe earthquake in Japan. *Science* 269:60–61
31. Ikeya M (2004) Earthquakes and Animals. World Scientific, Singapore, 294 pp
32. Ikeya M, Takaki S (1996) Electromagnetic fault for earthquake lightning. *Jpn Jour Appl Phys Part 2* 35(3A):355–357
33. Kamagawa M (2006) Preseismic lithosphere-atmosphere-ionosphere coupling. *Eos* 87:417, 424
34. Kamogawa M (2007) Reply to comment on preseismic lithosphere-atmosphere-ionosphere coupling. *Eos* 88:248
35. Kamogawa M, Ohtsuki YH (1999) Plasmon model for origin of earthquake related electromagnetic wave noises. *Proc Japan Acad* 75(Ser. B):186–189
36. Kamogawa M, Liu JY, Fujiwara H, Chuo YJ, Tsai YB, Hattori K, Nagao T, Uyeda S, Ohtsuki YH (2004) Atmospheric field variations before the March 31 2002 M6.8 Earthquake in Taiwan. *Terr Atmos Ocean Sci* 15:445–461
37. Kamogawa M, Ofuruton H, Ohtsuki YH (2005) Earthquake light: 1995 Kobe earthquake in Japan. *Atmos Res* 76:438–444
38. Keilis-Borok VI, Soloviev AA (eds) (2003) Nonlinear dynamics of the lithosphere and earthquake prediction. Springer, Heidelberg, 335 pp
39. Kopytenko YA, Matishvili TG, Voronov PM, Kopytenko EA, Molchanov OA (1993) Detection of ultra-low-frequency emissions connected with the Spitak earthquake and its aftershock activity, based on geomagnetic pulsation data at Dusheti and Vardzia observatories. *Phys Earth Planet Inter* 77:85–95
40. Kushida Y, Kushida R (2002) Possibility of earthquake forecast by radio observations in the VHF band. *J Atmos Electr* 22:239–255
41. Lighthill J Sir (ed) (1996) A critical review of VAN. World Scientific, Singapore, 376 pp
42. Liu J, Chen Y, Ho Y (2004) A study of lightning activities and  $M \geq 5.0$  Earthquakes in Taiwan during 1993–2002. *Eos Trans AGU* 85(47):T51B-0456 (Fall Meet. Suppl., Abstract)
43. Liu JY, Chen YI, Pulnits SA, Tsai YB, Chuo YJ (2000) Seismo-ionospheric signatures prior to  $M \geq 6.0$  Taiwan earthquakes. *Geophys Res Lett* 27:3113–3116
44. Liu JY, Chen YI, Chuo YJ, Tsai HF (2001) Variations of ionospheric total electron content during the Chi-Chi earthquake. *Geophys Res Lett* 28:1383–1386
45. Liu JY, Chen YI, Chuo YJ (2006) A statistical investigation of pre-earthquake ionospheric anomaly. *J Geophys Res* 111:A05304. doi:10.1029/2005JA011333
46. Lizunov G, Hayakawa M (2004) Atmospheric gravity waves and their role in the lithosphere-troposphere-ionosphere interaction 1109. *IEEJ Trans Fundam Mater* 124-A:1109–1120
47. Maeda K, Tomisaka T (1996) Decametric radiation at the time of

- the Hyogo-ken Nanbu earthquake near Kobe in 1995. *Geophys Res Lett* 23:2433–2436
48. Maekawa S, Horie T, Yamauchi T, Sawaya T, Ishikawa M, Hayakawa M, Sasaki H (2006) A statistical study on the effect of earthquakes on the ionosphere, based on the subionospheric LF propagation data in Japan. *Ann Geophys* 24:2219–2225
  49. Marenko VF (1989) Investigation of the relationship between seismic processes and disturbances to the lower ionosphere by means of VLF radio transmissions. Ph.D. Dissertation, USSR Academy of Sciences, Siberian Department, Irkutsk, 160 pp
  50. Michael AJ (1997) Testing prediction methods: Earthquake clustering versus the Poisson model. *Geophys Res Lett* 24:1891–1894
  51. Mizutani H, Ishido T, Yokokura T, Ohnishi S (1976) Electrokinetic phenomena associated with earthquakes. *Geophys Res Lett* 3:365–368
  52. Molchanov OA, Hayakawa M (1998) Subionospheric VLF signal perturbations possibly related to earthquakes. *J Geophys Res* 100:1691–1712
  53. Mulargia F, Gasperini P (1992) Evaluating the statistical validity beyond chance of VAN earthquake precursors. *Geophys J Int* 111:32–44
  54. Musha K (1932) Investigations into the luminous phenomena accompanying earthquakes. *Bull Earthquake Res Inst Tokyo Univ* 10:666–673
  55. Nagao T, Enomoto Y, Fujinawa Y, Hata M, Hayakawa M, Huang Q, Izutsu J, Kushida Y, Maeda K, Oike K, Uyeda S, Yoshino T (2002) Electromagnetic anomalies associated with 1995 Kobe earthquake. *J Geodynamics* 33:349–359
  56. Norwick AS (1996) The golden age of crystal defects. *Ann Rev Mater Sci* 26:1–19
  57. Ouzounov DP, Williams RG, Wohlman R (2000) A joint analysis of earthquake and lightning activity in the Southern California (1995–1999). *Eos Trans AGU* 81(19):S41B-08 (Spring Meet. Suppl. Abstract)
  58. Parrot M (ed) (2007) First results of the DEMETER micro-satellite. *Planet Space Sci* 54(5):411–558
  59. Pilipenko V, Shamimov S, Uyeda S, Tanaka H (2001) Possible mechanism of the over-horizon reception of FM radio waves during earthquake preparation period. *Proc Japan Acad* 77(Ser. B):125–130
  60. Pulnits S (2007) Natural radioactivity, earthquakes, and the ionosphere. *Eos* 88:217–218
  61. Pulnits S, Boyarchuk K (2005) Ionospheric precursors of earthquakes. Springer, p 315
  62. Pulnits SA, Boyarchuk KA, Hegai VV, Kim VP, Lomonosov AM (2000) Quasielectrostatic model of atmosphere-thermosphere-ionosphere coupling. *Adv Space Res* 26:1209–1218
  63. Pulnits SA, Legen'ka AD, Gaivoronskaya TV, Depuev VK (2003) Main phenomenological features of ionospheric precursors of strong earthquakes. *J Atmos Solar Terr Phys* 65:1337–1347
  64. Rodger CJ, Lilvered MA (2007) Comment on preseismic lithosphere-atmosphere-ionosphere coupling. *Eos* 88:248
  65. Rundle JB, Turcotte DL, Sammis C, Klein W, Shcherbakov R (2003) Statistical physics approach to understanding the multiscale dynamics of earthquake fault systems. *Rev Geophys* 41(4). doi:10.1029/2003RG000135
  66. Schreider SY (1990) Formal definition of premonitory seismic quiescence. *Phys Earth Planet Inter* 61:113–127
  67. Sobolev GA, Zakrzhevskaya NA, Kharin EP (2001) On a relation between seismicity and magnetic storms. *Phys Earth* 11:66–72
  68. Sornette D (2000) Critical phenomena in natural sciences. Springer, Berlin, 434 pp
  69. Terada T (1931) On luminous phenomena accompanying earthquakes. *Bull Earthq Res Inst Tokyo Univ* 9:225–255
  70. Tramutoli V, Di Bello G, Pergola N, Piscitelli S (2001) Robust satellite, techniques for remote sensing of seismically active areas. *Annali di Geofisica* 44:295–312
  71. Tributsch H (1982) When the snakes awake. MIT Press, Cambridge, 248 pp
  72. Tronin AA (1996) Satellite thermal survey – a new tool for the study of seismoactive regions. *Int J Remote Sens* 41:1439–1455
  73. Turcotte DL (1997) Fractals and chaos in geology and geophysics. Cambridge University Press, Cambridge, 398 pp
  74. Uyeda S (1996) Introduction to the VAN method of earthquake prediction, a critical review of VAN. World Scientific, London, Singapore, pp 3–28
  75. Uyeda S, Al-Damegh K, Dologlou E, Nagao T (1999) Some relationship between VAN seismic electric signals (SES) and earthquake parameters. *Tectonophysics* 304:41–55
  76. Uyeda S, Nagao T, Orihara Y, Yamaguchi Y, Takahashi T (2000) Geoelectric potential changes: Possible precursors to earthquakes in Japan. *Proc Nat Acad Sci USA (PNAS)* 97:4561–4566
  77. Uyeda S, Hayakawa M, Nagao T, Molchanov O, Hattori K, Orihara Y, Gotoh K, Akinaga Y, Tanaka H (2002) Electric and magnetic phenomena observed before the volcano-seismic activity 2000 in the Izu Island Region, Japan. *Proc Nat Acad Sci USA (PNAS)* 99(11):7352–7355
  78. Varotsos P, Alexopoulos K (1984) Physical properties of the variations of the electric field of the earth preceding earthquakes. *Tectonophysics* 110:73–125
  79. Varotsos P, Alexopoulos K (1986) Stimulated current emission in the earth and related geophysical aspects. In: Amelinckx S, Gevers R, Nihoul J (eds) *Thermodynamics of point defects and their relation with bulk properties*. North Holland, Amsterdam
  80. Varotsos P, Kulhanek O (eds) (1993) Measurements and theoretical models of the Earth's electric field variations related to earthquakes. *Tectonophysics* 224:1–288
  81. Varotsos P, Sarlis N, Lazaridou M, Kaporis P (1998) Transmission of stress induced electric signals in dielectric media. *J Appl Phys* 83:60–70
  82. Varotsos PA (2005) The physics of seismic electric signals. TerraPub, Tokyo, 338 pp
  83. Varotsos PA, Sarlis N, Skordas E (2002) Long-range correlations in the electric signals that precede rupture. *Phys Rev E* 66:011902
  84. Varotsos PA, Sarlis NV, Skordas ES, Lazaridou MS (2004) Entropy in the natural time domain. *Phys Rev E* 70:011106
  85. Varotsos PA, Sarlis NV, Skordas ES, Lazaridou MS (2005) Natural entropy fluctuations discriminate similar-looking electric signals emitted from systems of different dynamics. *Phys Rev E* 71:011110
  86. Varotsos PA, Sarlis NV, Skordas ES, Tanaka HK, Lazaridou MS (2006) Entropy of seismic electric signals: Analysis in natural time under time reversal. *Phys Rev E* 73:031114
  87. Wakita H, Nakamura Y, Notsu K, Noguchi M, Asada T (1980) Radon anomaly: A possible precursor of the 1978 Izu-Oshima-Kinkai Earthquake. *Science* 207:882–883
  88. Warwick JW, Stoker C, Meyer TR (1982) Radio emission associated with rock fracture: Possible application to the great Chilean earthquake of May 22 1960. *J Geophys Res* 87:2851–2859

89. Weron A, Burnecki K, Mercik S, Weron K (2005) Complete description of all self-similar models driven by Lévy stable noise. *Phys Rev E* 71:016113
90. Yamada I, Masuda K, Mizutani H (1989) Electromagnetic and acoustic emission associated with rock fracture. *Phys Earth Planet Int* 57:157–168
91. Yasui Y (1968) A study on the luminous phenomena accompanied with earthquakes (part 1). *Mem Kakioka Mag Obs* 13:25–61
92. Yasuoka Y, Igarashi G, Ishikawa T, Tokonami S, Shinogi M (2006) Evidence of precursor phenomena in the Kobe earthquake obtained from atmospheric radon concentration. *Appl Geochem* 21:1064–1072
93. Yoshida S (2001) Convection current generated prior to rupture in saturated rocks. *J Geophys Res* 106(B2):2103–2120
94. Yoshida S, Uyeshima M, Nakatani M (1997) Electric potential changes associated with slip failure of granite: Preseismic and coseismic signals. *J Geophys Res* 102:14883–14897
95. Yoshino T, Tomozawa I, Sugimoto T (1993) Results of statistical analysis of low-frequency seismogenic EM emissions as precursors to earthquakes and volcanic eruptions. *Phys Earth Planet Interi* 77:21–31

## Earth's Crust and Upper Mantle, Dynamics of Solid–Liquid Systems in

YASUKO TAKEI

Earthquake Research Institute, University of Tokyo,  
Tokyo, Japan

### Article Outline

Glossary

Definition of the Subject

Introduction

General Theoretical Framework to Describe  
the Dynamics of Solid–Liquid Composite Systems

Overview of Applications

Elastic Wave Propagation  
in a Solid–Liquid Composite System

Future Directions

Acknowledgments

Bibliography

### Glossary

**Partially molten rock** The partially molten state is a thermodynamic state between solidus and liquidus temperatures, where both solid and liquid phases co-exist. In the Earth's interior, partial melting of rocks occurs in the upper mantle and/or crust beneath volcanic areas.

**Melt** Liquid phase in partially molten rocks or completely molten rock above the liquidus temperature is called melt. Density of melt is about 10% lower than solid. Hence, melt phase in the partially molten rocks tend to ascend toward the Earth's surface.

**Aqueous fluid** H<sub>2</sub>O-rich fluid. In a subducting oceanic plate, at the depths of several tens of km, aqueous fluids are released by the dehydration of hydrated minerals. Aqueous fluids, having much lower density and viscosity than melts, tend to ascend due to the buoyancy force.

**Seismic tomographic image** A number of seismometer networks have been placed on the surface of the Earth to record the seismic wave propagation from seismic sources at depths to the surface. Using the traveltime data obtained from these observations, three-dimensional seismic velocity structures in the Earth can be obtained, with a process called seismic tomographic imaging. By using P and S wave traveltimes,  $V_P$  and  $V_S$  structures, respectively, can be obtained.

### Definition of the Subject

The dynamics of solid-liquid composite systems are of great relevance to many problems in the earth sciences, including how melts or aqueous fluids generated by partial melting or dehydration migrate through the mantle and crust toward the surface, how deformation and fracture in these regions are influenced by the existence of fluids, and also how these fluids can be observed in the seismic tomographic images. The mechanical and transport properties of the solid-liquid composite systems strongly depend on liquid volume fraction and pore geometry, such as pore shape, pore size, and a detailed porosity distribution. Therefore, the microstructural processes that control pore geometry influence macroscopic dynamics, and vice versa. This article introduces a general continuum mechanical theory to treat the macroscopic dynamics of solid-liquid composite systems with a special emphasis on how such interactions with pore geometry can be described. Although intensive experimental and modeling approaches have been performed to investigate the interactive evolution of pore geometry and matrix deformation or fluid flow, many problems still remain unsolved and the actual liquid content and pore geometries in the crust and mantle are poorly understood. Therefore, in the latter part of this article, by applying the general theoretical framework introduced in the former part to the seismic wave propagation, the determinability of porosity and pore geometry from seismic tomographic images is discussed in detail. Due to the recent advances in seismic tomography, we can obtain three-dimensional and highly resolved images of both  $V_P$  and  $V_S$  structures. From the  $V_P$  or  $V_S$  structure alone, neither porosity nor pore geometry can be determined independently. However, if both  $V_P$  and  $V_S$  structures are available, porosity and pore geometry can be determined independently, thus providing valuable information complementary to experimental and modeling approaches. A practical method to determine porosity and pore geometry from the  $V_P$  and  $V_S$  images is presented.

### Introduction

Melt segregation from partially molten mantle or crust to the surface is the fundamental process of volcanism. In arc volcanism, aqueous fluids derived from the dehydrating subducting slab migrate through the mantle wedge and play an important role in the melting process. Modeling approaches to these fluid migration processes have been developed to explain various chemical and petrological observations, e. g. [13,20,35,37]. To make in situ observations of these fluids, seismic and/or electromagnetic tomographic images of the partially molten regions are

produced. These images are then interpreted using methods built on experimental and theoretical studies of the effects of fluids on seismic and electromagnetic properties, e. g. [2,23,24,44]. The effects of fluids on the tectonic and/or volcanic earthquake source process have also been of great interests; recently, observations of deep low-frequency earthquakes and tremors have been considered to be indicators of the presence or active migration of fluids in the crust and mantle, e. g. [9,26,29].

Fluids in the crust and mantle exist as solid-liquid composite systems in which fluid-filled pores are included in the solid matrix. Solid-liquid composite systems are characterized by high structural sensitivity. When the liquid volume fraction increases from zero to a few tens of %, the mechanical and transport properties of the system change greatly from those of a solid to those of a liquid. These properties are not simply determined by the liquid volume fraction but also strongly depend on the geometry of the liquid-filled pores. Here, the term pore geometry represents not only pore shape, but also pore size, orientation, and homogeneous or heterogeneous porosity distribution. Therefore, liquid content and pore geometry strongly influence the melt segregation and/or matrix deformation dynamics. Pore geometry is not constant but can change interactively with the macroscopic dynamics. Experimental and modeling approaches have been performed to investigate the interactive evolution of pore geometry and matrix deformation or fluid flow, e. g. [11,35,39]. However, many problems, including a poor understanding of the underlying physics, discrepancies between the experimental and modeling results, and the issue of scaling laboratory results to km scale, remain unsolved and detailed forward approaches to the interactions are the subjects of future studies.

Remarkable progress has been made in the seismic observations of the solid-liquid composite systems in the crust and mantle. Due to the recent advances in seismic tomography, we can now obtain three-dimensional and highly-resolved images of both  $V_p$  and  $V_s$  structures. Because the seismic velocity depends on both liquid volume fraction and pore geometry, neither can be estimated independently from  $V_p$  or  $V_s$  alone. However, when both  $V_p$  and  $V_s$  are available, the liquid volume fraction and pore geometry can be estimated independently. By using  $V_p$  and  $V_s$  images, an inverse approach to estimate the actual pore geometry and fluid content in the crust and mantle can be performed, yielding complementary information to the experimental and modeling approaches mentioned above. Therefore, in the present introduction of the dynamics of solid-liquid composite systems, elastic wave propagation is discussed in detail with a special focus on

the determinability of fluid content and pore geometry from seismic tomographic data.

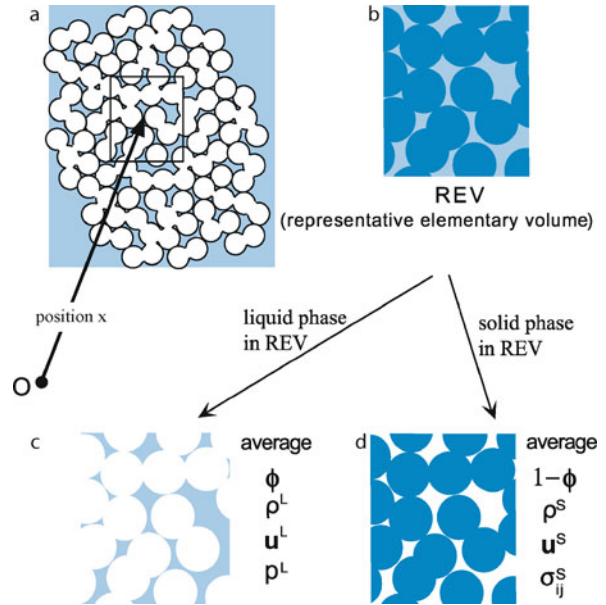
First, in Sect. “**General Theoretical Framework to Describe the Dynamics of Solid-Liquid Composite Systems**”, I introduce a general continuum mechanical formulation of the macroscopic dynamics of solid-liquid composite systems, with a special emphasis on how structural sensitivity is described. The elastic version of this theory, called “linear poroelasticity”, was developed by Biot and coworkers, e. g. [5,6,31,48]. This theory, which assumes infinitesimal strain in the solid phase, is applicable to the propagation of elastic waves. A more general version applicable to large deformations was developed based on the fluid dynamic theory, and applied to the melt segregation dynamics in partially molten mantle, e. g. [7,20]. On the one hand, in most reviews of the theory of linear poroelasticity, the governing equations are introduced empirically and are difficult to compare to the mass and momentum conservation equations used in the usual continuum mechanical formulation, e. g. [48]. However, in those reviews, detailed explanations are given for the meaning of the macroscopic constitutive relation that describes the structural sensitive character of the solid-liquid composite systems. On the other hand, in most reviews of the general fluid-dynamical formulation the mass and momentum conservations can be easily confirmed, but the physical meanings of the macroscopic constitutive relation and the structural-sensitive character are difficult to understand. Therefore, by taking advantage of both fields of study, I introduce here a general fluid-dynamical formulation with a detailed explanation of the macroscopic constitutive relation. The general formulation includes as a special case the theory of linear poroelasticity. The governing equations introduced in Sect. “**General Theoretical Framework to Describe the Dynamics of Solid-Liquid Composite Systems**” include several structural-sensitive parameters that are given as functions of liquid volume fraction and pore geometry. Provided an evolution equation of pore geometry is obtained in future studies, it will be possible to investigate the interaction between pore geometry and macroscopic dynamics by solving the governing equations including the constitutive relation together with the evolution equation.

In Sect. “**Overview of Applications**”, a brief summary of the various applications of the general theory introduced in Sect. “**General Theoretical Framework to Describe the Dynamics of Solid-Liquid Composite Systems**” is presented. In Sects. “**Derivation of Wave Equations**” to “**Dispersion and Attenuation of Waves in Solid-Liquid Composite Systems**”, the theory is applied to the elastic wave propagation in a solid-liquid composite system; we linearize the general formulation and derive the wave

equations (Sect. “**Derivation of Wave Equations**”). Based on the detailed descriptions of  $V_p$  and  $V_s$  obtained from the wave equations, the effects of liquid volume fraction, pore geometry, and liquid compressibility on the velocities are summarized systematically (Sects. “**Porosity and Pore Shape**” to “**Determinability of Porosity and Pore Shape from Elastic Wave Velocities**”). To assess the determinability of pore geometry, the usual forward modeling based on a priori assumed pore geometries, e.g. [15,17,18,27,41] are not enough, and a systematic treatment of general pore geometries is required, which is enabled by the introduction of the concept of equivalent aspect ratio (Sect. “**Porosity and Pore Shape**”). The determinability of porosity and pore geometry from seismic tomographic data is discussed in Sect. “**Application to Seismic Tomographic Images**”. In Sect. “**Dispersion and Attenuation of Waves in Solid-Liquid Composite Systems**”, so as to clarify the limitation in the application of the theoretical results, I briefly discuss the attenuation and dispersion of elastic waves. In this article, the term ‘liquid’ is used with the same meaning as ‘fluid’. Hence, ‘liquid’ in this paper includes ‘gas’.

### General Theoretical Framework to Describe the Dynamics of Solid-Liquid Composite Systems

A schematic illustration of a solid-liquid composite system considered in this article is shown in Fig. 1a. Length scales less than the grain size are referred to as “microscopic”; length scales greater than the grain size are referred to as “macroscopic”. Solid-liquid composite systems are characterized by large differences in mechanical properties between solid and liquid phases, and hence the stress and velocity fields that develop under external forces are usually highly heterogeneous at the microscopic scale. However, when studying macroscopic dynamics such as mantle-scale melt segregation and propagation of seismic waves with wavelengths much larger than the grain size, it is not practical to solve both the microscopic and macroscopic processes simultaneously. In this section, I review the theoretical framework to treat the macroscopic dynamics separately from the microscopic processes. In this theory, macroscopic dynamics of solid-liquid composite systems are described within the framework of continuum mechanics, using the macroscopic variables obtained by averaging the microscopic fields. The averages within the solid and liquid phases are taken separately, so that the theory can be applied to the phenomena involving a relative motion between the two phases. Although microscopic variables do not explicitly appear in the governing equations, several parameters included in these equa-



Earth's Crust and Upper Mantle, Dynamics of Solid-Liquid Systems in, Figure 1

**a** A solid-liquid composite system with a representative elementary volume (REV) at position  $x$ . **b** REV consisting of solid (thick blue) and liquid (light blue) phases. **c** Liquid phase in REV and the averaged quantities representing the macroscopic mechanical state of the liquid. **d** Solid phase in REV and the averaged quantities representing the macroscopic mechanical state of the solid

tions are sensitive to the microstructures and thus the microscopic processes do affect the macroscopic dynamics through these parameters. In this review, a special emphasis is given to these structurally sensitive parameters.

### Macroscopic Variables

We consider a region called REV (representative elementary volume), which is small enough to consider as a point in the macroscopic scale but large enough to contain a number of solid grains. Macroscopic quantities are defined by the average of the corresponding quantities in REV, where the averages within the liquid phase and those within the solid phase are taken separately. The averaging procedure can be defined by using a window function  $W$  and phase function  $A$ . By considering the REV as a cuboid with edge length  $L_i$ , the window function  $W(\mathbf{x})$  takes the value 1 at  $-L_i/2 \leq x_i \leq L_i/2$  ( $i = x, y, z$ ) and value 0 otherwise. The phase function  $A(\mathbf{x})$  takes the value 1 if the position  $\mathbf{x}$  is in the liquid phase, and value 0 in the solid phase;  $A$  and  $1 - A$  quantify the properties of the liquid and solid phases, respectively. Let  $a(\mathbf{x}, t)$  be a microscopic

field of the physical quantity  $a$ , which is a function of the position  $\mathbf{x}$  and time  $t$ . The phasic average of  $a$  in the liquid phase ( $a^L$ ) or solid phase ( $a^S$ ) is defined by

$$\begin{aligned} a^L(\mathbf{x}, t) &= \frac{1}{V\phi} \int W(\boldsymbol{\xi} - \mathbf{x}) A(\boldsymbol{\xi}, t) a(\boldsymbol{\xi}, t) dV_{\boldsymbol{\xi}} \\ a^S(\mathbf{x}, t) &= \frac{1}{V(1-\phi)} \int W(\boldsymbol{\xi} - \mathbf{x}) (1 - A(\boldsymbol{\xi}, t)) \cdot a(\boldsymbol{\xi}, t) dV_{\boldsymbol{\xi}}, \end{aligned} \quad (1)$$

where  $V(=L_x L_y L_z)$  represents the volume of REV and  $\phi$  represents the liquid volume fraction, defined as

$$\phi(\mathbf{x}, t) = \frac{1}{V} \int W(\boldsymbol{\xi} - \mathbf{x}) A(\boldsymbol{\xi}, t) dV_{\boldsymbol{\xi}}. \quad (2)$$

The volume integrals in Eqs. (1)–(2) are taken over whole of the solid-liquid system.

The dynamic state of the solid-liquid composite system at a spatially-fixed position  $\mathbf{x}$  is described by the following 7 macroscopic variables defined by Eqs. (1) and (2):

$\phi(\mathbf{x}, t)$ ...liquid volume fraction (nondimensional)

$\rho^L(\mathbf{x}, t)$ ...density of liquid ( $\text{kg/m}^3$ )

$\rho^S(\mathbf{x}, t)$ ...density of solid ( $\text{kg/m}^3$ )

$\mathbf{u}^L(\mathbf{x}, t)$ ...displacement of liquid (m)

$\mathbf{u}^S(\mathbf{x}, t)$ ...displacement of solid (m)

$p^L(\mathbf{x}, t)$ ...liquid pressure (Pa), with compression positive

$\sigma_{ij}^S(\mathbf{x}, t)$ ...solid stress (Pa), with tension positive.

It is rigorous to define  $\mathbf{u}^L$  and  $\mathbf{u}^S$  by using mass-weighted average [7]. However, for simplicity, the density heterogeneity within each phase is assumed to be small and the mass-weighted average is approximated by the phasic average.

## Governing Equations

The seven variables introduced in Sect. “**Macroscopic Variables**” are governed by the following seven equations:

mass conservation of liquid

$$\frac{\partial(\phi\rho^L)}{\partial t} + \nabla \cdot (\phi\rho^L\dot{\mathbf{u}}^L) = \Gamma \quad (A)$$

mass conservation of solid

$$\frac{\partial\{(1-\phi)\rho^S\}}{\partial t} + \nabla \cdot \{(1-\phi)\rho^S\dot{\mathbf{u}}^S\} = -\Gamma \quad (B)$$

intrinsic constitutive relation of liquid

$$\frac{\delta\rho^L}{\rho^L} = \frac{1}{k_L}\delta p^L \quad (C)$$

intrinsic constitutive relation of solid

$$\frac{\delta\rho^S}{\rho^S} = \frac{1}{k_S}\delta p^S \quad (D)$$

macroscopic constitutive relation of solid framework

$$\epsilon_{ij} = S_{ijkl} (\sigma_{kl}^S + p^L \delta_{kl}) - \frac{1}{3k_S} p^L \delta_{ij} \quad (E)$$

linear momentum conservation of liquid

$$\phi\rho^L\ddot{\mathbf{u}}^L = -\nabla(\phi p^L) + \phi\rho^L\mathbf{g} + \mathbf{I} \quad (F)$$

linear momentum conservation of solid

$$(1-\phi)\rho^S\ddot{\mathbf{u}}^S = \nabla \cdot \{(1-\phi)\sigma^S\} + (1-\phi)\rho^S\mathbf{g} - \mathbf{I}, \quad (G)$$

where  $\mathbf{I}$  ( $\text{N/m}^3$ ) in Eqs. (F) and (G) represents the interaction between the solid and liquid phases (the force applied to the liquid from the solid is taken positive), and is explicitly written as

$$\mathbf{I} = -\frac{\eta_L\phi^2}{k_\phi} (\dot{\mathbf{u}}^L - \dot{\mathbf{u}}^S) + p^L\nabla\phi. \quad (3)$$

The  $\Gamma$  ( $\text{kg/s/m}^3$ ) in Eqs. (A) and (B) represents the net mass flux from the solid to the liquid,  $k_L$  and  $k_S$  (Pa) in Eqs. (C), (D), and (E) represent the intrinsic bulk moduli of the liquid and solid, respectively,  $p^S = -(\sigma_{xx}^S + \sigma_{yy}^S + \sigma_{zz}^S)/3$  (Pa) in Eq. (D) represents the solid pressure (compression positive),  $\epsilon_{ij}$  in the left hand side of Eq. (E) represents the framework strain (extension positive), whose definition is given in Sect. “Equation (E)”,  $S_{ijkl}$  ( $\text{Pa}^{-1}$ ) in Eq. (E) represents the elastic compliance tensor of the solid framework,  $\mathbf{g}$  ( $\text{N/kg}$ ) in Eqs. (F) and (G) represents the gravitational acceleration vector, and  $\eta_L$  (Pa s) and  $k_\phi$  ( $\text{m}^2$ ) in Eq. (3) represent the liquid viscosity and permeability, respectively. For  $\alpha = S, L$ ,  $\dot{\mathbf{u}}^\alpha = D\mathbf{u}^\alpha/Dt$  and  $\ddot{\mathbf{u}}^\alpha = D\dot{\mathbf{u}}^\alpha/Dt$  represent velocity and acceleration vectors, respectively, where  $D/Dt = \partial/\partial t + \dot{\mathbf{u}}^\alpha \cdot \nabla$ . Also,  $\nabla \cdot \sigma = \partial\sigma_{ij}/\partial x_j = \sigma_{ij,j}$ . The summation convention for repeated subscripts is employed.

These equations, except for Eq. (E), are rigorously derived by averaging the mass and linear momentum conservation equations and constitutive relations required in the microscopic scale [7]. The physical meanings of Eqs. (A)–(D) are made clear by analogy with the standard fluid dynamic equations. The term  $\Gamma$  included in Eqs. (A)–(B) is zero unless melting/solidification or dissolution/precipitation occurs. The parameters that are sensitive to the microstructures are  $S_{ijkl}$  and  $k_\phi$  included in Eqs. (E)–(G). Therefore, in the following part of this

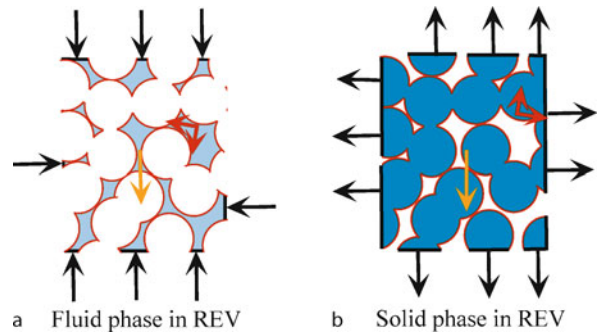


section, detailed discussions of these three equations are given. In Sect. “**Semi-Intuitive Derivations of Equations (E), (F), and (G)**”, derivations of Eqs. (E)–(G) are given in a semi-intuitive manner to clarify the physical meaning of these equations and the structural-sensitive parameters. Several important aspects of these equations are also discussed in Sects. “**Relation to the Effective Medium Theory**”–“**Fundamental Assumption for Stress Heterogeneity**”.

### Semi-Intuitive Derivations of Equations (E), (F), and (G)

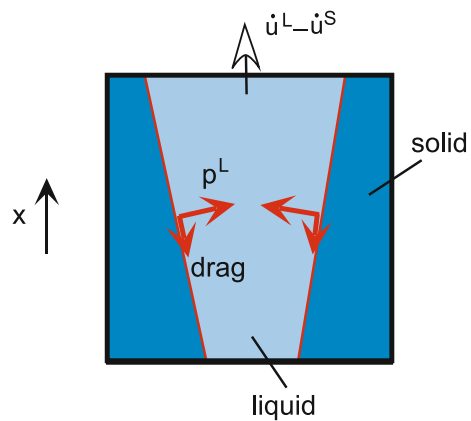
**Equations (F) and (G)** Equations (F) and (G) describe the linear momentum conservations for the liquid and solid phases, respectively, in the REV. The left hand sides (LHS) of these equations represent the acceleration terms. The right hand sides (RHS) represent the total forces applied on each system, including the body force and the surface force. The surface force is applied through the boundary surface of each system; this boundary is made of the boundary on the surface of REV (black boundary in Figs. 2a and 2b) and the boundary with the other phase (red boundary in Figs. 2a and 2b). The former boundary (black) exists within each phase and the latter boundary (red) exists on the phase boundary. In the RHS of Eq. (F) or (G), the second term represents the body force due to gravity, and the first and the third terms represent the surface forces applied through the black and red boundaries, respectively. The surface force through the red boundary is the interaction between solid and liquid, and hence the third term in the RHS of Eq. (F), I, and that of Eq. (G),  $-I$ , are of the same magnitude and of opposite sign.

Equation (3) shows that interaction I consists of two terms, corresponding to the contributions from the traction components tangential and normal to the phase boundary. The first term, corresponding to the tangential component, represents the viscous drag force proportional to the relative velocity of the two phases. The permeability  $k_\phi$  included in the proportionality constant depends on the detailed geometry of the liquid-filled pores. The second term in the RHS of Eq. (3) represents the contribution from the normal component of traction, which is determined by the liquid pressure. An intuitive explanation for the proportional dependence of this force on the porosity gradient is given in Fig. 3, in which a simple pore geometry is assumed. A more rigorous derivation of this term for general pore geometries is presented in Sect. “**Fundamental Assumption for Stress Heterogeneity**”, where the assumption needed to derive this term is clarified and its validity is discussed. In the dynamics of solid-liquid com-



Earth's Crust and Upper Mantle, Dynamics of Solid-Liquid Systems in, Figure 2

**a** Forces applied on the liquid phase in REV. **b** Forces applied on the solid phase in REV. Forces applied on each system consist of body force (orange arrow) and surface force (black and red arrows). The body force is due to gravity. The surface force is applied through the boundary surface of each system, which is divided into the boundary on the surface of REV (black boundary) and the boundary with the other phase (red boundary). The surface force applied through the red boundary is shown by dividing into normal and tangential components to the interface (red arrows). Solid stress is taken to be tension positive and liquid pressure is taken to be compression positive

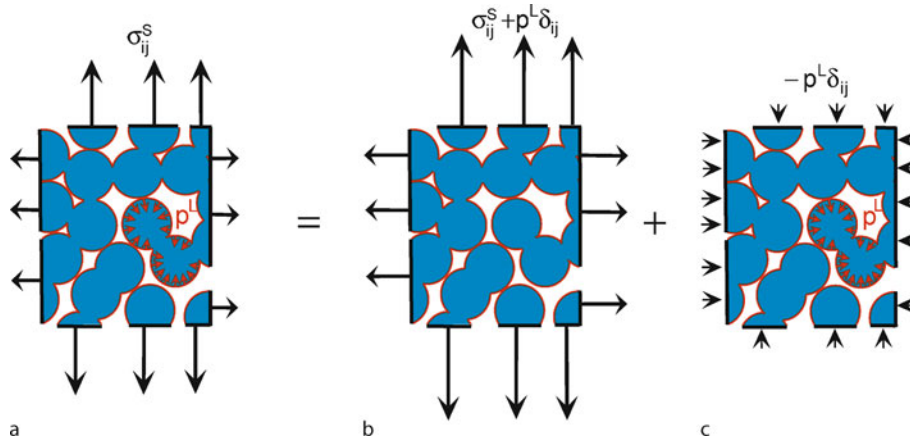


Earth's Crust and Upper Mantle, Dynamics of Solid-Liquid Systems in, Figure 3

Interaction between the solid and liquid phases through the phase boundary (red boundary) is schematically shown for a simple pore geometry with a porosity gradient. The forces applied from solid to liquid are shown. The total force of the force component normal to the phase boundary does not vanish if porosity gradient is not zero, and that tangential to the phase boundary does not vanish if average velocity of liquid relative to solid is not zero

posite systems, the motion of each phase is significantly affected by the interaction with the other phase through the phase boundary.

Note that Eqs. (F) and (G) are derived by implicitly assuming a connectivity of each phase (black boundary in



Earth's Crust and Upper Mantle, Dynamics of Solid–Liquid Systems in, Figure 4

**a** A stress state of the solid phase in REV, generally represented by the solid stress  $\sigma_{ij}^S$  (tension positive) and liquid pressure  $p^L$  (compression positive), is expressed by the superposition of **b** effective stress state, defined by solid stress  $\sigma_{ij}^S + p^L \delta_{ij}$  and liquid pressure 0, and **c** uniform stress state, defined by solid stress  $-p^L \delta_{ij}$  and liquid pressure  $p^L$ . Equation (E) in the text states that the framework strain under a given stress state is obtained by the superposition of the framework strain under the effective stress state and that under the uniform stress state

Figs. 2a and 2b). Therefore, we need to be careful in applying the present theory to end-member systems of suspensions of solid or isolated inclusions of liquid, in which one phase is dispersed in the other phase without connectivity.

**Equation (E)** The framework strain  $\epsilon_{ij}$  in the LHS of Eq. (E) is defined by using the macroscopic displacement of the solid  $\mathbf{u}^S$  as

$$\epsilon_{ij} = \frac{1}{2} \left( \frac{\partial u_i^S}{\partial x_j} + \frac{\partial u_j^S}{\partial x_i} \right), \quad (4)$$

and hence  $\epsilon_{ij}$  represents the macroscopic deformation of the solid framework. Equation (E) provides a macroscopic constitutive relationship between the framework strain and the macroscopic (averaged) stress state of the framework. The coefficient  $S_{ijkl}$  included in this equation is not only determined by the solid intrinsic properties, but shows a large structural sensitivity, which plays a significant role in the dynamics of solid-liquid composite systems. Equation (E) is one of the key equations characterizing the two-phase dynamics. I present here a semi-intuitive derivation of Eq. (E).

As shown in Fig. 4a, a macroscopic stress state of the solid framework is generally described by the solid stress  $\sigma_{ij}^S$  applied through the boundary on the surface of REV (black boundary) and liquid pressure  $p^L$  applied through the boundary with the liquid phase (red boundary). The framework strain under this stress state is calculated by the superposition of the framework strain under stress

$\sigma_{ij}^S + p^L \delta_{ij}$  applied through the black boundary while leaving the red boundary as a free surface (Fig. 4b) plus the framework strain under uniform pressure  $p^L$  applied to all boundaries (Fig. 4c). The assumption of the superposition is valid when the response of the solid framework to the applied stresses is linear. The term  $\sigma_{ij}^S + p^L \delta_{ij}$  is called the effective stress. Figure 4b shows that the elastic compliance tensor  $S_{ijkl}$  specifying the effect of effective stress on the framework strain describes the mechanical properties of the ‘skeleton’ (solid framework obtained by replacing the regions occupied by the liquid phase with a vacuum), which is not only determined by the intrinsic properties of solid but also strongly depends on the porosity and pore geometry. For a given solid-liquid composite system,  $S_{ijkl}$  is estimated by using experimental and/or modeling approaches, some of which are presented in a later section. The framework strain under a given pressure  $p^L$  applied to all boundaries (Fig. 4c) is identical to the strain of REV completely filled with solid ( $\phi = 0$ ) under uniform pressure  $p^L$ , and hence is determined only by the intrinsic bulk modulus of the solid phase (the last term in the RHS of Eq. (E)).

In the derivation mentioned above, the solid phase is assumed to deform elastically. For an isotropic system, Eq. (E) is written as

$$\epsilon_{ij} = \frac{1}{3k_{sk}} \left( \frac{\sigma_{kk}^S}{3} + p^L \right) \delta_{ij} + \frac{1}{2\mu_{sk}} \left( \sigma_{ij}^S - \frac{\sigma_{kk}^S}{3} \delta_{ij} \right) - \frac{1}{3k_S} p^L \delta_{ij}, \quad (E_e)$$

where  $k_{sk}$  and  $\mu_{sk}$  represent the bulk and shear moduli, respectively, of the skeleton. Similarly, when the solid phase deforms viscously, the macroscopic constitutive relation for an isotropic system is written as

$$\dot{\epsilon}_{ij} = \frac{1}{3\xi_{sk}} \left( \frac{\sigma_{kk}^S}{3} + p^L \right) \delta_{ij} + \frac{1}{2\eta_{sk}} \left( \sigma_{ij}^S - \frac{\sigma_{kk}^S}{3} \delta_{ij} \right), \quad (\text{E}_v)$$

where  $\xi_{sk}$  and  $\eta_{sk}$  represent the bulk and shear viscosities, respectively, of the skeleton. Because viscous deformation is usually large in amplitude, the volumetric deformation due to the intrinsic compressibility of the solid phase, which corresponds to the 3rd term in the RHS of Eq. (E<sub>c</sub>), is neglected in Eq. (E<sub>v</sub>). Similar to  $k_{sk}$  and  $\mu_{sk}$ ,  $\xi_{sk}$  and  $\eta_{sk}$  are not only determined by the intrinsic property of the solid phase but also strongly depend on the porosity and pore geometry. Although the intrinsic compressibility of the solid phase is neglected in Eq. (E<sub>v</sub>), the volumetric component of the framework strain rate controlled by  $\xi_{sk}$  cannot be neglected. This is because even when the solid phase is made of incompressible material, the solid framework can change its volume by changing the porosity. This demonstrates the essential difference between the skeleton property and the intrinsic property of the solid. A more general description of the viscous constitutive relation is given by neglecting the last term in the RHS of Eq. (E) and replacing  $\epsilon_{ij}$  and elastic compliance tensor  $S_{ijkl}$  by  $\dot{\epsilon}_{ij}$  and viscous compliance tensor  $S_{ijkl}^V$ , respectively.

### Relation to the Effective Medium Theory

The structural sensitivity of the skeleton properties  $S_{ijkl}$  plays an important role in the two-phase dynamics. To predict quantitatively the microstructural effects on the skeleton properties, an effective medium theory has been developed. However, the applicability of these theoretical results to Eq. (E) is not self-evident, because the definition of the macroscopic strain given by Eq. (4) is different from that of the average strain used in the effective medium theory. In the effective medium theory, it is well-known that the stress and strain fields of the solid phase are highly heterogeneous at the microscopic scale, so that the local stress can be largely different from the macroscopic stress. However, in the fluid dynamical formulation of the two-phase dynamics, this point is not emphasized and a confusion between microscopic and macroscopic stresses sometimes occurs (Sect. “**Fundamental Assumption for Stress Heterogeneity**”). Therefore, it is important to establish a connection between

the fluid dynamical formulation and the effective medium theory.

When the skeleton properties are calculated in the effective medium theory, the effective properties of the solid-liquid composites are calculated under either drained conditions, in which the liquid pressure is kept constant, or under dry conditions, in which the liquid phase with zero bulk and shear moduli is kept under undrained conditions. In both cases, the space which remains after subtracting the solid phase from REV is considered to be filled with the pore phase. The microscopic displacement field in the pore phase can be obtained from the undrained solution by setting the bulk and shear moduli of the liquid to zero. The displacement field in the solid phase is continuously connected with that in the pore phase and there is no relative motion between the solid and pore phases. The effective properties under the dry (or drained) condition is given by

$$\epsilon_{ij}^B = S_{ijkl}^{\text{dry}} \sigma_{kl}^B, \quad (5)$$

where  $\epsilon_{ij}^B$  and  $\sigma_{ij}^B$  represent the average strain and stress, respectively, over both solid and pore phases,

$$\begin{cases} \sigma_{ij}^B &= (1 - \phi) \sigma_{ij}^S \\ \epsilon_{ij}^B &= (1 - \phi) \epsilon_{ij}^S + \phi \epsilon_{ij}^P, \end{cases} \quad (6)$$

the superscript “B” means bulk, and

$$\epsilon_{ij}^\alpha = \frac{1}{2} \left\langle \frac{\partial u_i}{\partial \xi_j} + \frac{\partial u_j}{\partial \xi_i} \right\rangle_\alpha \quad (\alpha = S, P) \quad (7)$$

represents the phasic average of strain in the solid or pore phase, e. g. [21,50]. To clarify the relationship between  $S_{ijkl}$  and  $S_{ijkl}^{\text{dry}}$ , the relationship between  $\epsilon_{ij}$  and  $\epsilon_{ij}^B$  needs to be specified.

Let  $\langle a \rangle_\alpha$  ( $\alpha = S, L$ ) be the phasic average of the quantity  $a$  defined in Eqs. (1). Let  $\Sigma_{\text{black}}^\alpha$  ( $\alpha = S, L$ ) be the boundary of the phase  $\alpha$  on the surface of REV (black boundary in Figs. 2 and 4) and let  $\Sigma_{\text{red}}$  be the phase boundary in REV (red boundary in Figs. 2 and 4). Let  $n_i$  be the outward unit normal to these boundaries, where the positive direction at  $\Sigma_{\text{red}}$  is outward to the liquid phase. From the definition of phasic average  $\langle \cdot \rangle_S$ , we obtain

$$\begin{aligned} & \frac{\partial \langle u_i \rangle_S}{\partial x_j} \\ &= \frac{\partial}{\partial x_j} \left[ \frac{1}{V(1 - \phi)} \int W(\xi - \mathbf{x})(1 - A(\xi)) u_i(\xi) dV_\xi \right] \end{aligned}$$

$$\begin{aligned}
 &= \frac{\partial \phi}{\partial x_j} \frac{1}{V(1-\phi)^2} \int W(\boldsymbol{\xi} - \mathbf{x})(1 - A(\boldsymbol{\xi}))u_i(\boldsymbol{\xi})dV_{\boldsymbol{\xi}} \\
 &+ \frac{1}{V(1-\phi)} \frac{\partial}{\partial x_j} \left[ \int W(\boldsymbol{\xi} - \mathbf{x})(1 - A(\boldsymbol{\xi}))u_i(\boldsymbol{\xi})dV_{\boldsymbol{\xi}} \right] \\
 &= \frac{\langle u_i \rangle_S}{1-\phi} \frac{\partial \phi}{\partial x_j} \\
 &+ \frac{1}{V(1-\phi)} \int -\frac{\partial W(\boldsymbol{\xi} - \mathbf{x})}{\partial \xi_j} (1 - A(\boldsymbol{\xi}))u_i(\boldsymbol{\xi})dV_{\boldsymbol{\xi}} \\
 &= \frac{\langle u_i \rangle_S}{1-\phi} \frac{\partial \phi}{\partial x_j} \\
 &- \frac{1}{V(1-\phi)} \int \frac{\partial \{W(\boldsymbol{\xi} - \mathbf{x}) \cdot (1 - A(\boldsymbol{\xi}))u_i(\boldsymbol{\xi})\}}{\partial \xi_j} dV_{\boldsymbol{\xi}} \\
 &+ \frac{1}{V(1-\phi)} \int W(\boldsymbol{\xi} - \mathbf{x})(1 - A(\boldsymbol{\xi})) \frac{\partial u_i(\boldsymbol{\xi})}{\partial \xi_j} dV_{\boldsymbol{\xi}} \\
 &+ \frac{1}{V(1-\phi)} \int W(\boldsymbol{\xi} - \mathbf{x}) \frac{\partial (1 - A(\boldsymbol{\xi}))}{\partial \xi_j} u_i(\boldsymbol{\xi})dV_{\boldsymbol{\xi}} \\
 &= \frac{\langle u_i \rangle_S}{1-\phi} \frac{\partial \phi}{\partial x_j} + \left\langle \frac{\partial u_i}{\partial \xi_j} \right\rangle_S \\
 &+ \frac{1}{V(1-\phi)} \int W(\boldsymbol{\xi} - \mathbf{x}) \frac{\partial (1 - A(\boldsymbol{\xi}))}{\partial \xi_j} u_i(\boldsymbol{\xi})dV_{\boldsymbol{\xi}} \\
 &= \frac{\langle u_i \rangle_S}{1-\phi} \frac{\partial \phi}{\partial x_j} + \left\langle \frac{\partial u_i}{\partial \xi_j} \right\rangle_S \\
 &+ \frac{1}{V(1-\phi)} \int_{\Sigma_{\text{red}}} u_i(\boldsymbol{\xi})n_j(\boldsymbol{\xi})dS_{\boldsymbol{\xi}}, \quad (8)
 \end{aligned}$$

where the volume integral of factor  $\partial\{W(\boldsymbol{\xi} - \mathbf{x})(1 - A(\boldsymbol{\xi}))u_i(\boldsymbol{\xi})\}/\partial\xi_j$  in the RHS of the 4th equation is converted to the surface integral of  $W(\boldsymbol{\xi} - \mathbf{x})(1 - A(\boldsymbol{\xi}))u_i(\boldsymbol{\xi})$  on the outermost boundary of the solid-liquid system, which is zero because  $W(\boldsymbol{\xi} - \mathbf{x})$  is zero outside the REV.

Because the displacement field  $u_i$  of the solid phase can be continuously connected to  $u_i$  of the pore phase, by using Gauss's theorem, the integral over  $\Sigma_{\text{red}}$  in the third term on the RHS of the last equation of (8) can be rewritten in terms of the volume integral in the pore phase. Because the boundary of the pore phase is given by  $\Sigma_{\text{red}}$  and  $\Sigma_{\text{black}}^L$ , we thus obtain

$$\begin{aligned}
 &\frac{\partial \langle u_i \rangle_S}{\partial x_j} \\
 &= \left\langle \frac{\partial u_i}{\partial \xi_j} \right\rangle_S + \frac{\langle u_i \rangle_S}{1-\phi} \frac{\partial \phi}{\partial x_j} \\
 &+ \frac{1}{V(1-\phi)} \int_{\Sigma_{\text{red}} + \Sigma_{\text{black}}^L} u_i(\boldsymbol{\xi})n_j(\boldsymbol{\xi})dS_{\boldsymbol{\xi}} \\
 &- \frac{1}{V(1-\phi)} \int_{\Sigma_{\text{black}}^L} u_i(\boldsymbol{\xi})n_j(\boldsymbol{\xi})dS_{\boldsymbol{\xi}}
 \end{aligned}$$

$$\begin{aligned}
 &= \left\langle \frac{\partial u_i}{\partial \xi_j} \right\rangle_S + \frac{\langle u_i \rangle_S}{1-\phi} \frac{\partial \phi}{\partial x_j} \\
 &+ \frac{\phi}{(1-\phi)} \frac{1}{V\phi} \int W(\boldsymbol{\xi} - \mathbf{x})A(\boldsymbol{\xi}) \frac{\partial u_i}{\partial \xi_j} dV_{\boldsymbol{\xi}} \\
 &- \frac{1}{(1-\phi)} \frac{\partial (\phi \langle u_i \rangle_S)}{\partial x_j} \\
 &= \left\langle \frac{\partial u_i}{\partial \xi_j} \right\rangle_S + \frac{\phi}{1-\phi} \left\langle \frac{\partial u_i}{\partial \xi_j} \right\rangle_P - \frac{\phi}{1-\phi} \frac{\partial \langle u_i \rangle_S}{\partial x_j}, \quad (9)
 \end{aligned}$$

where  $u_i$  at the surface of REV ( $\Sigma_{\text{black}}^L$  and  $\Sigma_{\text{black}}^S$ ) is considered to be the same between the solid and pore phases, because there is no relative motion between these two phases. From Eqs. (4) and (9), we obtain

$$\begin{aligned}
 \epsilon_{ij} &= (1-\phi)\epsilon_{ij}^S + \phi\epsilon_{ij}^P \\
 &= \epsilon_{ij}^B. \quad (10)
 \end{aligned}$$

Therefore, a simple conversion formula can be obtained linking  $S_{ijkl}^{\text{dry}}$  and  $S_{ijkl}$

$$S_{ijkl}^{\text{dry}} = (1-\phi)S_{ijkl}. \quad (11)$$

We further discuss the peculiarity of Eq. (E) that it cannot be derived by the phasic average of the microscopic constitutive relationship. Because the microscopic stress and strain fields in the solid phase are related to each other by the intrinsic constitutive relationship of the solid phase, we may take the average of this relationship over the solid phase in REV, to obtain

$$\epsilon_{ij}^S = S_{ijkl}^{\text{intrinsic}} \sigma_{kl}^S, \quad (12)$$

where  $S_{ijkl}^{\text{intrinsic}}$  represents the solid intrinsic properties. By comparing Eq. (E) to Eq. (12), it is apparent that the significant difference between  $S_{ijkl}$  and  $S_{ijkl}^{\text{intrinsic}}$  comes from the significant difference between  $\epsilon_{ij}$  and  $\epsilon_{ij}^S$  in that  $\epsilon_{ij}$  represents not only  $\epsilon_{ij}^S$  but also  $\epsilon_{ij}^P$ . From the definition of  $\phi$ ,

$$\begin{aligned}
 \frac{\partial \phi}{\partial x_i} &= \frac{\partial}{\partial x_i} \left[ \frac{1}{V} \int W(\boldsymbol{\xi} - \mathbf{x})A(\boldsymbol{\xi})dV_{\boldsymbol{\xi}} \right] \\
 &= \frac{-1}{V} \int W(\boldsymbol{\xi} - \mathbf{x})n_i(\boldsymbol{\xi})dS_{\boldsymbol{\xi}}. \quad (13)
 \end{aligned}$$

Then, from Eqs. (4), (8), and (13),  $\epsilon_{ij}$  can be rewritten as

$$\epsilon_{ij} = \epsilon_{ij}^S + \frac{\int_{\Sigma_{\text{red}}} ((u_i(\boldsymbol{\xi}) - \langle u_i \rangle_S) n_j(\boldsymbol{\xi}) + (u_j(\boldsymbol{\xi}) - \langle u_j \rangle_S) n_i(\boldsymbol{\xi})) dS_{\boldsymbol{\xi}}}{2V(1-\phi)}. \quad (14)$$

Thus, a significant difference between  $\epsilon_{ij}$  and  $\epsilon_{ij}^S$  implies a significant contribution from the second term in the RHS

of (14). This means that the microscopic displacement field of the solid phase at the boundary with the liquid can be systematically different from the average displacement of the solid phase. An example of such systematic deviation can be seen in the compaction of the solid framework, where a contraction of the solid phase observed macroscopically is compensated by a displacement of the solid-liquid phase boundary into the pore space.

### Fundamental Assumption for Stress Heterogeneity

A fundamental assumption implicitly used in formulating the dynamics of solid-liquid composite systems is that the microscopic stress field is homogeneous in the liquid phase but can be heterogeneous in the solid phase. This point is rarely stated explicitly. Here, I discuss this assumption first with respect to the liquid phase and then with respect to the solid phase.

The second term in the RHS of Eq. (3) represents the total force due to the normal traction component applied to the liquid phase through the phase boundary (red boundary). This term is derived as

$$\begin{aligned}
& -\frac{1}{V} \int_{\Sigma_{\text{red}}} p(\xi) n_i(\xi) dS_\xi \\
&= -\frac{1}{V} \int_{\Sigma_{\text{red}} + \Sigma_{\text{black}}^{\text{L}}} p(\xi) n_i(\xi) dS_\xi \\
&\quad + \frac{1}{V} \int_{\Sigma_{\text{black}}^{\text{L}}} p(\xi) n_i(\xi) dS_\xi \\
&= -\frac{1}{V} \int_{\text{REV}} A(\xi) \frac{\partial p(\xi)}{\partial \xi_i} dV_\xi + \frac{1}{V} \int_{\Sigma_{\text{black}}^{\text{L}}} p(\xi) n_i(\xi) dS_\xi \\
&= -\phi \left\langle \frac{\partial p}{\partial \xi_i} \right\rangle_{\text{L}} + \frac{\partial(\langle p \rangle_{\text{L}} \phi)}{\partial x_i} \\
&= -\phi \frac{\partial \langle p \rangle_{\text{L}}}{\partial x_i} + \frac{\partial(\langle p \rangle_{\text{L}} \phi)}{\partial x_i} \\
&= p^{\text{L}} \frac{\partial \phi}{\partial x_i},
\end{aligned} \tag{15}$$

where the relationship

$$\left\langle \frac{\partial p}{\partial \xi_i} \right\rangle_{\text{L}} = \frac{\partial \langle p \rangle_{\text{L}}}{\partial x_i} \tag{16}$$

is assumed to obtain the 4th equation. The validity of this assumption is checked as follows. In the same manner as Eq. (8), we obtain

$$\frac{\partial \langle p \rangle_{\text{L}}}{\partial x_i} = \frac{-\langle p \rangle_{\text{L}}}{\phi} \frac{\partial \phi}{\partial x_i} + \left\langle \frac{\partial p}{\partial \xi_i} \right\rangle_{\text{L}} - \frac{1}{V\phi} \int_{\Sigma_{\text{red}}} p(\xi) n_i(\xi) dS_\xi. \tag{17}$$

From Eqs. (13) and (17),

$$\frac{\partial \langle p \rangle_{\text{L}}}{\partial x_i} = \left\langle \frac{\partial p}{\partial \xi_i} \right\rangle_{\text{L}} - \frac{1}{V\phi} \int_{\Sigma_{\text{red}}} (p(\xi) - \langle p \rangle_{\text{L}}) n_i(\xi) dS_\xi \tag{18}$$

is obtained. Equation (18) shows that the phasic average  $\langle \cdot \rangle_{\text{L}}$  is not exchangeable with the differential operator, and that Eq. (16) is not valid if the 2nd term in the RHS of Eq. (18) is non-negligible, that is, if the liquid pressure at the boundary with the solid phase is systematically different from the average. In solid-liquid composite systems, the stress heterogeneity within the liquid phase relaxes quickly and the liquid pressure is usually considered to be uniform in REV. Therefore, the 2nd term in the RHS of (18) is considered to be negligible. This also confirms the validity of an assumption implicitly used in Figs. 3 and 4: because the liquid pressure is homogeneous at the microscopic scale, then from the continuity of stress, the normal compressive stress of the solid phase at the solid-liquid interface is equal to the macroscopic liquid pressure  $p^{\text{L}}$ .

The inexchangeability between the differential operator and phasic average is described by Eq. (14) for the solid phase and by Eq. (18) for the liquid phase. Although the second term in the RHS of Eq. (14) was considered to be non-negligible, the corresponding term in Eq. (18) was considered to be negligible. The different treatments applied to the two phases are based on the fact that the stress heterogeneity within the liquid phase relaxes quickly but that the stress heterogeneity within the solid phase does not relax (elastic solid phase) or relaxes much more slowly (viscous solid phase). As can be seen from Fig. 4b, under nonzero effective stress, the traction applied to the surface of each solid grain is significantly different between the areas in contact with the liquid phase and the areas in contact with the neighboring grains, indicating that the microscopic stress field in each grain is highly heterogeneous. Such a heterogeneous stress field causes a systematic deviation of the microscopic displacement at the boundary with the liquid, making the second term in the RHS of Eq. (14) non-negligible.

For the solid phase, it is therefore important to recognize a possible difference between local and macroscopic stresses. However, there seems to be a confusion in some studies considering the effect of the solid-liquid interfacial tension  $\gamma_{\text{sl}}$ . When  $\gamma_{\text{sl}}$  is taken into account, the stress continuity condition required at the solid-liquid interface is replaced by the Laplace condition, e.g. [38]. Hence, the solid stress used in the Laplace condition is the local stress, which is locally determined by  $p^{\text{L}}$ ,  $\gamma_{\text{sl}}$ , and interfacial mean curvature, regardless of the macroscopic solid stress  $\sigma_{ij}^{\text{S}}$ . I emphasize this point because in the previous studies,

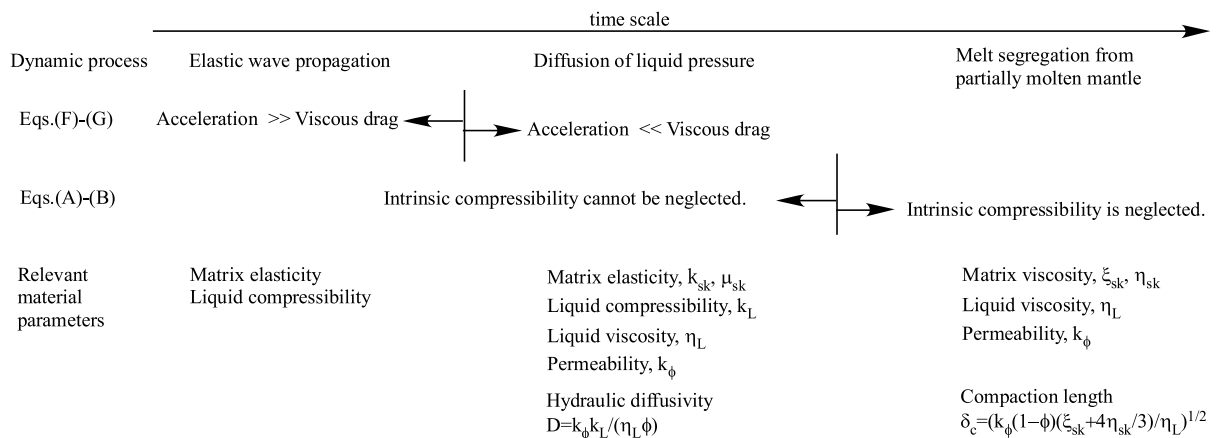
several questionable results were obtained due to the improper use of  $\sigma_{ij}^S$  in the Laplace condition, e. g. [38,46].

### Overview of Applications

The theoretical framework introduced in Sect. “General Theoretical Framework to Describe the Dynamics of Solid-Liquid Composite Systems” is applicable to various geophysical and geological phenomena occurring in solid-liquid two-phase systems. Typical applications are summarized in Fig. 5 by classifying processes into three categories based on their time scales. The propagation of elastic waves is investigated using the wave equations derived from Eqs. (A)–(G). The application to elastic wave propagation is discussed in detail in the following sections. The time-dependent evolution of a perturbation in liquid pressure in the porous media, and the interaction between liquid pressure and matrix deformation or fracture, have long been of interest in investigations of the possible occurrence of earthquakes due to dehydration, e. g. [51] and earthquake triggering, e. g. [16]. These processes have much longer time scales than the periods of elastic waves, and the acceleration terms in Eqs. (F) and (G), which play an essential role in the wave equations, are negligible compared to the viscous drag force included in the interaction term I. Then, as is well known, the evolution equation for liquid pressure is derived in the form of a diffusion equation, where the diffusivity is given by  $D = k_\phi k_L / (\eta_L \phi)$  (hydraulic diffusivity).

Melt segregation from a partially molten mantle has been of great interest in volcanology, petrology, and

geochemistry. This process occurs over much longer time scales than the processes mentioned above, and involves large viscous deformations. Accordingly, the intrinsic compressibilities of the constituent materials are neglected. By applying the governing equations (A)–(D), (E<sub>v</sub>), and (F)–(G) to a one-dimensional column of partially molten mantle with a homogeneous porosity distribution at  $t = 0$  ( $\phi = \phi_0$  at  $z \geq -H$  and  $\phi = 0$  at  $z < -H$ ), the initial stage of melt segregation was solved analytically [20]. In most of the column, the buoyancy force  $(1 - \phi)(\rho^S - \rho^L)g$  is in equilibrium with the viscous drag force  $\eta_L \phi (\dot{u}_z^L - \dot{u}_z^S) / k_\phi$ , while within the compaction length  $\delta_c$  from the bottom ( $-H \leq z < -H + \delta_c$ ), the buoyancy force balances with the compaction resistance of the solid framework. Therefore, when the compaction length  $\delta_c = \sqrt{k_\phi (1 - \phi) (\xi_{sk} + 4\eta_{sk}/3) / \eta_L}$  is smaller than  $H$ , which is the case for the mantle, the segregation velocity is determined by the permeability  $k_\phi$ . The permeability control is more clear in the steady-state model, in which the steady-state porosity structure develops to satisfy the balance between melt production rate and melt segregation rate, which is often assumed for the decompressional melting of the upwelling mantle below a ridge [30]. However, several instabilities inherent to the solid-liquid systems that result in a time-dependent evolution of the porosity distribution have also been reported. These include the propagation of melt as solitary waves or porosity waves, e. g. [3,34], and the unstable evolution of the perturbation in melt fraction under pure shear deformation of the solid matrix [39], in which both melt migration and matrix deformation are involved. These phenomena are de-



Earth's Crust and Upper Mantle, Dynamics of Solid-Liquid Systems in, Figure 5  
 Dynamic processes in solid-liquid composite systems. Typical assumptions adopted in applying the governing equations (A)–(G) in the text to these processes are shown, with the material properties relevant to the processes

scribed by Eqs. (A)–(D), (E<sub>v</sub>), and (F)–(G), where, in the solitary or porosity waves, the nonlinearity caused by the dependence of permeability on melt fraction ( $k_\phi \propto \phi^{n>1}$ ) plays an essential role and, in the latter instability, the dependence of  $\eta_{sk}$  on  $\phi$  plays an essential role. The possible occurrence of these instabilities has been of great interest, because the melt ascending velocity is significantly affected by the spatial distribution of porosity. The melt velocity and its spatial distribution determine the degree of chemical interaction between the melt and host rocks, and thus influence the major and trace element compositions. The microstructural dependences of permeability and viscous properties of the partially molten rocks affecting these instabilities are poorly understood and are the subject of future studies. The basic framework introduced in Sect. “General Theoretical Framework to Describe the Dynamics of Solid-Liquid Composite Systems” can be further extended to take into account additional factors, such as the chemical interaction between the fluid and host rocks, e. g. [1,36] or interfacial tension e. g. [32].

## Elastic Wave Propagation in a Solid-Liquid Composite System

### Derivation of Wave Equations

In this section, the general theoretical framework introduced in Sect. “General Theoretical Framework to Describe the Dynamics of Solid-Liquid Composite Systems” is applied to the elastic wave propagation in a solid-liquid composite system. The governing equations are linearized by considering the infinitesimal strain and displacement of a macroscopically homogeneous medium, and the wave equations are derived. The linearized equations are shown to be equivalent to the basic equations used in the theory of linear poroelasticity.

When we consider a macroscopically homogeneous solid-liquid composite system, the spatial and temporal variations in  $\phi$ ,  $\rho^L$ , and  $\rho^S$  are caused by the displacements  $\mathbf{u}^L$  and  $\mathbf{u}^S$ . Therefore, if  $\mathbf{u}^L$  and  $\mathbf{u}^S$  are infinitesimally small, such terms as  $\dot{\mathbf{u}}^L \cdot \nabla(\phi\rho^L)$  and  $\dot{\mathbf{u}}^S \cdot \nabla((1-\phi)\rho^S)$  are negligible as higher-order terms. Under these approximations, by substituting Eqs. (C) and (D) into (A)/ $\rho^L$ +(B)/ $\rho^S$ , and taking the time integration, we obtain

$$-\phi \nabla \cdot (\mathbf{u}^L - \mathbf{u}^S) = \frac{\phi}{k_L} p^L + \frac{1-\phi}{k_S} p^S + \nabla \cdot \mathbf{u}^S. \quad (19)$$

By using Eq. (E<sub>e</sub>),  $\nabla \cdot \mathbf{u}^S = \epsilon_{kk}$  in Eq. (19) can be expressed in terms of stresses. Then, Eq. (19) and Eq. (E<sub>e</sub>) are written

as

$$\begin{aligned} & \phi \nabla \cdot (\mathbf{u}^L - \mathbf{u}^S) \\ &= \phi \left( \frac{1}{k_S} - \frac{1}{k_L} \right) p^L + (1-\phi) \left( \frac{1}{K_b} - \frac{1}{k_S} \right) (p^S - p^L) \end{aligned} \quad (20)$$

$$\begin{aligned} \epsilon_{ij} &= \frac{(1-\phi)}{2N} \left( \sigma_{ij}^S - \frac{\sigma_{kk}^S}{3} \delta_{ij} \right) \\ &\quad - \frac{(1-\phi)}{3K_b} (p^S - p^L) \delta_{ij} - \frac{1}{3k_S} p^L \delta_{ij}, \end{aligned} \quad (21)$$

where  $K_b$  and  $N$  represent the bulk and shear moduli, respectively, of the skeleton and are related to  $k_{sk}$  and  $\mu_{sk}$  introduced in Eq. (E<sub>e</sub>) as  $K_b = (1-\phi)k_{sk}$  and  $N = (1-\phi)\mu_{sk}$ . In the theory of linear poroelasticity,  $K_b$  and  $N$ , rather than  $k_{sk}$  and  $\mu_{sk}$ , are commonly used. Similarly, by substituting Eq. (3) into Eqs. (F) and (G), and neglecting the effect of gravity, we obtain

$$\phi \rho^L \ddot{\mathbf{u}}^L = -\phi \nabla \cdot p^L - \frac{\eta_L \phi^2}{k_\phi} (\dot{\mathbf{u}}^L - \dot{\mathbf{u}}^S) \quad (22)$$

$$(1-\phi) \rho^S \ddot{\mathbf{u}}^S = (1-\phi) \nabla \cdot \sigma^S + \frac{\eta_L \phi^2}{k_\phi} (\dot{\mathbf{u}}^L - \dot{\mathbf{u}}^S). \quad (23)$$

Equations (20)–(23) are equivalent to the basic equations used in the theory of linear poroelasticity. Compared to the framework introduced in Sect. “General Theoretical Framework to Describe the Dynamics of Solid-Liquid Composite Systems”, the number of governing equations is reduced from 7 to 4, because the variables  $\rho^L$  and  $\rho^S$  are eliminated and  $\phi$  in Eqs. (20)–(23) can be treated as constant. In the theory of linear poroelasticity, Eq. (20) is called the constitutive relation for the relative motion between the two phases, and this relation is usually introduced empirically, e. g. [48]. The present derivation from the more general framework shows that Eq. (20) is based on the requirement of mass conservation and intrinsic constitutive relations. When the acceleration terms are negligible, Eqs. (22) and (23) are further rewritten as

$$\phi (\dot{\mathbf{u}}^L - \dot{\mathbf{u}}^S) = -\frac{k_\phi}{\eta_L} \nabla p^L \quad (24)$$

$$\nabla \cdot \sigma^B = 0, \quad (25)$$

where Eq. (24) represents Darcy's law, and the bulk stress  $\sigma_{ij}^B$  represents  $\sigma_{ij}^B = (1-\phi)\sigma_{ij}^S - \phi p^L \delta_{ij}$ .

By eliminating pressures and stresses from Eqs. (20)–(23), we obtain

$$\begin{aligned} (1-\phi) \rho^S \ddot{\mathbf{u}}^S &= P \nabla (\nabla \cdot \mathbf{u}^S) - N \nabla \times \nabla \times \mathbf{u}^S \\ &\quad + Q \nabla (\nabla \cdot \mathbf{u}^L) + \frac{\eta_L \phi^2}{k_\phi} (\dot{\mathbf{u}}^L - \dot{\mathbf{u}}^S) \end{aligned} \quad (26)$$

$$\phi \rho^L \ddot{\mathbf{u}}^L = Q \nabla(\nabla \cdot \mathbf{u}^S) + R \nabla(\nabla \cdot \mathbf{u}^L) - \frac{\eta_L \phi^2}{k_\phi} (\dot{\mathbf{u}}^L - \dot{\mathbf{u}}^S), \quad (27)$$

where  $P$ ,  $Q$ , and  $R$  are given by

$$\begin{cases} P = K_b + \frac{4}{3}N + \frac{(1 - \phi - \frac{K_b}{k_S})^2 k_S}{1 - \phi - \frac{K_b}{k_S} + \phi \frac{k_S}{k_L}} \\ Q = \frac{\phi(1 - \phi - \frac{K_b}{k_S})k_S}{1 - \phi - \frac{K_b}{k_S} + \phi \frac{k_S}{k_L}} \\ R = \frac{\phi^2 k_S}{1 - \phi - \frac{K_b}{k_S} + \phi \frac{k_S}{k_L}} \end{cases} \quad (28)$$

By taking the curl of Eqs. (26) and (27), and using the expressions  $\Omega_S = \nabla \times \mathbf{u}^S$  and  $\Omega_L = \nabla \times \mathbf{u}^L$ , we obtain wave equations for the shear component;

$$\begin{cases} (1 - \phi)\rho^S \ddot{\Omega}_S = N \nabla^2 \Omega_S + \frac{\eta_L \phi^2}{k_\phi} (\dot{\Omega}_L - \dot{\Omega}_S) \\ \phi \rho^L \ddot{\Omega}_L = -\frac{\eta_L \phi^2}{k_\phi} (\dot{\Omega}_L - \dot{\Omega}_S) \end{cases} \quad (29)$$

By taking the divergence of Eqs. (26) and (27), and using the expressions  $e_S = \nabla \cdot \mathbf{u}^S$  and  $e_L = \nabla \cdot \mathbf{u}^L$ , we obtain wave equations for the longitudinal component;

$$\begin{cases} (1 - \phi)\rho^S \ddot{e}_S = P \nabla^2 e_S + Q \nabla^2 e_L + \frac{\eta_L \phi^2}{k_\phi} (\dot{e}_L - \dot{e}_S) \\ \phi \rho^L \ddot{e}_L = Q \nabla^2 e_S + R \nabla^2 e_L - \frac{\eta_L \phi^2}{k_\phi} (\dot{e}_L - \dot{e}_S) \end{cases} \quad (30)$$

The elastic wave propagation in a solid-liquid composite system was first formulated by Biot [5,6]. The wave equations (29)–(30) are almost the same as those obtained by Biot [5,6], except for the acceleration terms, which are slightly different. This is because the interaction  $\mathbf{I}$  given in Eq. (3) does not take into account the effect of the relative acceleration between the solid and liquid phases. If an additional term proportional to the relative acceleration is added to the RHS of Eq. (3), then the same equations as Biot [5,6] can be derived. The proportionality constant attached to the relative acceleration is called tortuosity; tortuosity represents the deviation from the straight pore channel.

The elastic waves obtained by solving Eqs. (29)–(30) are dispersive and dissipative due to the relative motion between the solid and liquid phases. However, as shown below in Sect. “Dispersion and Attenuation of Waves in Solid-Liquid Composite Systems”, the characteristic frequency for the dispersion and attenuation is much higher than the seismic frequency range. Therefore, in predicting

the seismic wave velocities, the solutions obtained at the low-frequency limit are of importance. The wave solutions at the low-frequency limit do not involve relative motions, because the velocity terms,  $\dot{\Omega}_L - \dot{\Omega}_S$  and  $\dot{e}_L - \dot{e}_S$ , if any, dominate the acceleration terms, and hence are not dispersive nor dissipative. The longitudinal and shear wave velocities at the low-frequency limit are obtained as

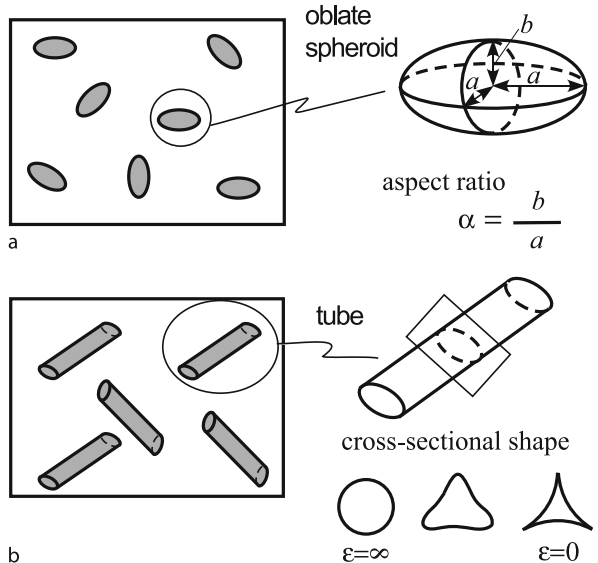
$$V_P = \sqrt{\frac{K_b + \frac{4}{3}N + \frac{k_S(1 - K_b/k_S)^2}{1 - \phi - K_b/k_S + \phi k_S/k_L}}{\bar{\rho}}} \quad (31)$$

$$V_S = \sqrt{\frac{N}{\bar{\rho}}}, \quad (32)$$

where  $\bar{\rho} = (1 - \phi)\rho^S + \phi\rho^L$  represents the average density of the medium. Without relative motion, neither permeability nor tortuosity affect the velocities. Therefore, Eqs. (31)–(32) are exactly the same as the results of Biot [5,6].

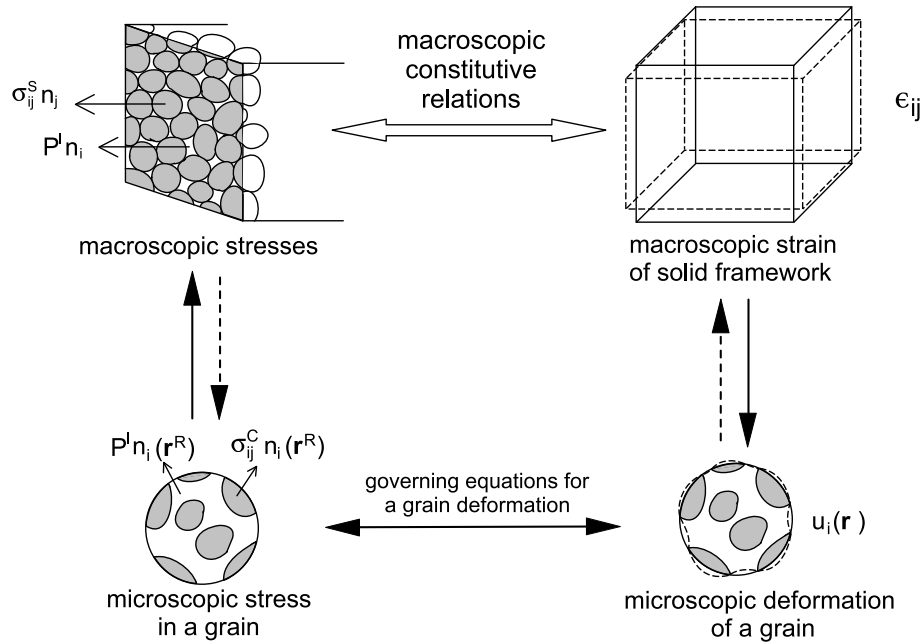
### Porosity and Pore Shape

Because the bulk and shear moduli of the skeleton,  $K_b$  and  $N$ , included in Eqs. (31)–(32) depend not only on porosity but also on pore geometry, various models assuming various pore geometries have been developed to predict  $K_b$  and  $N$  quantitatively, e. g. [15,17,18,27,41]. The oblate spheroid model (Fig. 6a), tube model (Fig. 6b),



Earth's Crust and Upper Mantle, Dynamics of Solid-Liquid Systems in, Figure 6  
**Inclusion models of a solid-liquid composite system. a Oblate spheroid model. b Tube model**





Earth's Crust and Upper Mantle, Dynamics of Solid-Liquid Systems in, Figure 7  
 Granular model of a solid-liquid composite system, showing the procedures to derive macroscopic constitutive relation, Eq. (E), based on the microscopic deformation of each grain [41]

granular model (Fig. 7), and crack model are four representative models in which analytical results can be obtained for  $K_b$  and  $N$  (Table 1). All but the granular model are inclusion models in which the liquid phase is modeled by inclusions contained in a continuum solid phase, and  $K_b$  and  $N$  are derived based on the effective medium theories. In the granular model, the constitutive equation (E) is derived directly. The results from the different theories can be compared based on Eq. (11). Although the connectivity of the liquid phase is not guaranteed in the inclusion models, when considering waves in the low-frequency regime where relative motion between solid and liquid does not occur, connectivity of the liquid phase is not important.

Here, by assuming a random orientation and homogeneous distribution of the pores, the macroscopic properties are assumed to be isotropic. Then,  $K_b$  and  $N$  are derived as functions of the porosity  $\phi$  and aspect ratio  $\alpha$  ( $\alpha$ =short radius/long radius) for the oblate spheroid model, as functions of the porosity  $\phi$  and parameter  $\varepsilon$  for the tube model, where  $\varepsilon$  represents the cross-sectional tube shape (Fig. 6b), as functions of the contiguity  $\varphi$  for the granular model, and as functions of the crack density parameter  $\kappa$  for the crack model. The contiguity  $\varphi$  used in the granular model is defined by the ratio of the grain-to-grain contact area relative to the total surface area of each grain;

thus,  $\varphi = 0$  when there is no grain-to-grain contact, and  $\varphi = 1$  when there is no liquid or pore phase. The crack density parameter  $\kappa$  used in the crack model is defined by  $\kappa = n_\kappa a_\kappa^3$ , where  $a_\kappa$  represents the radius of the circular crack and  $n_\kappa$  represents number density. Walsh [47] showed that in the limit of small aspect ratio,  $K_b$  and  $N$  obtained from the oblate spheroid model depend only on the crack density parameter  $\kappa = 3\phi/(4\pi\alpha)$ , and the results of the oblate spheroid model and crack model become equivalent. Therefore, the crack model can be included in the oblate spheroid model as a special case of small aspect ratio.

In the granular model, the dependence of contiguity  $\varphi$  on porosity  $\phi$  first needs to be assessed in order to specify the dependences of  $K_b$  and  $N$  on the porosity  $\phi$ . For a random packing of elastic spheres, which is often used as a model of soil, the total area of the elastic contacts of the spheres increases with raising confining pressure, and  $\varphi$  and  $\phi$  are derived as functions of confining pressure [8]. However, when the temperature is higher than a few hundreds °C, such an elastic model is not realistic. In the deep crust and mantle, solid grains are single crystals and the contiguity is related to the area of (liquid-free) grain boundaries. Because the grain boundary and crystal-liquid interface both have interfacial energies, there ex-

Earth's Crust and Upper Mantle, Dynamics of Solid-Liquid Systems in, Table 1  
Microstructural Models for Solid-Liquid Composites

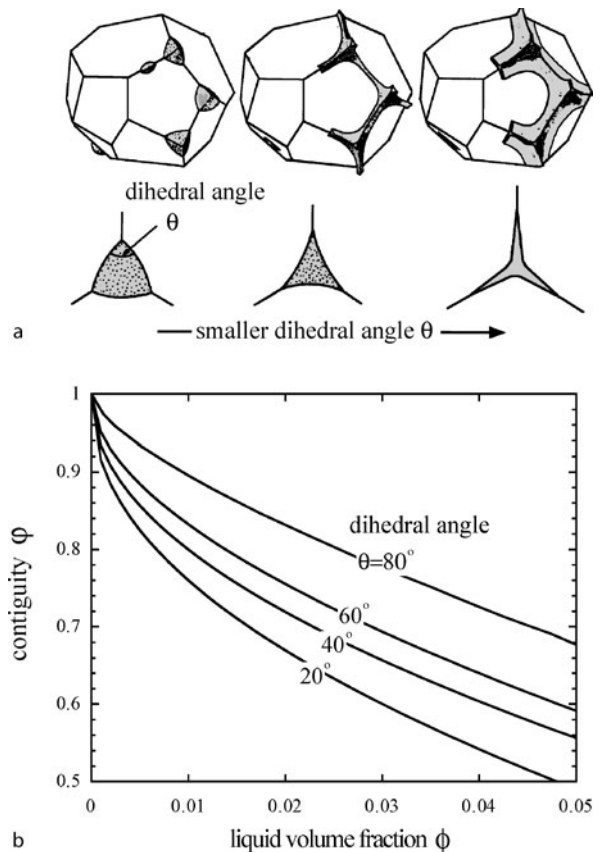
Model	Structural parameters		References
Oblate spheroid	Porosity $\phi$	Aspect ratio $\alpha$	e. g., [4] <sup>1</sup>
Tube	Porosity $\phi$	Tube geometry $\varepsilon$	[17]
Granular	Contiguity $\varphi$		[41]
Equilibrium geometry <sup>2</sup>	Porosity $\phi$	Dihedral angle $\theta$	[41]
Crack	Crack density parameter $\kappa$		[27]

1: Typographical errors in the former studies were corrected.

2: Equilibrium geometry model is a special case of granular model.

ists an equilibrium shape of the liquid phase which minimizes the total interfacial energy of the system. The equilibrium shapes were actually observed in the high T and high P experiments for various rock + melt and rock + aqueous fluid systems e. g. [10]. Therefore, the relationship between contiguity  $\varphi$  and porosity  $\phi$  is derived by assuming the equilibrium shape of the liquid phase. The granular model under this assumption is called the equilibrium geometry model (Table 1). Under a given liquid volume fraction  $\phi$ , the equilibrium shape is controlled by the dihedral angle  $\theta$ , which is determined by the grain boundary energy  $\gamma_{ss}$  and crystal-liquid interfacial energy  $\gamma_{sl}$  as  $\gamma_{ss}/\gamma_{sl} = 2 \cos(\theta/2)$  (Fig. 8a). The theoretical results of von Bargen and Waff [45] show that under a given  $\phi$ , the equilibrium contiguity is smaller for smaller  $\theta$  (Fig. 8b). By substituting  $\varphi$  obtained as functions of  $\phi$  and  $\theta$  into the results of the granular model,  $K_b$  and  $N$  in the equilibrium geometry model can be derived as functions of  $\phi$  and  $\theta$ . Most rock + melt systems have  $\theta$  between 20–40° and most rock + aqueous fluid systems have  $\theta$  between 40–100° [10]. When  $\theta \leq 60^\circ$ , a connected liquid network develops along the grain edges at  $\phi > 0$ . Although the tube model considers such grain edge tubules, the parameter  $\varepsilon$  in the tube model cannot be quantitatively related to the dihedral angle  $\theta$ .

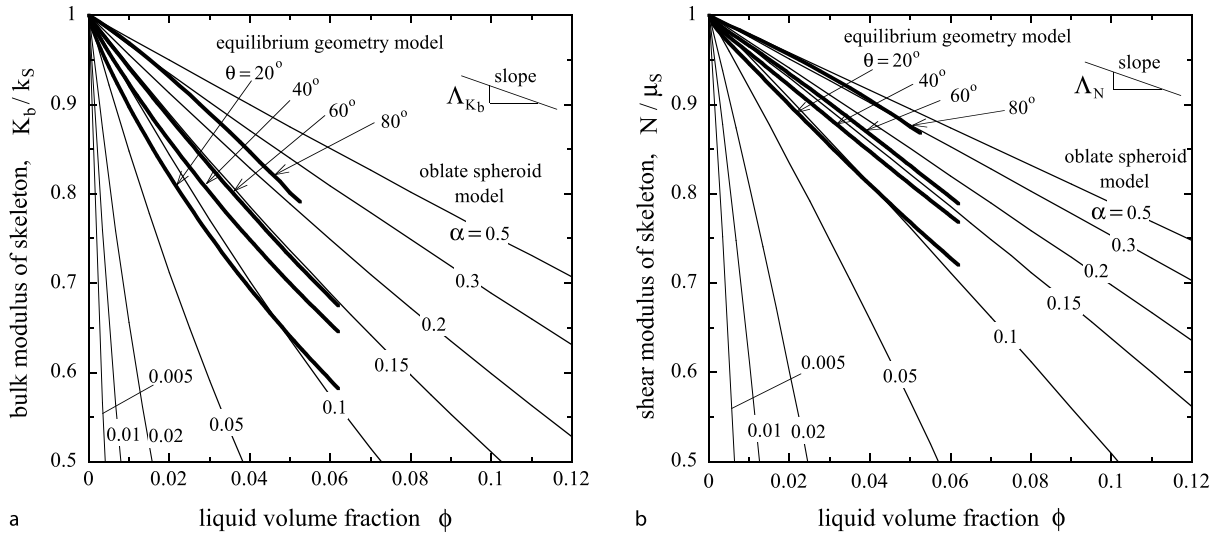
Now, solid-liquid composite systems are characterized in terms of the two parameters: porosity  $\phi$  and pore shape  $x$  ( $= \alpha, \varepsilon$ , or  $\theta$ ). Because a random orientation and homogeneous distribution of the pores are assumed for simplicity, and since  $K_b$  and  $N$  do not depend on pore size, pore geometry is parameterized only by the shape. Porosity  $\phi$  and pore shape  $x$  are generally not dependent but can vary independently governed by different physics. For example, in a texturally equilibrated system, the dihedral angle  $\theta$  is determined by thermodynamic conditions such as temperature, pressure, and chemical compositions, whereas  $\phi$  can vary mechanically through flow-in or flow-out of the liquid. In a rock + water system stressed



Earth's Crust and Upper Mantle, Dynamics of Solid-Liquid Systems in, Figure 8

**a** Equilibrium geometry of the liquid phase characterized by dihedral angle  $\theta$ . **b** Contiguity  $\varphi$  versus liquid volume fraction  $\phi$  calculated theoretically for the equilibrium geometry with dihedral angle  $\theta$

under undrained condition, the pore shape can vary by fracture, while  $\phi$  is kept constant. In most solid-liquid systems in the Earth, neither the porosity nor the pore shape are known.



Earth's Crust and Upper Mantle, Dynamics of Solid-Liquid Systems in, Figure 9

**a** Bulk modulus  $K_b$  and **b** shear modulus  $N$  of skeleton versus porosity  $\phi$ , for the oblate spheroid model with various aspect ratios  $\alpha$  and for the equilibrium geometry model with various dihedral angles  $\theta$

Because the parameter describing the pore shape is different in different models, it seems difficult to investigate the effects of pore shapes systematically. However, introduction of the concept of “equivalent aspect ratio” enables us to treat various pore shapes systematically. When the porosity  $\phi$  is small,  $K_b$  and  $N$  are closely approximated by linear functions of  $\phi$ ,

$$\begin{cases} \frac{K_b}{k_s}(\phi, x) = 1 - \phi \Lambda_{K_b}(x) \\ \frac{N}{\mu_s}(\phi, x) = 1 - \phi \Lambda_N(x) \end{cases} \quad (x = \alpha, \varepsilon, \theta), \quad (33)$$

where  $k_s$  and  $\mu_s$  represent the intrinsic bulk and shear moduli, respectively, of the solid, and the proportionality coefficients  $\Lambda_{K_b}$  and  $\Lambda_N$  are functions of pore shape  $x$  (Fig. 9). In other words, the effects of porosity and pore shape on  $K_b$  and  $N$  can be separated in such simple forms as given in Eq. (33), in which the pore shape given by  $x$  is characterized in terms of two parameters  $\Lambda_{K_b}$  and  $\Lambda_N$ . If a tube model with  $\varepsilon$  (or an equilibrium geometry model with  $\theta$ ) has almost the same values of  $\Lambda_{K_b}$  and  $\Lambda_N$  as the oblate spheroid model with  $\alpha$ ,

$$\begin{cases} \Lambda_{K_b}(x) \simeq \Lambda_{K_b}(\alpha) \\ \Lambda_N(x) \simeq \Lambda_N(\alpha) \end{cases} \quad (x = \varepsilon, \theta), \quad (34)$$

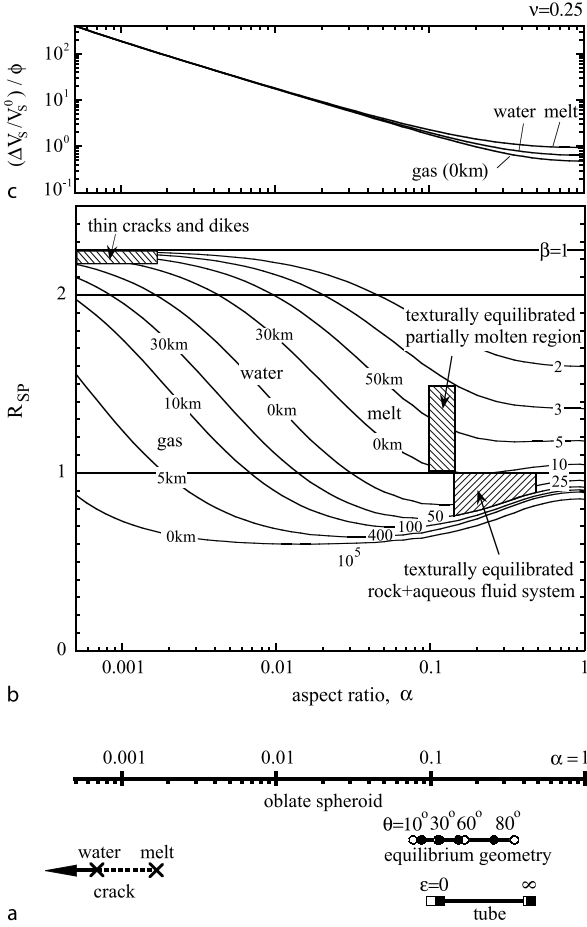
these two models yield almost the same values of  $K_b$  and  $N$  for a given  $\phi$ . This also means that  $V_P$  and  $V_S$  calculated with Eqs. (31)–(32) are almost the same in these two mod-

els. Therefore, the value of  $\alpha$  satisfying Eqs. (34) is called the equivalent aspect ratio of  $x$ ; this aspect ratio guarantees the equivalence between the different pore shapes in predicting  $V_P$  and  $V_S$ .

Figure 10a shows the relationships between the different models in terms of equivalent aspect ratio. The solid and open symbols represent the equivalent aspect ratios determined from  $\Lambda_{K_b}$  and  $\Lambda_N$ , respectively. The difference between these symbols is small, indicating that, in a practical sense, one value of equivalent aspect ratio satisfying both equations in (34) can be determined. Figure 10a shows that the tube model with  $\varepsilon = 0$ , equilibrium geometry model with  $\theta = 30^\circ$ , and oblate spheroid model with  $\alpha = 0.1$  are all equivalent. Rigorously speaking,  $\Lambda_{K_b}$  and  $\Lambda_N$  depend on the intrinsic Poisson's ratio  $\nu$  of the solid phase. The results shown in Fig. 10a are calculated for  $\nu = 0.25$ . Fortunately, however, the effects of  $\nu$  are almost the same in all models, and the equivalent aspect ratio can be determined almost independently of  $\nu$ . The present method to determine the equivalent aspect ratio from Eqs. (33)–(34) is applicable to general isotropic solid-liquid systems. By using the equivalent aspect ratio, general pore geometries can be treated systematically.

#### Determinability of Porosity and Pore Shape from Elastic Wave Velocities

It can be shown that when the porosity  $\phi$  is small, the effects of  $\phi$  on the skeleton properties  $K_b$  and  $N$  can be



Earth's Crust and Upper Mantle, Dynamics of Solid-Liquid Systems in, Figure 10

**a** Equivalence of the equilibrium geometry model, tube model, and crack model to the oblate spheroid model is shown by the equivalent aspect ratio  $\alpha$ . Solid and open symbols are the equivalent aspect ratios determined from  $\Lambda_{K_b}$  and  $\Lambda_N$ , respectively. **b**  $R_{SP}$ , representing the ratio between  $V_S$  and  $V_P$  perturbations,  $(\Delta V_S/V_S^0)/(\Delta V_P/V_P^0)$ , versus pore aspect ratio  $\alpha$ , for various fluid compressibilities  $\beta = k_S/k_L$ . **c** Proportionality constant between  $V_S$  perturbation  $\Delta V_S/V_S^0$  and porosity  $\phi$  versus pore aspect ratio  $\alpha$ , for various liquid types (gas, water, and melt)

closely approximated by linear functions of  $\phi$ . In the same manner, when  $\phi$  is small, the effects of  $\phi$  on the velocities  $V_P$  and  $V_S$  can be closely approximated by linear functions of  $\phi$ . Let  $\Delta V_P = V_P^0 - V_P$  and  $\Delta V_S = V_S^0 - V_S$  be reductions in  $V_P$  and  $V_S$ , respectively, caused by liquid-filled pores, where  $V_P^0 = \sqrt{(k_S + 4\mu_S/3)/\rho^S}$  and  $V_S^0 = \sqrt{\mu_S/\rho^S}$  represent the intrinsic elastic wave velocities of the solid phase. Let  $\Delta V_P/V_P^0$  and  $\Delta V_S/V_S^0$  be perturbations in  $V_P$  and  $V_S$ , respectively. By substituting Eqs. (33) into Eqs. (31)–(32) and neglecting higher-order terms in  $\phi$  ( $\phi^n$

with  $n \geq 2$ ), we obtain

$$\begin{cases} \frac{\Delta V_P}{V_P^0} = \left[ \frac{(\beta-1)\Lambda_{K_b} + \frac{4}{3}\gamma\Lambda_N}{1 + \frac{4}{3}\gamma} - \left(1 - \frac{\rho^L}{\rho^S}\right) \right] \frac{\phi}{2} \\ \frac{\Delta V_S}{V_S^0} = \left[ \Lambda_N - \left(1 - \frac{\rho^L}{\rho^S}\right) \right] \frac{\phi}{2}, \end{cases} \quad (35)$$

where  $\beta = k_S/k_L$  and  $\gamma = \mu_S/k_S$ . Without loss of generality,  $\Lambda_{K_b}$  and  $\Lambda_N$  can be treated as functions of the equivalent aspect ratio  $\alpha$ . Equations (35) demonstrate that the velocity perturbations are affected by the five non-dimensional parameters  $\phi$ ,  $\alpha$ ,  $\beta$ ,  $\gamma$ , and  $\rho^L/\rho^S$ . Because a possible variation in  $\gamma$  in response to a variation in the intrinsic Poisson's ratio  $\nu$  of the solid phase is small,  $\gamma$  can be fixed to 0.6 ( $\nu = 0.25$ ). Also, as shown below, the effect of  $\rho^L/\rho^S$  on the perturbations is small. Therefore, in a practical sense, the velocity perturbations are controlled by the three non-dimensional factors: liquid volume fraction  $\phi$ , pore aspect ratio  $\alpha$ , and liquid compressibility  $\beta$ . If only one of  $\Delta V_P/V_P^0$  and  $\Delta V_S/V_S^0$  is known,  $\phi$  cannot be determined without knowing  $\alpha$  (and  $\beta$ ). However, if both  $\Delta V_P/V_P^0$  and  $\Delta V_S/V_S^0$  are known, significant constraints can be placed on  $\phi$ ,  $\alpha$ , and/or  $\beta$ . A practical method to obtain these constraints is presented below.

First, we introduce  $R_{SP}$ , representing the ratio of the perturbations in  $V_S$  and  $V_P$ . From Eq. (35),  $R_{SP}$  is written as

$$\begin{aligned} R_{SP} &= \frac{\Delta V_S/V_S^0}{\Delta V_P/V_P^0} \\ &= \frac{\Lambda_N - \left(1 - \frac{\rho^L}{\rho^S}\right)}{\frac{(\beta-1)\Lambda_{K_b} + \frac{4}{3}\gamma\Lambda_N}{1 + \frac{4}{3}\gamma} - \left(1 - \frac{\rho^L}{\rho^S}\right)}. \end{aligned} \quad (36)$$

$R_{SP}$  can be closely related to the  $V_P/V_S$  ratio frequently used in seismology: when  $R_{SP} < 1$ , the perturbation (reduction positive) in  $V_P$  is larger than that in  $V_S$  and hence the  $V_P/V_S$  ratio decreases; when  $R_{SP} > 1$ , the perturbation in  $V_S$  is larger than that in  $V_P$  and thus the  $V_P/V_S$  ratio increases. Because  $R_{SP}$  is independent of the liquid volume fraction  $\phi$ , this factor provides a useful insight into the effects of pore shape  $\alpha$  and liquid compressibility  $\beta$  on the velocity perturbations.  $R_{SP}$  is sometimes written as  $d \ln V_S/d \ln V_P$ . Figure 10b shows  $R_{SP}$  versus pore aspect ratio  $\alpha$  for various compressibility  $\beta$  of pore fluids. Values of  $\beta$  are estimated as 5–10 for rock + melt systems, 10–40 for rock + water systems, and 50–10<sup>5</sup> for rock + ideal

## Earth's Crust and Upper Mantle, Dynamics of Solid-Liquid Systems in, Table 2

### Liquid bulk modulus $k_L$

depth, km	$P$ , GPa	$T$ , °C	$k_L$ , GPa		
			gas <sup>1</sup>	water <sup>2</sup>	melt <sup>3</sup>
0	$10^{-4}$	20	$1.3 \times 10^{-4}$ ( $\beta = 10^5 - 10^6$ ) <sup>4</sup>	2.2 (18–50)	7–25 (4–10)
5	0.15	75	0.2 (200–600)	3.1 (13–40)	
10	0.3	150	0.4 (100–300)	1.8 (22–66)	
35	1	500	1.3 (30–100)	4.5 (9–25)	
70	2				20–40 (3–6)

1: Adiabatic bulk modulus estimated by 1.3  $P$ .

2: Isothermal bulk modulus estimated at each  $(P, T)$  condition. Data from Schäfer [33].

3: Data from Stolper et al. [40]. Data at  $P = 2$  GPa are estimated from  $\partial k_L / \partial P = 6 - 7$ .

4: Numerals in the parentheses show  $\beta = k_S / k_L$  evaluated for  $k_S \approx 40 - 120$  GPa

gas systems in the 0–50 km depth range (Table 2). Values of  $\rho^L / \rho^S$  are estimated as 0.92, 0.33, and 0 for rock + melt, rock + water, and rock + ideal gas systems, respectively. The effect of  $\rho^L / \rho^S$  on  $R_{SP}$  is small, and practically the same figure as Fig. 10b can be obtained by simply assuming  $\rho^L / \rho^S = 1$  ([42] Fig. 4). Figure 10b shows that for a given pore shape  $\alpha$ ,  $R_{SP}$  increases with decreasing liquid compressibility  $\beta$ . Figure 10b also shows that for a fixed liquid compressibility  $\beta$ ,  $R_{SP}$  varies significantly with the variation of pore shape  $\alpha$ . When  $\beta$  is fixed to 25, for example,  $R_{SP}$  is smaller than 1 for moderate values of pore aspect ratio ( $\alpha > 0.03$ ), larger than 1 for small aspect ratio ( $< 0.03$ ), and larger than 2 for very small aspect ratio ( $< 0.0016$ ). Therefore,  $R_{SP}$  provides a good seismological indicator of pore shape.

When  $\Delta V_P / V_P^0$  and  $\Delta V_S / V_S^0$  are obtained from seismological observations or laboratory experiments,  $R_{SP}$  is calculated by taking the ratio of these two. If we know whether the liquid phase is melt, water, or gas, Fig. 10b can be used for estimating the equivalent aspect ratio  $\alpha$  from  $R_{SP}$  under known  $\beta$ . Without any additional information about the liquid phase,  $\alpha$  is estimated from  $R_{SP}$  under an assumed  $\beta$ . Figure 10c shows the proportionality coefficient between  $\Delta V_S / V_S^0$  and  $\phi$  (the 2nd equation of 35) versus  $\alpha$  for various liquid types. By applying  $\alpha$  estimated from  $R_{SP}$  to Fig. 10c, the liquid volume fraction  $\phi$  can be determined from  $\Delta V_S / V_S^0$ .

Figures 10a–10c are a complete summary of the effects of liquid volume fraction  $\phi$ , pore shape  $\alpha$ , and liq-

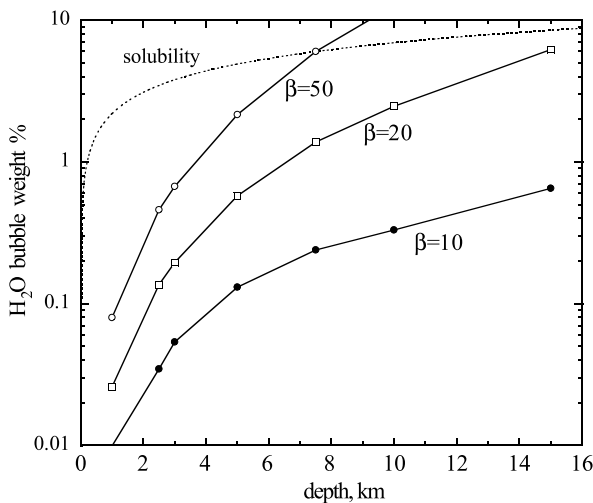
uid compressibility  $\beta$  on  $V_P$  and  $V_S$ . Using these figures, we can combine and confirm our understandings from the previous forward approaches, such as the different effects of gas, water, and melt, on the  $V_P / V_S$  ratio [27,49]. Also, Figs. 10b–10c represent a practical method in the inverse approach to constrain  $\phi$ ,  $\alpha$ , and/or  $\beta$  from the observation of  $V_P$  and  $V_S$ . Figures 10a–10c are based on Eqs. (33), (35), and (36) in which the higher-order terms in  $\phi$  ( $\phi^{n \geq 2}$ ) are neglected. A possible variation of  $R_{SP}$  with  $\phi$  caused by the higher-order terms in  $\phi$  can be obtained by calculating  $(\Delta V_S / V_S^0) / (\Delta V_P / V_P^0)$  directly from Eqs. (31) and (32) without using Eq. (33). Using results from the latter calculations, it is confirmed that the present method based on the linearized equations is valid for  $\Delta V_S / V_S^0 \leq 0.35$  for the equilibrium geometry model, which corresponds to  $\phi \leq 0.15$ , and valid for  $\Delta V_S / V_S^0 \leq 0.2$  for the oblate spheroid, tube, and crack models. The applicable range is large for the equilibrium geometry model because the higher-order effects neglected in obtaining Eq. (33) cancel those neglected in obtaining Eq. (35). Compared to the equilibrium geometry model, the applicable range is smaller for the oblate spheroid, tube, and crack models. To use the present method outside the applicable range, a small modification is required ([42], Appendix B).

### Application to Seismic Tomographic Images

The information on pore geometry obtainable from seismic tomographic data is the equivalent aspect ratio. Whether the actual geometry is an oblate spheroid or tube, for example, cannot be determined from seismological data. A reasonable interpretation of the derived equivalent aspect ratio requires a knowledge of probable pore geometries in the Earth. Two end-member images of pore geometry have been inferred based on experimental and field observations: one is the equilibrium geometry characterized by a dihedral angle, and the other is thin dikes and veins. In the equilibrium geometry, the pore size is much smaller than the grain size and the permeability is small. For rock + water systems, because the dihedral angle is usually larger than  $60^\circ$ , the equilibrium geometry may have nearly zero permeability. Dikes and veins can develop on much larger scales than the grain size and hence may have much larger permeability than the equilibrium geometry. Therefore, the particular geometry realized between these two end-members significantly affects the liquid migration velocity of the buoyancy ascent. As discussed in Sect. “Overview of Applications”, whether the liquid phase is in the equilibrium shape or not can provide us with valuable information about the degree of interaction between pore geometry and macroscopic dynam-

ics. It is therefore desirable to distinguish between these two end-members using seismic tomographic data. Because the equivalent aspect ratio for the equilibrium geometry is about a factor of 100 larger than that for the thin cracks and dikes (Fig. 10a), the expected values of  $R_{SP}$  are significantly different between these two end-member geometries (hatched regions in Fig. 10b); for rock + water systems,  $R_{SP}$  is  $<1$  for the equilibrium geometry and  $>2$  for thin cracks and dikes; for rock + melt systems,  $R_{SP}$  is 1–1.5 for the equilibrium geometry and  $>2$  for thin cracks and dikes. Therefore,  $R_{SP}$  can be used as a seismological indicator of the textural equilibrium, and the information on pore geometry obtained from this indicator can provide us a valuable constraint on actual fluid migration processes in the Earth. Low-velocity regions observed in the mantle wedge beneath Northeastern Japan subduction zone show a systematic change in  $R_{SP}$  with depth. Nakajima et al. [24] applied the method introduced in Sect. “**Determinability of Porosity and Pore Shape from Elastic Wave Velocities**” to seismic tomographic data and inferred a systematic change in pore geometry from an equilibrium geometry at a depth of  $\sim 90$  km to thin cracks and dikes at a depth of  $\sim 65$  km.

Beneath volcanic areas, low-velocity regions with lower  $V_P/V_S$  ratio than the surrounding regions are sometimes observed at depths of several km [22]. Rock + melt systems usually have  $\beta$  smaller than 10. This means that  $R_{SP}$  is larger than 1 regardless of  $\alpha$  (Fig. 10b) so that the



Earth's Crust and Upper Mantle, Dynamics of Solid-Liquid Systems in, Figure 11

The amount of H<sub>2</sub>O bubble (weight %) in melt at which the compressibility of the mixture is equal to  $\beta = 10, 20$ , or  $50$ , is shown as a function of depth

observed low  $V_P/V_S$  ratio cannot be explained by the melt-filled pores. At these shallow depths, however, the melt phase can be a mixture of melt and H<sub>2</sub>O vapor, because H<sub>2</sub>O initially dissolved in the melt in the deeper reaches of the subduction zone starts to exsolve. Here, we briefly discuss such situation, which is not considered in Fig. 10b and Table 2. Because of the high temperature (900–1000°C) of melt, the water phase in the melt is much more compressible than the estimates in Table 2. Figure 11 shows the fraction of water (wt%) above which  $\beta$  of the mixture exceeds 10, 20, or 50. It is shown that at a depth of 3–4 km,  $\beta \geq 20$  occurs for the melt containing 0.2–0.5 wt% water as vapor phase, which is realistic in the subduction zone. Therefore, from Fig. 10b,  $R_{SP} < 1$  can occur at large  $\alpha$  and can explain the observed reduction in the  $V_P/V_S$  ratio. If the H<sub>2</sub>O vapor in the melt phase can be detected by the low  $V_P/V_S$  ratio, we can obtain valuable constraints on the water content of melts and on the evolution of a magma chamber in the crust. However, because melt viscosity is considered to be high in shallow magma chambers, it is important to be careful about wave dispersion, as discussed below.

### Dispersion and Attenuation of Waves in Solid-Liquid Composite Systems

In deriving the elastic wave velocities in Sect. “**Derivation of Wave Equations**”, it was implicitly assumed that the frequencies of the seismic waves are lower than the characteristic frequencies of several relaxation processes inherent to solid-liquid composite systems. If this assumption is not valid, the relaxation processes affect the wave velocities due to dispersion. To assess the applicability of the theoretical results presented in Sect. “**Determinability of Porosity and Pore Shape from Elastic Wave Velocities**”, I present here a brief discussion of such relaxation processes. To show the mutual relationship between dispersion and attenuation, I consider the relaxation mechanism predicted from Eqs. (29)–(30). This mechanism was first studied by Biot [5,6] and is hereafter referred to as the Biot mechanism. The behavior obtained for the Biot mechanism describes the fundamental characteristic of relaxation. Several other relaxation mechanisms inherent to the solid-liquid composites also affect wave propagation and these are further summarized below.

Let  $\omega$  and  $k$  be the angular frequency and wave number, respectively. By substituting the traveling wave solutions  $\Omega_S = \Omega_S^0 e^{-i(\omega t - kx)}$  and  $\Omega_L = \Omega_L^0 e^{-i(\omega t - kx)}$  into Eq. (29), the dispersion relation, under which non-trivial solutions exist, is obtained as

$$\left(\frac{k}{\omega}\right)^2 = \frac{\rho_U}{N} \cdot f(\omega), \quad (37)$$

where the complex function  $f(\omega) = f_1(\omega) + if_2(\omega)$  is explicitly written as

$$\begin{cases} f_1(\omega) = 1 + \frac{\Delta}{1 + \left(\frac{\omega}{\omega_c}\right)^2} \\ f_2(\omega) = \frac{\Delta \cdot \left(\frac{\omega}{\omega_c}\right)}{1 + \left(\frac{\omega}{\omega_c}\right)^2}, \end{cases} \quad (38)$$

and  $\rho_U = (1 - \phi)\rho^S$ ,  $\rho_R = (1 - \phi)\rho^S + \phi\rho^L$ ,  $\Delta = (\rho_R - \rho_U)/\rho_U$ , and  $\omega_c = \eta_L\phi/(k_\phi\rho^L)$ . The phase velocity  $V$  and attenuation  $Q^{-1}$  are defined by  $k/\omega = V^{-1}(1 + i/(2Q))$ . By assuming  $Q^{-1}$  to be small ( $f_2 \ll f_1$ ), we obtain

$$\begin{cases} V = \sqrt{\frac{N}{\rho_U \cdot f_1(\omega)}} \\ Q^{-1} = \frac{f_2(\omega)}{f_1(\omega)}. \end{cases} \quad (39)$$

The phase velocity  $V$  and attenuation  $Q^{-1}$  given by Eqs. (39) are shown in Fig. 12 as functions of the normalized frequency  $\omega/\omega_c$ . Both dispersion and attenuation occur near  $\omega/\omega_c = 1$  and the total amplitude of dispersion and peak value of  $Q^{-1}$  are equal;

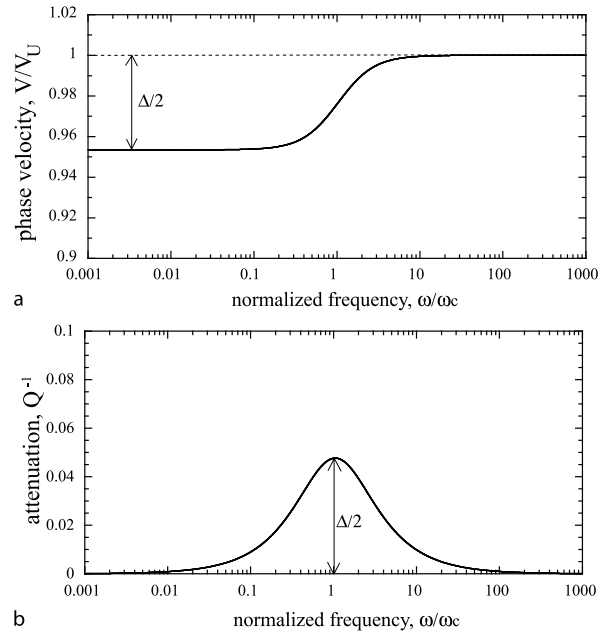
$$Q_{\text{MAX}}^{-1} = \frac{\Delta}{2} = \frac{V_U - V_R}{V_U}, \quad (40)$$

where  $V_U$  (unrelaxed velocity) represents  $V$  at  $\omega/\omega_c \gg 1$  and  $V_R$  (relaxed velocity) represents  $V$  at  $\omega/\omega_c \ll 1$ .

Although obtained for the Biot mechanism, Eqs. (38) and (39) describe the fundamental characteristics of dispersion and attenuation regardless of the individual mechanism. These equations simply mean that the dispersion and attenuation are caused by a relaxation process. When a response  $J(t)$  (e. g., strain) of a medium to a constant unit force (in the form of a Heaviside function  $H(t)$ ) applied at  $t \geq 0$  is not instantaneous but shows a time delay expressed in the form of

$$J(t) = J_U [1 + \Delta \cdot (1 - e^{-\omega_c t})] \cdot H(t), \quad (41)$$

this phenomenon is called relaxation. In other words, Eq. (41) gives a phenomenological model of relaxation. The relaxation process is characterized by relaxation strength  $\Delta$  and relaxation time scale  $\omega_c^{-1}$ . The time derivative of Eq. (41),  $\dot{J} = J_U[\delta(t) + \omega_c \Delta e^{-\omega_c t} H(t)]$ , yields the impulse response, where the Fourier transform of  $\dot{J}/J_U$  is equal to  $f(\omega) = f_1(\omega) + if_2(\omega)$  with  $f_1(\omega)$  and  $f_2(\omega)$  given by Eq. (38). Therefore, Eq. (38), which is called Debye equation [25], is a phenomenological model of relaxation in the frequency domain.



Earth's Crust and Upper Mantle, Dynamics of Solid-Liquid Systems in, Figure 12

**a** Frequency-dependent phase velocity (dispersion) and **b**  $Q^{-1}$  (attenuation) described by Debye equations (38)

Solid-liquid composite systems show several relaxation mechanisms caused by the liquid phase. One example is the Biot mechanism exemplified above, in which the density of the system relaxes from  $\rho_U$  to  $\rho_R$ . At  $\omega/\omega_c \gg 1$ , due to a dominant effect of liquid inertia, a wave field cannot cause motion in the liquid phase. Hence,  $\rho_U$  is associated only with the solid mass. At  $\omega/\omega_c \ll 1$ , due to a dominant effect of viscous drag force, relative motion does not occur between the solid and liquid phases. Hence,  $\rho_R$  is associated with the total mass. The characteristic frequency  $\omega_c$  is estimated as 200 kHz for water-saturated sandstone ( $k_\phi = 10^{-12} \text{ m}^2$ ,  $\eta_L = 10^{-3} \text{ Pa s}$ ,  $\rho^L = 10^3 \text{ kg/m}^3$ , and  $\phi = 0.2$ ). Because  $k_\phi$  in the mantle is usually smaller than the value for sandstone,  $\omega_c$  is usually much higher than the seismic frequency. The Biot mechanism for the longitudinal waves also has the value of  $\omega_c$  much higher than the seismic frequency. Therefore, the assumption of the relaxed state is valid for the Biot mechanism.

Another relaxation mechanism is squirt flow [19]. When the pore shape is not spherical, changes in pore pressure induced by the elastic waves depend on the aspect ratio and orientation of each pore. Therefore, liquid pressure becomes heterogeneous at the microscopic scale, and this pressure heterogeneity is relaxed by viscous flow of the liquid between pores (squirt flow). This causes a relaxation of the skeleton moduli from  $N_U$  and  $K_{bU}$  to  $N_R$  and

$K_{bR}$ . The relaxation strength of  $N$ ,  $(N_U - N_R)/N_R$ , is generally larger than that of  $K_b$ ,  $(K_{bU} - K_{bR})/K_{bR}$ , and therefore the effect of squirt flow is larger on  $V_S$  than on  $V_P$ . In Sect. “**General Theoretical Framework to Describe the Dynamics of Solid-Liquid Composite Systems**”, the liquid pressure was assumed to be homogeneous within REV and hence the velocities obtained in Sect. “**Derivation of Wave Equations**” represent relaxed velocities. When  $\alpha$  is large, the relaxation strength is small and the difference between the relaxed and unrelaxed velocities is not significant. As  $\alpha$  becomes smaller, the relaxation strength becomes larger, and hence the relaxed velocities can be used only when liquid pores are not isolated and the frequency of the waves are much lower than  $\omega_c = k_S \alpha^3 / \eta_L$  [28]. For water ( $\eta_L = 10^{-3}$  Pa s) and basaltic melt ( $\eta_L = 1 - 10^3$  Pa s),  $\omega_c$  approaches the seismic frequency when  $\alpha$  is smaller than  $10^{-2} - 10^{-3}$ . If the pore orientation is not random, even under the relaxed state of squirt flow, the pore pressure is different from that estimated for the random orientation. Therefore, when  $V_P$  and  $V_S$  derived in Sect. “**Determinability of Porosity and Pore Shape from Elastic Wave Velocities**” are applied to  $\alpha < 10^{-2} - 10^{-3}$ , the connectivity of the pores, characteristic frequency of squirt flow, and the randomness of pore orientation should all be checked.

As exemplified in the previous section, the presence of  $H_2O$  vapor in the melt phase can relax the liquid compressibility from that of a pure melt ( $\beta_U$ ) to that of a water-melt mixture ( $\beta_R$ ). Figure 11 is obtained by assuming that the characteristic frequency  $\omega_c$  is much higher than the seismic frequency. However,  $\omega_c$  decreases with increasing viscosity of the melt. For andesitic and rhyolitic melts, the characteristic frequency is close to or lower than the seismic frequency range [12].

### Future Directions

One practical problem limiting the determinability of porosity and pore geometry from seismological data lies in the difficulty of accurately estimating  $\Delta V_P$  and  $\Delta V_S$ . For example, a low-velocity anomaly in the upper mantle is generally caused by the superposition of high-temperature anomaly and partial melting. In order to determine porosity and pore geometry from these data,  $\Delta V_P$  and  $\Delta V_S$  associated only to the existence of liquid-filled pores (hereafter referred to as poroelastic effect) should be estimated by accurately correcting the data for the temperature effect. Recent experimental studies on the elastic properties of melt-free olivine polycrystals performed in the seismic-frequency range have demonstrated that at  $T > 1000^\circ\text{C}$  and at such low frequency, the temperature effect consists of both anharmonic and anelastic effects [14]. Unlike the an-

harmonicity, anelasticity cannot be measured by the usual experimental methods using ultrasonic waves. Because experimental data on anelasticity are still limited and the detailed mechanism of anelasticity with or without melt is poorly understood, correction of the data for the temperature effect is difficult. Also, in the crust, accurate estimation of  $\Delta V_P$  and  $\Delta V_S$  for the poroelastic effect is difficult because the effect of anelasticity has not been assessed under crustal conditions and also because the lithological heterogeneity is considered to be larger than in the mantle. Therefore, the separation of poroelastic, temperature, and lithological effects affecting the velocity perturbations is an important subject of future study. Recently, not only the  $V_P$  and  $V_S$  structures but also the three-dimensional  $Q_P$  and/or  $Q_S$  (seismic attenuation) structures and two-dimensional or three-dimensional electrical conductivity structures have become available. These structures provide additional information on liquid-filled pores, temperature anomaly, and/or lithological heterogeneity. The  $Q_P$  and/or  $Q_S$  structures, for example, are important in constraining the magnitude of the anelastic effect [24]. Although the separation of individual factors is difficult to determine from velocity structures alone, additional information from  $Q$  and/or electrical conductivity structures will be very valuable.

### Acknowledgments

The original and more simplified form of this article was published in Japanese [43]. I especially thank Tokyo Geographical Society, for the permission to use a modified version of figures and limited text. I thank S. Nagumo for helpful discussions. I also thank B. K. Holtzman and B. Chouet for reading the manuscript and providing helpful comments.

### Bibliography

1. Aharonov E, Whitehead JA, Kelemen PB, Spiegelman M (1995) Channeling instability of upwelling melt in the mantle. *J Geophys Res* 100:20433–20450
2. Baba K, Chave AD, Evans RL, Hirth G, Mackie RL (2006) Mantle dynamics beneath the East Pacific Rise at  $17^\circ\text{S}$ : Insights from the mantle electromagnetic and tomography (MELT) experiments. *J Geophys Res* 111:B02101. doi:10.1029/2004JB003598
3. Barcion V, Richter FM (1986) Nonlinear waves in compacting media. *J Fluid Mech* 164:429–448
4. Berryman JG (1980) Long-wavelength propagation in composite elastic media 2: Ellipsoidal inclusions. *J Acoust Soc Am* 68:1820–1831
5. Biot MA (1956) Theory of propagation of elastic waves in a fluid-saturated porous solid, 1, Low-frequency range. *J Acoust Soc Am* 28:168–178



6. Biot MA (1956) Theory of propagation of elastic waves in a fluid-saturated porous solid, 2, Higher frequency range. *J Acoust Soc Am* 28:179–191
7. Drew DA (1983) Mathematical modeling of two-phase flow. *Annu Rev Fluid Mech* 15:261–291
8. Duffy J, Mindlin RD (1957) Stress-strain relations and vibrations of a granular medium. *J Appl Mech* 24:585–593
9. Hasegawa A, Yamamoto A (1994) Deep low-frequency micro-earthquakes in or around seismic low-velocity zones beneath active volcanoes in northeastern Japan. *Tectonophysics* 233:233–252
10. Holness MB (1997) Surface chemical controls on pore-fluid connectivity in texturally equilibrated materials. In: Jamtveit B, Yardley B (eds) *Fluid flow and transport in rocks*. Chapman and Hall, London, pp 149–169
11. Holtzman BK, Groebner NJ, Zimmerman ME, Ginsberg SB, Kohlstedt DL (2003) Stress-driven melt segregation in partially molten rocks. *Geochem Geophys Geosyst* 4:8607, doi:10.1029/2001GC000258
12. Ichihara M (1997) Mechanics of viscoelastic liquid containing bubbles; implications to the dynamics of magma. Ph D thesis, Univ. of Tokyo (in Japanese)
13. Iwamori H (1994)  $^{238}\text{U}$ ,  $^{230}\text{Th}$ ,  $^{226}\text{Ra}$ - and  $^{235}\text{U}$ ,  $^{231}\text{Pa}$  disequilibria produced by mantle melting with porous and channel flows. *Earth Planet Sci Lett* 125:1–16
14. Jackson I, Fitz Gerald JD, Faul UH, Tan BH (2002) Grain-size-sensitive seismic wave attenuation in polycrystalline olivine. *J Geophys Res* 107(B12):2360, doi:10.1029/2001JB001225
15. Kuster GT, Toksöz MN (1974) Velocity and attenuation of seismic waves in two-phase media, 1, Theoretical formulations. *Geophysics* 39:587–606
16. Masterlark T, Wang HF (2002) Transient stress-coupling between the 1992 landers and 1999 Hector Mine, California, earthquakes. *Bull Seism Soc Am* 92:1470–1486
17. Mavko GM (1980) Velocity and attenuation in partially molten rocks. *J Geophys Res* 85:5173–5189
18. Mavko G, Mukerji T, Dvorkin J (1998) *The Rock Physics Handbook*. Cambridge University Press, New York
19. Mavko GM, Nur A (1975) Melt squirt in the asthenosphere. *J Geophys Res* 80:1444–1448
20. McKenzie D (1984) The generation and compaction of partially molten rock. *J Petrol* 25:713–765
21. Mura T (1987) *Micromechanics of defects in solids*, 2nd edn. Martinus Nijhoff Publishers, Dordrecht
22. Nakajima J, Hasegawa A (2003) Tomographic imaging of seismic velocity structure in and around the Onikobe volcanic area, northeastern Japan: implications for fluid distribution. *J Vol Geotherm Res* 127:1–18
23. Nakajima J, Matsuzawa T, Hasegawa A, Zhao D (2001) Three-dimensional structure of Vp, Vs, and Vp/Vs beneath the northeastern Japan arc: Implications for arc magmatism and fluids. *J Geophys Res* 106:21843–21857
24. Nakajima J, Takei Y, Hasegawa A (2005) Quantitative analysis of the inclined low-velocity zone in the mantle wedge of northeastern Japan: A systematic change of melt-filled pore shape with depth and its implications for melt migration. *Earth Planet Sci Lett* 234:59–70
25. Nowick AS, Berry BS (1972) *Anelastic relaxation in crystalline solids*. Academic Press, New York
26. Obara K (2002) Nonvolcanic deep tremor associated with subduction in Southwest Japan. *Science* 296:1679–1681
27. O'Connell RJ, Budyanskiy B (1974) Seismic velocities in dry and saturated cracked solids. *J Geophys Res* 79:5412–5426
28. O'Connell RJ, Budyanskiy B (1977) Viscoelastic properties of fluid-saturated cracked solids. *J Geophys Res* 82:5719–5735
29. Ohmi S, Obara K (2002) Deep low-frequency earthquakes beneath the focal region of the Mw 6.7 2000 Western Tottori Earthquake. *Geophys Res Lett* 29:1807, doi:10.1029/2001GL014469
30. Ribe N (1985) The deformation and compaction of partially molten zone. *Geophys J R astr Soc* 83:487–501
31. Rice JR, Cleary MP (1976) Some basic stress diffusion solutions for fluid-saturated elastic porous media with compressible constituents. *Rev Geophys* 14:227–241
32. Riley GN, Kohlstedt DL (1991) Kinetics of melt migration in upper mantle-type rocks. *Earth Planet Sci Lett* 105:500–521
33. Schäfer K (ed) (1980) *Landolt-Börnstein Numerical Data and Functional Relationships in Science and Technology, New Series IV, vol 4, High-Pressure Properties of Matter*. Springer, Berlin
34. Scott DR, Stevenson DJ (1984) Magma solitons. *Geophys Res Lett* 11:1161–1164
35. Spiegelman M, Kelemen PB (2003) Extreme chemical variability as a consequence of channelized melt transport. *Geochem Geophys Geosyst* 4:1055, doi:10.1029/2002GC000336
36. Spiegelman M, Kelemen PB, Aharonov E (2001) Causes and consequences of flow organization during melt transport: The reaction infiltration instability in compactible media. *J Geophys Res* 106:2061–2077
37. Spiegelman M, McKenzie D (1987) Simple 2-D models for melt extraction at mid-ocean ridges and island arcs. *Earth Planet Sci Lett* 83:137–152
38. Stevenson DJ (1986) On the role of surface tension in the migration of melts and fluids. *Geophys Res Lett* 13:1149–1152
39. Stevenson DJ (1989) Spontaneous small-scale melt segregation in partial melts undergoing deformation. *Geophys Res Lett* 16:1067–1070
40. Stolper E, Walker D, Hager BH, Hays JH (1981) Melt segregation from partially molten source regions: The importance of melt density and source region size. *J Geophys Res* 86:6261–6271
41. Takei Y (1998) Constitutive mechanical relations of solid-liquid composites in terms of grain-boundary contiguity. *J Geophys Res* 103:18183–18203
42. Takei Y (2002) Effect of pore geometry on Vp/Vs: From equilibrium geometry to crack. *J Geophys Res* 107(B2):2043, doi:10.1029/2001JB000522
43. Takei Y (2005) A review of the mechanical properties of solid-liquid composites. in *Japanese, J Geography* 114(6):901–920
44. Tsumura N, Matsumoto S, Horiuchi S, Hasegawa A (2000) Three-dimensional attenuation structure beneath the northeastern Japan arc estimated from spectra of small earthquakes. *Tectonophysics* 319:241–260
45. von Bargen N, Waff HS (1986) Permeabilities, interfacial areas and curvatures of partially molten systems: Results of numerical computations of equilibrium microstructures. *J Geophys Res* 91:9261–9276
46. Waff HS (1980) Effects of the gravitational field on liquid distribution in partial melts within the upper mantle. *J Geophys Res* 85:1815–1825
47. Walsh JB (1969) New analysis of attenuation in partially melted rock. *J Geophys Res* 74:4333–4337
48. Wang HF (2000) Theory of linear poroelasticity with applica-

- tions to geomechanics and hydrogeology. Princeton University Press, Princeton, New Jersey
49. Watanabe T (1993) Effects of water and melt on seismic velocities and their application to characterization of seismic reflectors. *Geophys Res Lett* 20:2933–2936
  50. Watt JP, Davies GF, O'Connell RJ (1976) The elastic properties of composite materials. *Rev Geophys* 14:541–563
  51. Wong T-F, Ko S, Olgaard DL (1997) Generation and maintenance of pore pressure excess in a dehydration system 2, Theoretical analysis. *J Geophys Res* 102:481–852

## Evacuation as a Communication and Social Phenomenon

DOUGLAS GOUDIE

Australian Centre for Disaster Studies, School of Earth and Environmental Science, James Cook University, Townsville, Australia

### Article Outline

Glossary

Definition of the Subject

Introduction

Concepts, Language and Mathematically Modeling the Propensity to Evacuate

Mathematical Modeling

Effective Risk Communication

Integrating Theory and Implementation

Institutional Barriers to Greater Community Self-Help Experiences and Lessons – Some Case Studies

Discussion

Acknowledgment

Bibliography

### Glossary

**Community** A group of neighbors or people with a commonality of association and generally defined by location, shared experience, or function [59].

**Community empowerment** Internally and externally nurtures a community to accept that residents live in a hazard zone, and they choose to do things as a group to maximize their safety.

**Community safety group** Existing community groups (such as Neighborhood Watch) and individuals, working with formal response organizations form a coherent affiliation in and near a hazard zone, to help maximize safety and care for all community members.

**Disaster** The interface between an extreme physical event and a vulnerable human population [81].

**Disaster lead time** The time taken from first detection of a natural disaster threat to the likely time of impact on humans or human structures.

**Disaster threat** A natural extreme event which may impact on a community.

**Effective risk communication** . That which motivates people to maximize their own safety.

**Emergency** An actual or imminent event which endangers or threatens to endanger life, property or the environment, and which requires a significant and coordinated response [55].

**Evacuation** People relocating to safely escape hazardous disaster impacts. To move from a high danger zone to relative safety.

**Hazard** A source of potential harm or a situation with a potential to cause loss. A situation or condition with potential for loss or harm to the community or environment [55]. Hazard is synonymous with ‘source of risk’ [25].

**Hazard zone** Defined geographic areas which may be subject to a natural disaster impact of flood, bush-fire, storm surge, destructive winds, earthquake, landslide or damaging hail. Hazard zones include major accident sites, including industrial, transport or mining precincts; or biological or terrorist threat or impact, or from-source predicted area(s) of pandemic spread.

**Mitigation** Any efforts taken which may reduce the impact of a threat.

**Prevention** Measures to eliminate or reduce the incidence or severity of emergencies [55].

**Ramp-up preparations** The final set of preparations and precautionary evacuations taken ahead of a forecast disaster impact. This includes earlier final actions than precipitated by formal organizations.

**Risk treatment options** Measures that modify the characteristics of hazards, communities and environments to reduce risk, e. g. prevention, preparedness, response and recovery [55].

**Vulnerability** comprises ‘resilience’ and ‘susceptibility’. ‘Resilience’ is related to ‘existing controls’ and the capacity to reduce or sustain harm. ‘Susceptibility’ is related to ‘exposure’ [25].

### Definition of the Subject

This article intends to show how system and complexity science can contribute to an understanding and improvement of evacuation processes, especially considering the roles of engaged communities at risk, the concepts of community self-help, and clear communication about local threats and remedies.

This article shows researchers in Complexity and Systems Science (CSS) a social sciences approach to maximize effective and precautionary evacuation, maximize safety, minimize loss and speed full recovery. The computational and analytical modeling tools of CSS may be considered to apply to a complex interaction of community awareness, inclination to accept the reality of a natural disaster threat, along with achieving background and final preparations to maximize safety and recovery from a natural disaster impact. This article may stimulate CSS

researchers to develop detailed models of the complex systems and complexity of melding information from Weather Bureaus and Disaster Managers, via contacts and intervening media to communities at risk, with the shared social goal of maximizing safety. This social sciences task requires cross-disciplinary approaches of respect and response.

The old disaster management model lacked the predictive and rapid communication systems now available and developing in disaster predictive models (such as a flood maps). An approach to modeling the great complexity of human behavior responding to threat is provided. Such a model must include people's prior knowledge of a threat type, and consider such fine detail as the overarching language used in a country with threat zones, and the dominant languages of all under threat. It is hoped this article stimulates CSS models to further engage in this social good of helping people get safe and stay safe through natural disasters by providing predictive tool to Authorities to better inform and encourage those at risk to action, including the possible need for precautionary evacuations ahead of a predicted impact.

Disaster management in Australia, and increasingly, globally, is focused on mitigation as part of a 'threat continuum', from acceptance that some locations are vulnerable to a hazard impact, through to recovery [13]. Emergency warnings and a possible need to evacuate are embedded as 'spikes' on that continuum. Thus, this article stresses the importance of developing ways; incentives, to mobilize aware at-risk community members to precautionary self-evacuation. For this to happen, people need to know and internalize the reality that they are in a hazard zone.

Thus, in the cost-effective philosophy of engendering self-help, the process of understanding the complexity of achieving the shared social goal in maximizing safety and minimizing loss is to engender creation of empowered communities with a high motivation for safety-oriented and precautionary action. This is likely to lead to minimized loss and disruption, and maximized recovery. This article details many elements of that process, and invites detailed development of the Sustainability Implementation Research to achieve that goal through CSS.

To model the path to collective safety, the complexity of the dynamics at play need to be clarified: impact preparedness, including possible evacuation, is a communication and social issue.

This article demonstrates that acknowledgment of hazard zones, developing community acceptance of threat and needed action needs to be at the individual, household and community levels. Evacuation modeling is needed

only for those whose homes may be at real threat of a disaster impact. For those living in a hazard zone, a fully informed community, who have internalized the reality of the threat and have worked for maximum background preparation, and have mechanisms to receive alerts and warnings of a looming threat; a community predisposed to precautionary evacuations will result.

Capturing this complexity is the challenge for modelers. Evacuation is about hazard zone residents actively monitoring a looming threat via refined communication channels detailed in this article, within a developed social predisposition to act. Some examples are given. For consideration by scientists and students internationally, this article introduces the Communication Safety Triangle and the Seven Steps to Community Safety on the Preparedness Continuum, within the new research frame of Sustainability Implementation Research (SIR).

## Introduction

The purpose of this article is to share with modelers and complexity and system scientists the social and communication issues of modeling effective safety strategies to a natural disaster threat. It is hoped, through the approaches and processes described in this article, that modelers will more clearly link physical threats with warnings and community engagement.

This article first looks at the definitions and language used in risk communication and effective warnings, leading to informal and formal evacuations, then considers some Australian policies relating to emergency management. Theories related to risk communication are presented, with examples of evacuation issues provided from Indigenous communities, and from non English speaking households. The needed conceptual shift to self-help is placed within the larger theoretical frame of paradigms and paradigms shifts.

An example of including residents to internalize threats is given, followed by a more general example of transport evacuation.

A discussion of international evacuation issues precedes a broader view of some of the institutional barriers which may restrict the uptake of the seven step approach to an aware, informed community, relying on accurate information and choosing to self evacuate as a precaution. These issues are discussed, considering effective ways of allowing people to know that they are at risk so they are inclined to evacuate themselves, as a practice. These approaches can be used or tested by other scholars. Recommendations and a summary of the key issues to maximize voluntary and safe evacuations finish this article.

### Some Key Evacuation Issues

In an era of increasing social self-help [7,55] and community empowerment [13,25,26], a part of global efforts to embrace Ecologically Sustainable Development [8,62] is to increasingly see evacuation as a social phenomenon.

This article provides a conceptual frame, the Communication Safety Triangle (CST) which includes responsible media telling people at risk what they need to know and *seven steps to community safety* (7SCS). The 7SCS help guide emergency managers and modelers to treat the possible need for evacuation as a decision-making process where the community should be aware of the potential threats and receive clear, detailed and reliable information on the possible need to evacuate, so most residents in a high impact zone self-evacuate in a precautionary way, as a practice. Examples provided in this article illustrate this new, *sustainability implementation research* (SIR) approach to disaster management.

Warnings precede a perceived need to evacuate. In the USA, the need for an integrative approach to warnings is identified: “There is a major need for better coordination among the warning providers, more effective delivery mechanisms, better education of those at risk, and new ways for building partnerships among the many public and private groups involved” [63].

Sustainability implementation is the way to achieve a sustainable future. Disaster management, effective risk communication and community self-help provide a stark and comprehensible example of what sustainability implementation means, how it will benefit societies, and will help channel us into a safer, more sustainable future. Within this approach, scientists and students become major agents for sustainability implementation, as models can illustrate the needed paths to achievement.

### Evacuation Overview

There are three types of disasters which may require evacuation: human induced and natural disasters with and without lead times. This article is focused on precautionary evacuations ahead of natural disasters with lead times (Table 1). The research and conceptual frames draw on and are applicable to all communities in hazard zones. Defining, modeling and effectively communicating to at-risk residents and travelers about specific geographic hazards and safety-oriented behaviors are core elements of successful precautionary evacuations.

Within Table 1, evacuation may be a response to an emerging hazard of indicated intensity, direction and speed. This article advocates development of informed and directed communities which will actively respond to

a communicated threat, with the vulnerable moving themselves early to places of safety, or being helped by other community members to move.

### Concepts, Language and Mathematically Modeling the Propensity to Evacuate

Precautionary self evacuation pivots on knowing who is at risk. With the help of researchers, modelers; planning and community involvement can prepare people and their valuables to minimize loss. Very public hazard maps will help people internalize that they are in a hazard zone, and what they should do. The concept is not new or alarming: every public building has emergency exit routes marked, with all that implies. Air flight comes with the mandatory emergency preparation presentation. Minute or recurrent risks to where we live or travel to should be no less public, nor more alarming than a fire drill in a public building.

### Hazard Types – Little or Considerable Warning Times

Hazards may or may not have lead times (Table 1). Sudden onset threats include: tsunami, earthquake, major eruption, major toxic spill or discharge, mine disaster, terrorist threat or attack. Signaled threats include: – cyclone, flood, fire and destructive winds. This article focuses on disasters with sufficient warning periods to be able to evacuate the vulnerable away from the predicted worst impact areas. Table 1 that with sudden onset impacts, sheltering to survive the first few minutes is critical, then moving to open; stable ground is important to avoid further aftershocks or landslide. Always take direction from local authorities. For natural disasters with lead times, Table 1 shows that, if you will clearly be safe where you are, stay put, and prepare as best you can. If, in the worst case, you may not be safe, move early (Table 1). The vulnerability of individuals needs to be considered. In a bush fire, for instance, the general advice is: if the property is well prepared [7], stay and defend. If, however, you may not be able to cope with the psychological terror of staying through the fire front, fully prepare your property, and then leave early (Table 1). If the exit route may be blocked by flood waters, or by dense smoke or fire, evacuation needs to precede that obstruction. In the ‘background’ phase of community disaster preparation, all such possible obstacles to a clear escape route (including gridlock congestion) need to be factored in to the timing of precautionary evacuation (Table 1).

Finally, as seen in the Woodgate Beach example of Sect. “**Experiences and Lessons – Some Case Studies**”, the level of community support in ensuring all residents are aware of and prepared for the ‘ramp-up’ phase of a possible disaster impact, and receive the warning as soon as

Evacuation as a Communication and Social Phenomenon, Table 1  
Evacuation decision matrix – short to long warning times

Evacuation decision matrix, Evacuation around <i>sudden onset impact</i>			
Hazard:	Landslide	Earthquake	Tsunami
<b>Possible safety-oriented response</b>	Stay in strong structure. Move across slope as soon as possible.	Get into the open or shelter under strong structure.	Immediately flee to higher ground.

Precautionary Evacuation (PE) decisions with lead time – signaled threats			
Considerations for evacuation decision	Hazard		
	Destructive wind/cyclone	Fire	Flood
<b>1. Vulnerability of present environment</b>	If likely to be in a storm surge, must PE; If shelter weak, must PE	House material, surrounds, water available. If poor, PE	May be inundated= PE; may be cut off, consider PE
<b>2. If 1 OK: vulnerability of individuals</b> (e. g. weak)	PE first	Asthmatics and less able: PE	Judgments of flood height. If in any doubt, PE
<b>3a. Distance to safe shelter</b>	The further, PE earlier	The further, PE earlier	The further, PE earlier
<b>3b. Safety along exit route</b>	Know in preparation	Know in preparation	Know in preparation
<b>3c. Means of travel</b>	Reliability and suitability	Reliability and suitability	Reliability and suitability
<b>4. Community cohesion</b>	Help available	Help available	Help available

possible, the able-bodied will help the less able to get out of harm's way early in the threat period, as a safety-oriented practice, minimizing the demands of the formal response groups as the likelihood of impact increases.

All effective communication involves sending signals and having them received and processed, then incorporated into the receiver's world view [27,40,41,76,81,82]. Effective risk communication [4,13], Sect. "Effective Risk Communication", motivates people to act to maximize their own safety. This may not happen if people have not internalized that the threat is real. They are in denial. Alternatively, people may be ignorant of the threat, or why it should be taken seriously. Salter [77] categorizes ignorance from pure ignorance to acts of ignoring.

### Disaster Definitions

A disaster may be seen as a negative impact of a hazard on a community as measure of vulnerability. The language of disaster mitigation evolved and is increasingly practiced since the late 1990s i. e. [13,96]. Risk is seen as a function of probability and consequence, related to exposure and the level of force embedded in the threatening hazard.

Boughton [6] points out that natural disasters are usually extremely rare for the individuals concerned but they can cause massive impacts. Because Australia is so vast, overall there are reasonably frequent natural disasters. However, in most locations they are rare indeed. Boughton [6] argues that a "natural disaster" is a natu-

ral event in which the community life is seriously and traumatically disrupted. Embracing the way forward with 'structural mitigation', "... a key step in preventing natural disasters is to prevent building damage." [6]. Like "disasters", "community" is difficult to define [84]. As developed in Sect. "Integrating Theory and Implementation", "community" is the collection of people in a close geographic area, particularly focused on near neighbors and supportive friends of those in or near a known hazard zone.

### What to Communicate

Having the right words or approaches in place *as policy* does not automatically guarantee community safety-oriented responses to disruptive warnings. Since 1989 the approach to cope with disasters has been *prevention, preparedness, response and recovery* training courses. This helps focus all concerned on the temporal sequence. The language could perhaps be refined to talk about acceptance that a threat exists, background, then 'ramp-up' (final) preparations to safe shelter; impact, then orderly return and recovery. Classically, 'response' was seen as the final, near impact flurry of 'lock down' activity [57].

Deeper than language is our conceptual frame (Sect. "Integrating Theory and Implementation"). If researchers encouraged disaster managers to move 'response' behavior back a day; a few hours to 'precautionary, early response', many of Lewis's [57] legitimate concerns would be addressed. Some disaster managers may see themselves

as dramatic figures in the early impact phase of a disaster, rather than calm, precautionary minimizers of risk. This culture has changed greatly over the last decade (Sect. “**Integrating Theory and Implementation**”). For many reasons, there is a clear divide between the US Federal response to Cyclone Katrina (2005) and to the California fires (October, 2007). This shift from passive to active can be encouraged by modelers. As part of effective risk communication to encourage precautionary impact preparedness, research with remote Indigenous communities and recently arrived, non-English speaking refugees (Sect. “**Integrating Theory and Implementation**”) showed the importance of accurate, plain English, and the use of images.

Yates [93] argues mitigation efforts need to be refined to make sure they are focused on issues from the relevant local communities. Much of the problem of non response seems that ‘the message’ to take care does not effectively get through to the target [68]. It is to do with communication, with signals sent, signals received, and their interpretation.

The types of received and interpreted messages are explored specifically in Sect. “**Integrating Theory and Implementation**”, which develops the intellectual foundation to more fully understand the semiotics of risk communication. The following outlines the concepts and relevance of semiotics.

**Semiotics** is *the study of signs* including words, sounds and such things as ‘body language’. By the ‘message conveying’ principles of semiotics we can understand how various authors on disasters and evacuations approach the topic, and which cultural signs and symbols they manipulate. For example, a sign showing a person running up slope ahead of an exaggerated tsunami wave contains all the preferred imagery to tell people what to do in a tsunami warning. Images, as seen for locating oneself in airports or finding evacuation stairs, do not use words.

The next section lays the foundations for a mathematical modeling of propensity for householder self-evacuation, centered on knowledge and refined communications acting on residents in hazard zones who are predisposed to act in their own safety.

## Mathematical Modeling

### Foundations of the Household Safety Preparedness and Action Index: the HSPA Index

This section does not consider refined algorithms to define how successfully people may respond to *Authority’s Perceived Need to Evacuate (APNE)*. This section considers the form and ingredients of how this could be done; the elements and their possible interactions to predict the

success of APNE. Unfortunately, in the previously cited Canberra 2003 bush fires, the APNE did not appear universally, nor was it properly conveyed to the residents in dire risk.

Rohrmann [75] has clearly mapped a process from the warning signal to the hoped-for response. The information and processing flow to internalized inferences is called intuitive heuristics [74]. Renn & Rohrmann [74] basically argue that people process probabilistic information on the likelihood of them needing to move, aligned and convergent with researchers like Thompson [87]. Hence the need for the seven step process (Fig. 8) to trigger a paradigm shift for people in danger zones.

As seen in formula 1, there are finite evacuation triggers (ET), ranging from no lead time  $LT_0$ , such as an earthquake, to a LT of days ( $LT_{xd}$ ), such as a cyclone. So a key component of a successful response to an APNE is the time in which to disseminate the warning ( $LT_{0-xd}$ ; Table 1) Also, Impact Severity (IS) is central to safety and loss.

As detailed in Sect. “**Integrating Theory and Implementation**” and Fig. 8; *The seven steps to community safety*, there is a convincing body of research which indicated that People’s Propensity to Respond (PPR) to an emergency warning (whether that response be to finalize preparation to stay and defend the property or evacuate as a precaution) is strongly linked to their Acceptance they are In a Hazard Zone (AIHZ).

Community Resilience (linkage and inter-support and communication – CR) is also important; along with a potential evacuee’s Knowledge Base of the hazard and safety-maximizing behavior (KB). Thinking Through to Recovery (TR) is also important. As per points 6 and 7 in the 7SCS; Fig. 8, the medium of warning delivery, and the quality of the warning information (Medium & Message: MM) is a factor in the response of those under threat. Overall, Sorenson and Mileti [81] believe that increased credibility of the warning source means the specific warning will be more effective. They also provide evidence that the electronic mass media produced the most believable public warnings. This underlines why development of formal links between all types and levels of media information about hazards and preparation from the weather bureau and emergency managers is so important. Having formal links with the media, coupled with web-delivered ‘real-terrain’ simulations of the hazard will help produce a powerful and effective ‘active warning’ regime.

Issues of Exit Route(s) (ER), Possible Travel Mode (PTM) may be crucial. A normally free-flowing exit route may be blocked by the disaster itself, or others trying to flee. This may involve accidents. People’s psychological and physical states of Well-Being (WB) will also affect po-

tential effective evacuation decisions and accomplishment. Authors like Rohrmann [75] have attempted to stylistically model some of this complexity.

Many of the earlier recognized impediments to full preparedness, like unrealistic optimism (Heller et al., [46]) are incorporated into one overarching factor of formula (1): People’s Propensity to Respond (PPR), including the centrally important feature of individual households fully accepting they are in a hazard zone, and that they need to engage in background preparation, ‘listen up’ around a hazard threat, and undertake their own, precautionary final preparations.

Cutter et al. [17] developed 11 factors indicating vulnerability caused by a major disaster, using principal components analysis within factor analysis. The resultant Social Vulnerability Index (SoVI), necessarily weights the 11 components according to the percentage of variation of analysis of counties on the USA as to resident’s vulnerability. Personal wealth, age and housing density head up the contributions of the 11 variables used to determine the SoVI of USA counties. Cutter’s SoVI contributes to the likelihood that proper preparations; including for a timely evacuation, would occur, either at a broad regional or more individual level. This SoIV is included in the HPSAI below. Its weighting remains to be tested. A final issue overarching many others is Institutional Barriers to change (IB), detailed in Sect. “Institutional Barriers to Greater Community Self-Help”. The support or other-

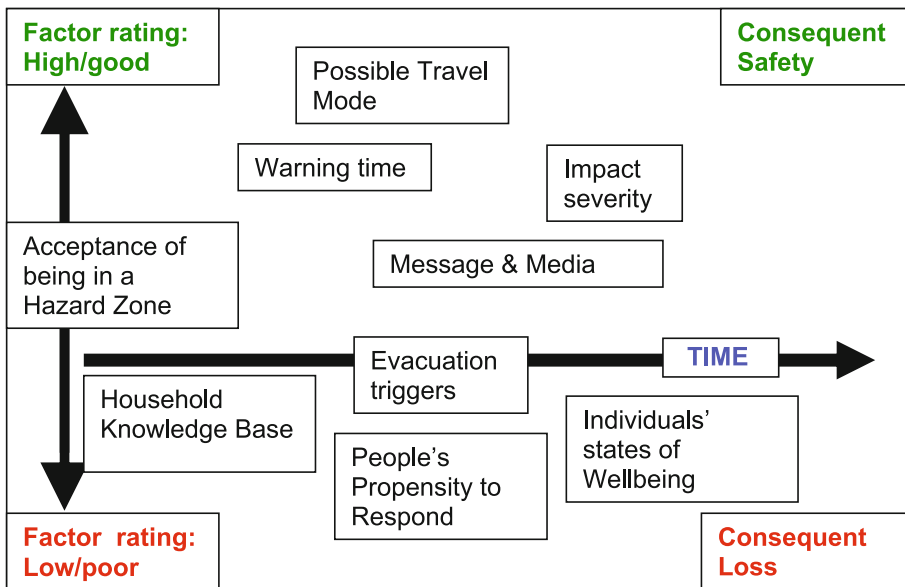
wise of the media can play a critical role in effective crisis communication – most messages are likely to be through the media, so the media is part of the systems complexity blend: Media Support (MS).

From the weight of SIR and the conceptual framework described in this article, the following generic formula expressing the above factors likely to impact on householder propensity for precautionary evacuations may entice other, more theoretical researchers, to develop the needed algorithms. What follows is purely a design base for others to develop predictive modeling of *People’s Propensity and Capacity to Successfully Evacuate (PPCSE – or safely and actively stay)* PPCSE is better named the *Household Preparedness and Safety Action Index: the HPSA Index* . The following is a synthesis of Bayesian Logic (Hoeting et al. [45]) and Eigenvalues.

**People’s Propensity and Capacity to Successfully Evacuate (PPCSE) or Household Preparedness and Safety Actions Index (HPSAI):**

$$\begin{aligned}
 &HPSAI \propto \\
 &f(PPR)(AIHZ)(LT_{0-xd}) \\
 &(APNE)(ET)(IS)(AIHZ)(CR)(TR)(MM)(ER) \\
 &(PTM)(WB)(SoIV)(IB)(MS)
 \end{aligned}
 \tag{1}$$

Formula (1) anticipates constants to ‘weigh’ each factor according to its contributing importance on the resultant



Evacuation as a Communication and Social Phenomenon, Figure 1  
 Factors in the Eigen plane determining *Household Preparedness and Safety Action Index*



HPSAI (see [45] p 384) for detail of the Bayesian modeling approach.

Resultant safety (*safety* being a combination of successfully evacuating to a safe place, or sheltering within the impact zone in a safe place) of those under threat will be greatest (Fig. 1) where the factors of formula (1) tend to intersect in the positive quadrant of an Eigen plane (Fig. 1) [94]. Equally, outcomes of loss will occur with, for example, short lead times, insufficient warnings of evacuation triggers or lack of belief that residents or travelers are in a hazard zone. A journal search showed few links between disasters and Eigenvalues. An exception is Fowler et al. (2007, [32]). With global warming and increased populations increasingly encroaching into obvious hazard zones, modeling to influence planning and to maximize HPSAI must be a major growth industry.

### Calibrating the Model: the Greek Fires, August '07

[98] or California in late October '07 [99] show that if various of the above variables or likely variable clusters or Eigenvalues from factor analysis have low values, people will not be placed in a strong, precautionary position of safety. This was true of fire-threatened residents in Canberra in 2003 (<http://www.abc.net.au/canberra/bushfires/>), all ultimately leading back to the mathematical weight which needs to be given to each point on in the 7SCS: especially accepting that the threat is real (and possibly imminent).

Once the above-described preparedness paradigm is internalized, it is difficult to experience any disasters news coverage without feeling some despair: cars submerged or floating in flood waters; people fleeing for their lives ahead of imminent immolation; any reports of lost lives through natural disasters with lead times of hours or longer. There are few excuses for this to remain so. In most parts of the developed world, there are prevailing technologies to detect and rapidly convey threat. Researchers are well placed to conceive, trial and link the threat information directly to prepared populations; to close the gap between knowing and acting to maximize safety, as a social exercise of community engagement and communication refinement.

Cyclone Larry [54] showed that, with general community acceptance of the reality of the threat, good warnings and a lead time of about 20 h, combined with well developed disaster management and media cooperation in getting the needed messages to people well primed to the 7SCS, there was no loss of life, although about 5000 people were directly subjected to destructive winds of a category 4/5 cyclone. Residents of the impact zone

generally followed the 7SCS, supporting the general approach of formula (1): the community was prepared and acted in a precautionary way to maximize their own safety.

The above model design offers elements/factors to maximize *People's Propensity and Capacity to Successfully Evacuate (PPCSE – or safely and actively stay); the Household Safety Preparedness and Action Index (HSPAI)*, so strengthening a paradigm among Agencies and hazard-zone residents of 'self-help' and 'community engagement' approaches; less passive than earlier approaches used in considering people's vulnerability. This generic approach is supported by model testing by Schadschneider et al. (► **Evacuation Dynamics: Empirical Results, Modeling and Applications**).

The Australian examples detailed in this article, including the national shift to mitigation (COAG 2004 [13]), best illustrated in the Woodgate Beach example (Sect. "Experiences and Lessons – Some Case Studies"), show that the conceptual frame and thus variables chosen to mathematically model evacuation propensity and capacity are more important than the results of any model using 'passive' variables or variable clusters. The advantages and means of achieving this shift from community passivity to partnerships; place- and people-based community empowerment and self-help, form the basis of this article. The above mathematical approach may help validate and guide this necessary conceptual shift.

The next section considers policy, laws, concepts and possible safety-oriented actions that words, images and stories convey to vulnerable residents in hazard zones. In Australia, policy and legislative requirements are indicated, which aim to responsibly maximize community safety [13,24,25,26,27,28,44,70]. Some new concepts, epitomized by new language phrases (with their embedded meanings) are discussed in Sect. "Integrating Theory and Implementation", especially *Community safety groups*. New phrases include 'Social Burnoffs' and 'Practice evacuations'. Like "The Communication Safety Triangle" and "Seven steps to community safety", it is hoped these expressions enter local, state and national lexicons and modeling in disaster management.

## Effective Risk Communication

### Policy and Laws on Hazard Preparedness and Evacuation

All risk communication operates within government policies and legislation. Forced evacuation may be normal in some jurisdictions for all threats; outlawed in others. Policy may encourage passivity, or be energetically pro-active

in achieving the 7SCS, both in urban planning and building materials, and community engagement. In Australia, for instance, there are many federal and regional guides to urban growth, e. g. [2]. A new urban paradigm is, perhaps, best expressed as: “Think globally, act locally, respond personally” ([2] pnp, [3] p 1). This paradigm shift can easily embrace and be informed by computer modeling of individual’s behavior, demonstrated in recent Springer’s Journal of Systems Science and Complexity articles, such as Kikuchi T & Nakamori Y [97]. Agent model analysis to explore effects of interaction and environment on individual performance [97]

From all decisions of urban development being in the hands of ‘experts’, government policy increasingly requires a dialog with local residents, through public participation. The first barrier to ecologically sustainable urban development is “... belief systems – doubting a problem exists, or supporting the status quo. [Solutions include] ... consciousness raising campaigns, public participation in decision-making, demonstration projects and incentives and disincentives” ([76] p 46). This is the KB of formula (1).

This article stresses that response models, Authorities and all residents are constrained (and empowered) by national and regional laws and policies. Research can only work within a prevailing frame, and will be welcomed by or influence prevailing policy. It is thus useful to have clear goals to achieve social or environmental ‘good’ [13,44,70]. This requires greater community involvement in mitigation, and responsible media helping to mitigate disaster impacts (MS of formula 1). In Australia, the national radio broadcaster has a binding agreement with the weather bureau and disaster managers to read issued warnings verbatim; report and engage community members in safety oriented behavior, by providing relevant information. Local community media are often willing to develop the same responsible functions.

### Risk Communication’s Core Goals

Risk communicators advise “individuals and communities to respond appropriately to a threat in order to reduce the risk of death, injury, property loss and damage.”[42]. Risk communicators and disaster managers need to formally work closely with the media to maximize social benefit.

Although it may be intended that information flows and is received accurately; that desired behavior will result, and that only communication techniques are important [49], humans tend to be irrational and optimistic, and only hear what they want to hear. It is not what

our message *is*, but what, if anything, the listener *does* with our message. To have any chance of ‘success’, information needs to have meaning which is shared between those who construct and send the warning, and those for whom the warning is meant to inform and motivate to action.

‘Action statements’ (what the at-risk person, family or community should actually do to minimize damage) are seen as central to the whole purpose of risk communication [78]. Kasperson & Stallen [49], along with Salter [77,78] and others detail risk communication messages in terms of content, clarity, understandability, consistency, relevance, accuracy, certainty, frequency, channel, credibility, public participation, ethnicity, age, gender, roles, responsibility, elements, sequencing, synopsis, prognosis, location, action, warning timing, and action statements. It is not a case of saying: “a category three cyclone will pass over Smithfield”. It is more a case of making sure the members of Smithfield hear that message, and feel moved to and competent to take well-understood personal risk-minimizing actions.

### Knowledge, Self, World Views and Messages

For precautionary evacuations to be successful, the seven steps to safety (Sect. “**Integrating Theory and Implementation**”) have occurred; or people have been coerced by authorities, or they have seen others depart, and felt insecure, so they leave as a follower of the ‘innovation’ of the ‘norm’: precautionary evacuation. This occurred in the desktop exercise in an isolated settlement in March 2007, detailed in Sect. “**Experiences and Lessons – Some Case Studies**”.

It is now difficult to understand why people needed prompting to take evasive action against the forecast Brisbane floods in 1974, but many ignored the prompts. Authorities believed houses and roads would be flooded, so people should finalize travel or evacuation early [10]. For those in the flood zone, 88% of a later survey sample reported evacuating their home. Some took this step very early, but 67% of respondents only made preparations immediately before leaving home. Twelve percent only left on foot or in boats *after waters entered the main living areas of their homes*. Almost 22% of respondents said they made no preparations, mainly because the threat was not recognized in time [10]. This well documented ‘poor’ crisis communication can help calibrate formula 1.

A key safety message surrounding floods is not to enter flood waters. Figure 2 is one example of using images to help convey the reality of a threat, and the pitfalls of belated action.



Evacuation as a Communication and Social Phenomenon, Figure 2  
 What flood-threatened residents need to consider (Photos courtesy Townsville City Council)

### Effectively Conveying Meaning

Because language is pivotal to acceptance of risk and conveyed warnings of need for evacuations, it is important that all participants reasonably agree on word meanings. This section starts in the comfort zone of simple definitions, then considers semiotics, considers core issues of world views (paradigms), and finishes with the uneasy realities of our knowledge base, our epistemological orientation. Sect. **“Integrating Theory and Implementation”** shows that some cultures do not carry the background cultural experience to absorb the meaning of cyclone or bushfire – there is no shared experience or ‘stories’ of the power of these extreme forces of nature. Some of the underlying knowledge foundations of context, intent and behavioral motivation need to be considered – how humans construct, transfer, acquire and use knowledge.

Imparting meteorological knowledge, then warnings, to target audiences to engender safety-oriented responses is a complex exercise in social empowerment, explored in Sect. **“Experiences and Lessons – Some Case Studies”**. As information promulgators, information and warning sources should understand a little of *Perception* (the raw data from the outside world entering an organism via any of the five senses), *Cognition* (internal processing, analyzing, information storage and processing), *Attitudes* (how we think and feel about particular issues, implying a predisposition to specific action), *Language use* and links to *Behavior*.

An intellectual framework to risk communication is provided by Rohrman [75]. It considers ‘the message features’: the recipient features, social influences and context which influences individual risk assessment and management, including preventative action. Rohrman and Handmer’s publications [41,42,43,75] inform the Communication Safety Triangle and the *seven steps to community safety* provided in this article.

### Philosophy for Policy Review: Crying Wolf or Worse – Applying the Precautionary Principle

From the 1990s a strong issue of debate in risk communication has been “the right to know” [4]. Some disaster managers wish to avoid false alarms, which may cause ‘concern fatigue’. This can be seen as an institutional barrier to change (Sect. **“Institutional Barriers to Greater Community Self-Help”**). ‘Avoiding undue alarm’ is in conflict with the right to know, and the precautionary principle of ESD.

The ‘precautionary’ approach is supported by the Economic Commission for Asia and the Pacific (ECAP), the World Meteorological Organization and the Red Cross Societies. The alerting of the community and its responsible authorities must begin, at least provisionally, as soon as the existence of a tropical cyclone over the seas bordering the country is known ([24] p 16). According to ECAP et al. [24], the warning challenge is less clear for predicted localized downpours and flash flooding – how much effort should be taken to warn – what is the message, how do you keep it to the affected area, and what do you want people to do? These questions resonated in Australia after billion dollar hail damage in Sydney in April 1999, or damaging flash floods in Melbourne in December 2003.

### Precautionary Evacuations

Handmer [43] reports an evacuation of 250,000 Dutch ahead of a flood threat in 1995. Eighty-eight percent of people surveyed in broad post-emergency surveys “believe that evacuation was appropriate.” [43] p 24. In part this may be because of floods experienced two years prior. Good skills in dealing with the mass media appear to have helped in the effective precautionary evacuation. The Dutch experience showed a willingness to evacuate again in future, even though the threatening flood waters did not inundate to the level feared. This compares favorably

with the 2005 boat owners' responses to Cyclone Ingrid in Port Douglas, North Queensland (Sect. "**Institutional Barriers to Greater Community Self-Help**"). Both Handmer and Goudie's research [33,36,38,41,42,43,52,53,54] show clearly that people would rather practice (make a precautionary evacuation) than incur loss. It was treated as a learning experience. The 'boy who cried wolf' argument is not acceptable. This is an important message researchers can test and convey to partnering Disaster Managers.

### Have Clear, Consistent Messages

Part of the *seven steps to community safety* is clear, reliable, explicit languages and images. Salter et al. [78] point out that the use of meteorological category systems such as 'minor', 'moderate' or 'major' carry unambiguous information about the level of disruption likely from a particular flood. Language used should not be for the convenience of the warning agencies. Its function is to convey clear unambiguous messages to the threatened public.

### Use Past Events to 'Make the Threat Real'

The purpose of risk communication is to make people perceive the threat as real, and to successfully motivate safety-oriented action. Boughton (1992, p 6) argues that awareness of hazards and disasters can be fostered by "drawing attention to media coverage of hazards in other places". Images of large scale floods and evacuation in Holland in 1995 [42] may help prompt a future flood evacuation.

### The Changing Politics of Risk Communication

Risk communication is often laden with values and political implications (Sect. "**Institutional Barriers to Greater Community Self-Help**"). For instance, in the 1990s in Cairns, north Queensland, it was argued that the reason for not having detailed local cyclone surge inundation maps made available at the corner store level was that such information may have a negative impact on local land prices. This appeared more a political decision than an attempt at effective risk communication. In 2007, such maps are freely available on the local Council web site [9].

### Sirens or Not

No or short lead time disasters are a powerful argument for warning sirens to alert people, perhaps at 2 am, to "listen to local media now". Lives will be saved in future tsunamis and bush fires with the reintroduction of sirens, as insistent triggers to "find out more now". However, "large numbers of sirens are needed to cover populated areas and to be loud enough to be heard indoors by most people. Sirens are expensive to install and maintain and can only

provide limited information" [63] p 33. Fixed public address systems or those on vehicles may be used. People who hear such sirens will be encouraged to tune to local media, and to phone others in the threat zone, to warn them of the alert. Sorenson and Mileti [81] believe sirens are most effective if used on populations without other ways of receiving the warning. Wider use would appear prudent, especially with short lead-time threats.

### Media Roles

As community media moves from 'sensationalism' [13, 71], 2007 research by Goudie confirms that community media organizations now want to become responsible conveyors of safety-oriented information to people at risk. After the Canberra fires of 2003, where 300 homes and four lives were lost in the nation's 'bush capital', all local media signed binding agreements with emergency authorities to faithfully convey provided safety information. The Communication Safety Triangle envisages local media and householders drawing detailed local threat information from the internet, with media conveying that directly to readers, watchers and listeners. This forms the basis of future 'world best practice' risk communication. In the preparation for threat impact, the reliable information will help people make informed decisions, rather than remaining trapped and inactive in uncertainty.

Risk communication is complex, involving many values, predispositions and distorting lenses. Rohrmann's [75] fine risk communication explanations show that we may tell the people at risk, but they may not interpret as intended. Clear, consistent warnings in plain English, with clear images of the threat, showing safety-oriented behavior are needed from reliable sources. Warnings, seen on a continuum of risk and preparation actions, should be able to be discussed and reinforced with information from other sources. This is most likely to produce safety-oriented behavior, with the constrained and clear assistance of the media, as responsible agents for community safety.

Given the previously fraught nature of risk communication, Sect. "**Institutional Barriers to Greater Community Self-Help**" provides a conceptual framework, with Australian examples of current community safety theory and implementation, preceding some detailed examples.

### Integrating Theory and Implementation

This section introduces a triangular model to best ensure community safety. With disaster managers central, the three elements are the community, local media and the internet. Also, there is a continuous spectrum of seven steps,

from accepting there is a risk, through early preparation, final (ramp up) preparation, which may include evacuation, surviving the hazard impact and achieving smooth recovery. An exploration of motivation leads to a discussion of 'world view', including some insights into remote Indigenous communities [36], and recently arrived, non-English speaking immigrants. The conceptual frame and steps are intended to help guide risk communicators and potential modelers through the issues of acceptance, preparations, evacuation and functional recovery. Within CSS, all these factors have some bearing on consequent behavior.

### Changing Values and Roles

In an attempt to understand why we support or ignore certain messages relevant to our safety, and thus facilitate modeling, this section considers the Dominant Social Paradigm and the New Environmental Paradigm. A paradigm is a coherent world view, "a mental image of social reality that guides expectations in a society" ([23] p 10, [29]). Shared paradigms change. The underlying philosophy shift toward disaster mitigation generates a distinct policy move toward safe, self-helping communities. Thus, the goal of disaster managers, effective risk communicators and computer modelers becomes one of championing or demonstrating the efficacy of self-help techniques and information to and within communities.

### World Views Which Accept Responsibilities of Living in a Hazard Zone

'Why do we do what we do?' has long been a central exploration of psychology. Precautionary safety behavior, in-

Behavior is explained by:

1. a person's *position in a social structure*,
2. with *constraints and incentives* as generators of *values*, which lead to
3. *general beliefs*,
4. *world view*,
5. *specific beliefs and attitudes*, generating
6. *intent*, which helps explain
7. *Behavior*.

Evacuation as a Communication and Social Phenomenon, Figure 3  
Stern's 1995 behavioral explanation model

cluding self-evacuations, may depend on a person's world view. Stern et al. [83] developed a simple and elegant model to help explain human behavior, starting with a person's position in a social structure, with constraints and incentives which generate values. Values determine general beliefs, leading to a consistent world view, specific beliefs and attitudes, which predisposes intent and helps explain behavior ([83] Fig. 3). Loosely translated, behavior results from a linked sequence starting with: the where, when and to whom of an individual's birth, the surrounding circumstances and 'cultural sheath' of their early childhood experiences, leading to acceptance or rebellion against prevailing social norms, determining an evolving world view. Once we have our coherent world view, beliefs and values give us our 'predisposition to act', our 'intent' which precedes action/behavior.

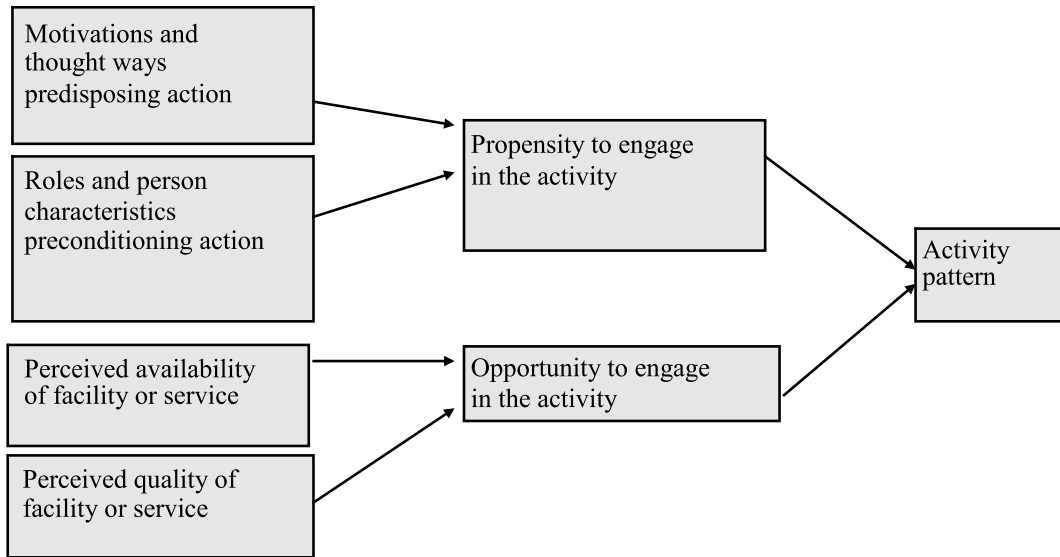
Many things now mold and modify world views, including the media, and normative world views are mutable, changeable. Aligned starkly with our shared and collective biological survival urge, and much concerted efforts by scientists and conservationists for decades; publicizing by people like Al Gore, and the authority of England's uptake of the 2006 Stern Report, there is now a global paradigm shift to the urgent need for behavioral and technological change to minimize the gross impacts of global warming, pertinent to increased disasters and needed evacuation readiness. Modeling specific hazard zones' strengths and weaknesses and broadcasting prediction simulations to the internet and TV will help mobilize those at risk.

### Why We Do What We Do

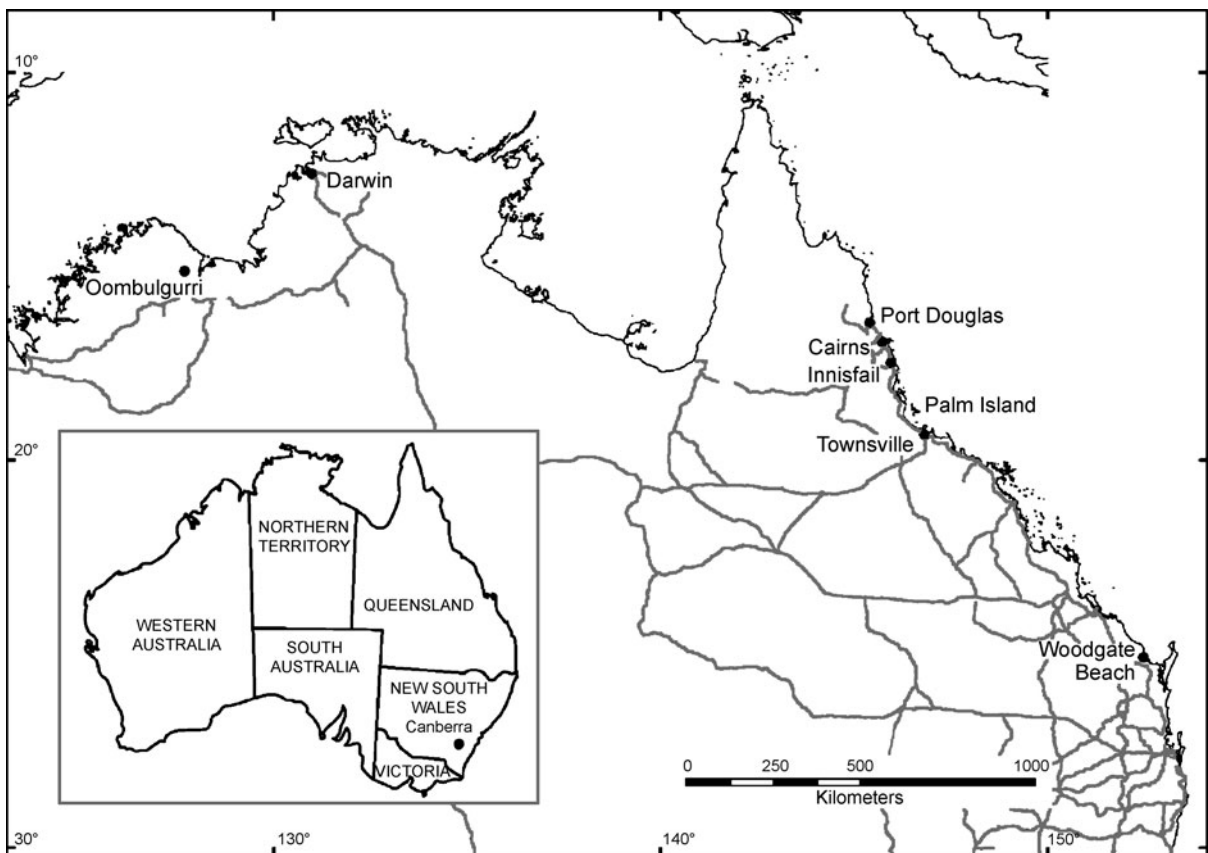
A behavioral model of causality ([91] Fig. 4) shows relationships between reported attitudes and actual behavior. However, a complex model proposed by Kitchin [51], with the strength of including social and environmental interactions shows why we do what we do: Kitchin's model includes a person's 'working and long term' memory. Internal information is processed within 'real world' context, such as cost [82], which may be processed within the 'it can't happen to me' cognitive frame [61] of subjective reality. These complex but quantifiable attributes will contribute to modeling maximum likelihood to accept, prepare and act to maximize safety.

### Learning from 18 Remote Indigenous Communities

Much of the rest of this section underlines why modelers will need to deal in great detail with 'cultural' aspects of massed responses to safety threats.



Evacuation as a Communication and Social Phenomenon, Figure 4  
Possible determinants of activity patterns (from [91])



Evacuation as a Communication and Social Phenomenon, Figure 5  
Australian study sites

The Stern model (Fig. 3) helps explain why there is such a strong sense of self-help in remote Indigenous communities [27,29,80,94] helps explain why there is such a strong sense of self-help in remote Indigenous communities [27,29,80,94]. Elders decide responses to threats, there are historic and immediate constraints, generating a value system where community members need to look out for each other.

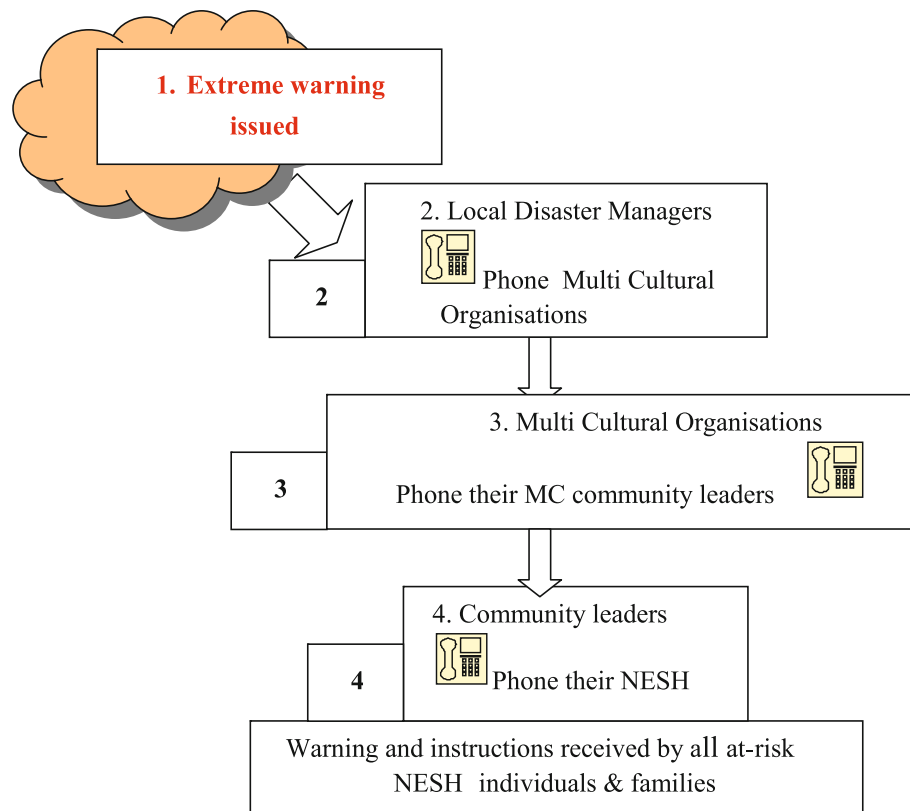
In many traditional Australian Aboriginal stories, most people drown in flood disasters, often as punishment when people do not take care of each other [36,39,64,87, 88]. If general beliefs embrace self help, including the felt need to ensure safety, and a resultant intent to achieve community safety, this should lead to safety-centered behavior.

Each community ([36] Fig.5) was not greatly concerned by weather extremes (values), but each relied on and respected their traditional reading of threats, and information from the Bureau of Meteorology (the Bureau). The world view is that flooding or worse may happen, and that the Bureau will provide adequate warning. Responses

may be inhibited by the Social Vulnerability Index (SoVI) of formula 1.

### Community Links, Phone Trees and the Web in Risk Communication for Non-English Speaking Households (NESH)

Many multicultural organizations and focus group meetings in 2005 helped develop the warnings phone tree model for NESH in Fig. 6. NESH rarely listen to the English language media. With about 30,000 NES people arriving in Australia each year [18], there is Federal government recognition of special emergency management needs [28]. This way of getting evacuation warnings through was developed once interviews revealed that practically all NESH have a mobile phone, and are closely connected with their nearest government funded Multicultural Organization and their 'community leader'; hence the phone tree. Modeling is of and for the real world, so research like this reported NESH study is needed to see what 'complex systems' may be possible. Further recommen-



Evacuation as a Communication and Social Phenomenon, Figure 6  
Multicultural phone tree disaster warnings model

dations to develop a multilingual warning web site with a simple guide to the seven steps, in relevant languages, to be accessed and used by MCO and NESH training sessions is being considered by authorities. This provides another example of the internet's latent role in effective risk communication.

The Australian Northern Territory has cyclone preparedness kits with action guides in 8 languages; the NSW Rural Fire Service has information in 27 languages. This shows the importance of modeling all possible communication avenues, encapsulated as Medium & Message: MM in the general formula for Household Preparedness and Safety Actions Index (HPSAI).

### An Holistic Approach to Community Risk Management

The seven steps for effective warnings involves community networking, 'responsible' media, and, increased web use. A 'continuum' approach to hazards which may lead to a need to evacuate is important to modelers because it is important to all people in hazard zones: they will be able to access more accurate, detailed and timely information of any looming threat, assisting Emergency Managers because a self-help public will decrease demands on formal evacuation, rescue and property protection [93]. Recovery

will be less arduous [67] if hazard impact is minimized. Reducing disaster impacts will reduce costs, to national benefit. This gives strength to supporting the very demanding goal of developing the model of formula 1.

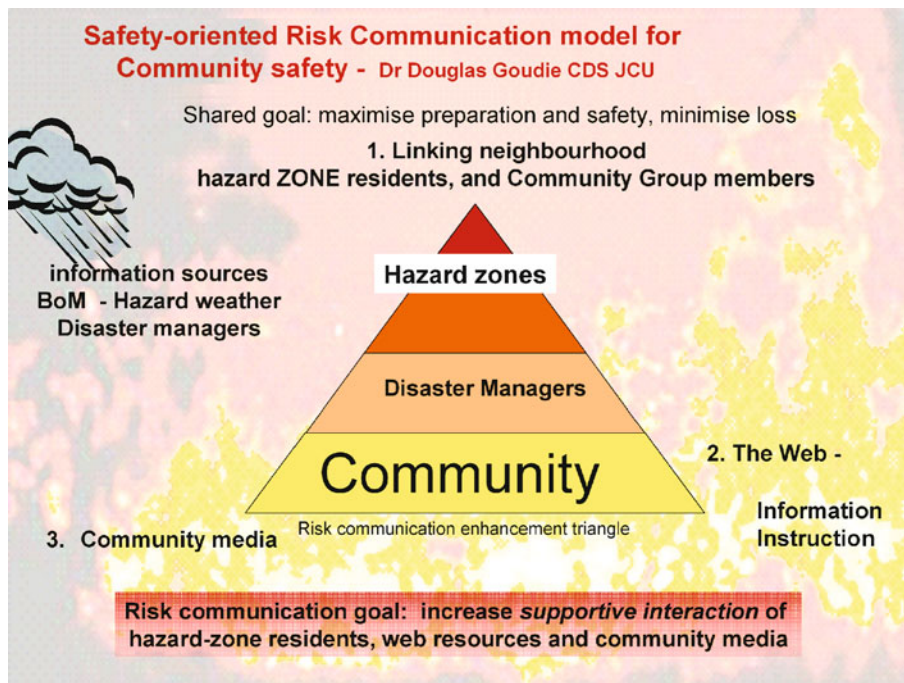
### Risk Managers

Researchers can work with risk managers to achieve "... better, timely warnings and advice on safe action during fire events" [3]. With risk managers central, Fig. 7 suggests that enhancing community links [7]; web information for residents and media outlets, and cooperation of community media with fire managers [89] will more empower householders to embrace self-help in fire safety. Figure 7, the Community Safety Triangle, with the "Seven steps to community safety" (Fig. 8) provides the conceptual frame to enable maximum safety modeling.

### Core of Risk Communication for Community Safety Through Natural Disasters

#### Building Community Links and Refining Media Delivery will Change the Household Preparedness and Safety Actions Index

The disaster safety benefits of enhanced community links fit positively with broader social policy [47]. Stronger com-



Evacuation as a Communication and Social Phenomenon, Figure 7  
The Communication Safety Triangle



**Goal: Maximise safety and recovery, and minimise loss in hazard zones.**

1. Encourage those in hazard zones to accept that the risks are real.
2. Help create an aware, informed community, predisposed to safety-oriented action, as a precaution; as a practice.
3. Encourage information-sharing and support among friends, neighbours, family.
4. Provide ‘what to do’ (action) information, via reliable sources, including web and community media delivered for background and preparation.
5. Encourage people to think right through to impact and recovery.
6. When a threat is closing in, warning messages will clearly convey: *this is real, this is coming at me. I need to make safe where I am, or move early to somewhere much safer. I will not travel during the impact period.*
7. Provide timely, effective threat warnings and fine location and forecast weather detail, and recommended local responses.

Evacuation as a Communication and Social Phenomenon, Figure 8

**Seven steps to community safety**

munity links will help ensure that threat information is easily accessed and shared; and the need to actively self-protect is internalized at the neighborhood level. Residents will benefit by enhanced community links (with such innovations as social burn offs and Community Safety Groups) in improving their general quality of local social interaction. Changing Community Resilience (CR of the formula) will then change overall preparedness, so pilot tests of change to CR will show quantifiable changes to the Household Preparedness and Safety Actions Index. Disaster web site managers will be providing a product which is rationalized to reduce national duplication of core information. This will positively change Medium & Message, MM, of the formula. Residents with few English language skills will benefit by having disaster information delivered, via the web, in language and concepts that can be internalized and acted upon.

Handmer [43] suggests that a flood, for instance is actually ‘owned’ by the communities at risk. Individuals and organizations within these communities actively seek out information and mobilize their personal networks for action. In this way of looking at the warning process, the

Evacuation as a Communication and Social Phenomenon, Table 2

**Cyclone zone people have much personal contact**

Contact with other relatives	% of total response (rounded)
Yes	25
Lots	20
Mobile contact	15
Landline Phone	30
No	10
Frequency of neighbor contact	
Often or lots	50
A bit	20
Helped/contacted during eye	15
None	15

warning specialists act as mediators between the threat and the threatened. Local knowledge is used and the whole response process remains focused on safety and loss-minimization. For this plausible vision to become fact, local capacity building will need to proceed apace with these more formal efforts to inform and maximize community safety.

The importance of ‘informal’ information sources and community links are shown in Table 2, from about 200 people interviewed immediately after Cyclone Larry in 2006 [53], showing that about 90% are deeply dependent on social and family support, if only for reassurance.

Finally for community enhancement within the CST model, there is current and lucid national disaster mitigation policy [13] [http://www.dotars.gov.au/localgovt/ndr/nat\\_disaster\\_report/naturaldis.pdf](http://www.dotars.gov.au/localgovt/ndr/nat_disaster_report/naturaldis.pdf): in the statement of the paradigm shift [13], p 13, “Principle 7 – Reform Commitment 7: develop jointly improved national practices in community awareness, education, and warnings which can be tailored to suit State, Territory and local circumstances.”

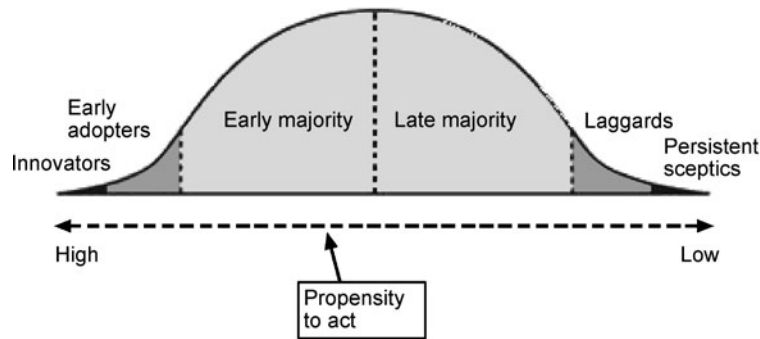
**Community Media**

“The Media: Media organizations, particularly public and private radio and television organizations, have responsibilities in ensuring that timely and appropriate warnings and advice on disasters is broadcast to communities at the request of relevant authorities. They also have a role to play in educating the community about natural disaster issues” ([13] p. 18).

**The Web**

*Rationalized web information delivery:* One national generic provider and up keeper of core information,

The stages of change from the Diffusion of Innovations theory. The vortex shape has no mathematical basis. Its purpose is to illustrate the point.



Evacuation as a Communication and Social Phenomenon, Figure 9  
O'Neil's 2004 innovation uptake model

known to all, is recommended; to which state and local government will add their unique detail. A rationalized disaster information web delivery can properly incorporate broad contents, fine (and zoom-in findable) and multi lingual information.

Simulations of predicted near-term fire or other threat movement based on the above, akin to micro scale modeling of cyclone impact forecasts will help revolutionize the way people react to fire and flood threat. With about one in three households having the internet, prior to likely power loss residents can draw on enhanced fine detail of the current fire situation, to assist in their actual decision making. Before Cyclone Larry struck, residents with web access to the Bureau copied cyclone forecast maps and distributed them to neighbors [53]. All the preceding is encapsulated in Figs. 8 and 9.

### Images Tell a Story

Images such as cyclone forecast track maps in risk communication convey more information than words alone [36,53]. With this feedback from real residents, it becomes clear that linking fire zone residents, for instance, to educative web fire images becomes an important goal (e. g. [100]).

### Community Story Telling Bonds Neighbors

The multi-pronged approach [33,52,54] fosters community 'story-telling' links [50,59,90] relating to community self-help and safety. Strengthening community bonds, working with enhanced web resources and community media [14,72] will increase community internalization of risk [31], enhancing the likelihood of safety-oriented responses, leading to and possibly including evacuation.

People need to accept the reality of the threat, indeed, feel some anxiety about the threat to help drive the intent to seek more information, or the intent to prepare [68].

With easier access to relevant web information; with greater detail of current fire behavior and *nowcasts*; fire zone residents and community radio announcers can describe the looming threat, helping timely preparations and the monumental 'stay or go' decision. Community radio, like the National public radio, the ABC, will deliver authoritative and timely risk communication directly from the refined web information.

Theory and emerging practice converge on using refined web-delivered material to households and their neighbors, and to local media, to inform and motivate residents through the whole continuum of the seven steps to community safety. The next section considers institutional barriers to change, followed by extensive research findings and lessons from Australia in disaster management to develop effective warnings and self evacuations.

### Institutional Barriers to Greater Community Self-Help

Institutional barriers to change take many forms, from 'unconsciously' avoiding consideration of the extreme event as 'too hard', to a misuse of the power relationships within bureaucracies because of fear of change, or malice. Clear examples of the 'too hard syndrome' mingled with entrenched vested interests has been denial of links between smoking and cancer, or delayed uptake of global warming mitigation.

There were Institutional Barriers for disaster managers to consider land-based flooding from torrential rains preceding a cyclone (hurricane/typhoon) [34] in many cy-

clone-prone areas. The Queensland State Planning policy [70] now explicitly refers to Probable Maximum Flood. Including maximum flooding is now part of emergency manager's internal planning base. This means the Exit Routes (ER, formula 1) are now considered in planned evacuations.

Since the early 90s researchers like Boughton [6] have suggested having drills for schools and other institutions in readiness for possible earthquakes, cyclones or other hazards. This author supports precautionary evacuation practices. There are, of course, liability barriers to easily undertaking practice evacuations. There is also uncertainty of threat, which may restrain some risk managers [86].

Institutions have cultures which may passively express antipathy to a paradigm shift toward sustainability, whilst being required to usher in sustainability [19,37]. Emergency Management Australia, the peak national body, has the slogan: *Safe, sustainable communities*. However, there is a multitude of conservative forces representing the Dominant Social Paradigm restricting innovation, despite sustainability policy. People engaging in sustainability implementation may come into conflict with institutional representatives of the old paradigm.

With most development taking place in urban settings, concepts of urban sustainability attempt to merge two different fields of human endeavor – how we modify our landscape, our built environment, and how we behave in that environment. Disaster mitigation is obliged to fuse these two seemingly disparate fields. There may be resistance to that. That resistance needs to be included in any mass disaster movement modeling. The following subsections indicate some reasons for IBs which modelers may need to consider in the IB factor.

### **Political Insecurity, Real Estate Values, or Undue Alarm**

The tension between people's right to know, government duty of care and politicians' perceptions of probability of risk on 'their shift' form a complex interplay. If maximizing safety is the shared goal, all risk communication theory suggests people will behave appropriately with the realistic threat information, motivation and 'how to' instruction on safety-oriented behavior leading into, through and recovering from a natural disaster impact.

### **Paradigms of Politics: the Real Estate Industry and Vested Interests**

Broughton [6] argues that if people know of the threats, they are likely to support politicians who make sound de-

isions for community survival. Some plans are made but they are hidden away for various financial and political reasons. This is a case in which attitudinal changes on the part of those communities may change the priorities of the decision makers and promote the interests of the community.

Those most empowered to assist in implementing projects that require independence, initiative, local links and knowledge may be the ones who prove most obstructive to innovation [1,48,56,69].

*Action Research* [16] sets out to research, develop and document socially and environmentally desired outcomes within sustainability principles. In this inclusive and 'engagement-focused' research, gaining support from the top is crucial to the success of any organizations' efforts at 'social change' [66]. If innovation uptake is viewed as a normative uptake process [12,65] and Fig. 9, then the gradual, long term paradigm shift to sustainability [30] is plausible and has a conceptual frame.

Uptake of the Communication Safety Triangle and the seven steps to community safety (Figs. 7 and 8) will make them the disaster managers' 'norm' over time, simply because that is the direction of social evolution in disaster management. It fits social policy and advanced disaster management approaches, and the technology and willingness of regional media and residents is there. All that is needed is the testing and roll-out by more innovative disaster managers. The same applies for Community Safety Groups and variants of *over-managed, locally controlled near-house "social burn-offs"* in the case of fire risk management.

Bureaucratic processes need to support the stated goals of their own work units, but individuals may be ambiguous, contradictory, belated, bullying, ill informed and, due perhaps to time pressure or arrogance, quite destructive. Such individuals may be 'corporate psychopaths' (<http://www.abc.net.au/catalyst/stories/s1360571.htm>). Instead, innovators may be seen as threatening; troublesome. Any innovative pilot project must develop strategies that will maximize change within their unique situation. Ideally, that happens within a supportive parent organization. The next section discusses the emergent issues of effective risk communication and precautionary self-evacuations.

### **Experiences and Lessons – Some Case Studies**

This section considers disaster impacts and impacts from north Queensland and from fire zones in SE Australia to illustrate the generally positive points of approaches already outlined: living flesh for modelers to understand how sub-

tle, complex but do-able successful community preparedness and willingness to act for maximum safety can be. Researching these events, gaining community and disaster managers feedback on risk communication; working closely with the Australian Bureau of Meteorology over 15 years, through the Federal disaster ‘paradigm shift’ to mitigation ([13] p 13) all inform the emergent CST and seven steps to community safety.

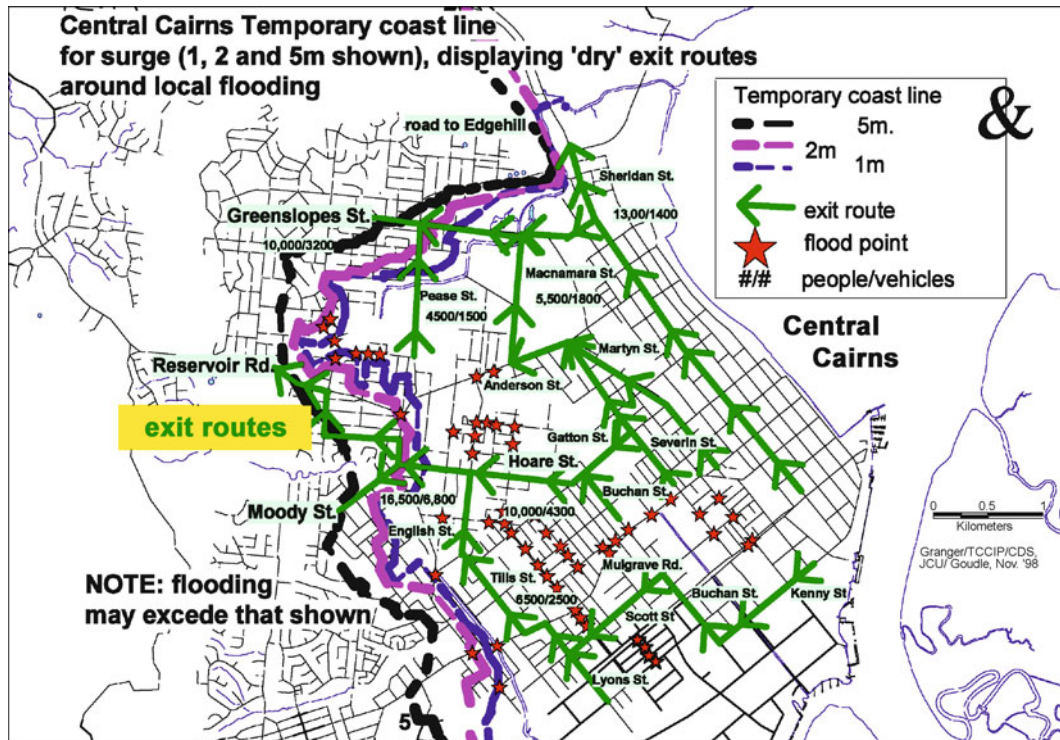
**Cairns and Storm Surge Considering Exit Routes (ER, Formula 1)**

Cairns City, North Queensland, is centered on land less than 2 m above high tide, and subject to cyclone (storm) surge of up to 5 m. A storm surge tracks just behind the eye of a cyclone, a low mound of sea water, perhaps 50 km wide and up to 3 m above normal sea level. It may flow overland for perhaps 4 h, as destructive winds to 280 km/h tear at structures, and churning seas; both laden with pounding debris, behave as battering rams and missiles. Because cyclones have such long lead times from modern electronic detection to landfall, results of this deadly combination [60] will be widespread destruction and loss

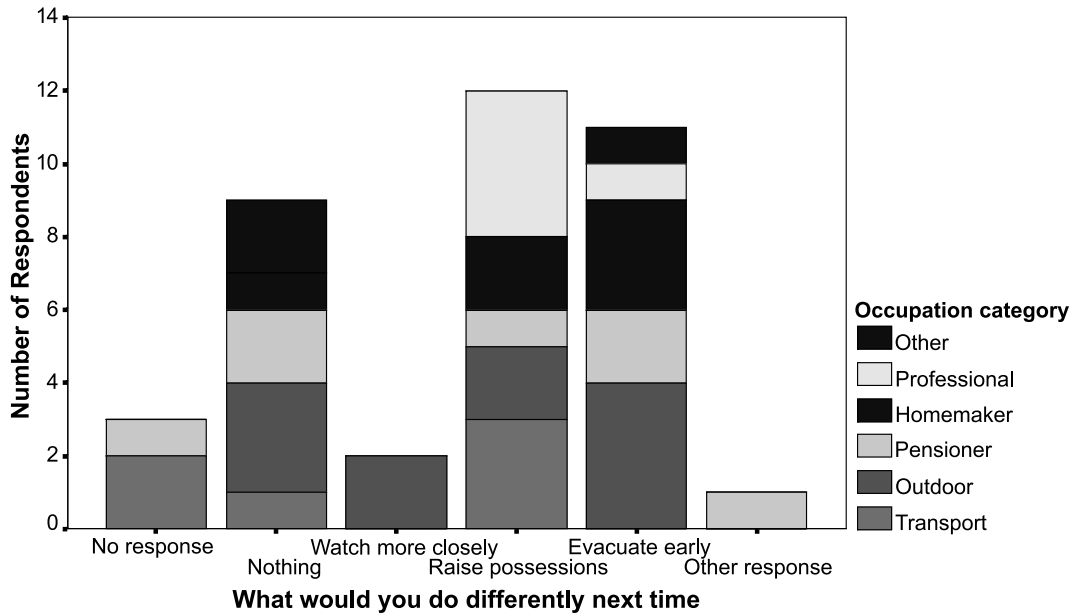
of infrastructure, but not life, with precautionary evacuations [38,72]. At the macro-observation scale, people and vehicles in motion within confining environments tend to behave like fluids, capable of modeling. Constrictions of flow paths cause congestion, as described by Helbing et al., (► **Pedestrian, Crowd and Evacuation Dynamics**), so early and precautionary evacuation will help minimize the likelihood of grid-lock.

Land-based flooding may be a core issue to effective evacuation [34,38,95], with up to 40,000 people in the vulnerable central city area and northern beach suburbs needing early, precautionary evacuation [79], Fig. 10. Cairns City Council now has a storm surge map on its public web site ([101], posted in 2006), like the public and informative flood map for the Redlands Shire, SE of Brisbane, Queensland ([102]).

Within the philosophical and moral frame of people’s “right to know”, these local governments are using the internet to inform people they are in a hazard zone, the first step in making the threat real to those residents. They have made the paradigm shift to providing fine-detailed background information which says to residents: “you are in a hazard zone, you may need to do things. Lis-



Evacuation as a Communication and Social Phenomenon, Figure 10  
 Cairns flood prone evacuation routes and schemata of wave reach and temporary coast line from 1998 information



Evacuation as a Communication and Social Phenomenon, Figure 11  
 Cloncurry reflective of likely responses to future floods

ten for warnings and be prepared to act in a precautionary way.”

Figure 11 shows that after a devastating flood in 1997, the flood-affected residents of Cloncurry, NW Queensland town would do things differently, with better warnings, when faced with rising flood waters. Remote automatic flood monitoring devices were requested, but the local downpour over a fully flooded, vast and flat landscape appeared to be the cause of the flood rising 2 m higher than any flood in the prior hundred years. Precautionary evacuation of the low-lying homes would have prevented much heartache and loss of valued possessions [52].

In 2005, north Queensland was threatened by cyclone Ingrid. Newspaper-reading residents were left in no doubt about the threat (Fig. 12); a good example of the clear warning role played by the media, and non-language image used to convey meaning (Sect. “Discussion”).

A portion of a Weather Bureau warning media bulletin for Cyclone Ingrid follows, showing, verbatim, what the nation radio broadcaster, the Indigenous Radio Network and most responsible media outlet relayed on. This is clear information, including the possible impacts (M& M of formula 1).

Media: For immediate broadcast. Transmitters in the area Cape Grenville to Cooktown are requested to use the Standard Emergency Warning Signal.

TOP PRIORITY

TROPICAL CYCLONE ADVICE NUMBER 14

Issued by the Bureau of Meteorology, Brisbane

Issued at 10:56 am on Wednesday the 9th of March 2005

A Cyclone WARNING is current for communities between Cape Grenville and

Cooktown. The warning extends inland across central Cape York Peninsula.

A Cyclone WATCH is current for coastal and island communities on the eastern Gulf of Carpentaria between Weipa and Kowanyama.

The watch south to the Gilbert River Mouth has been canceled.

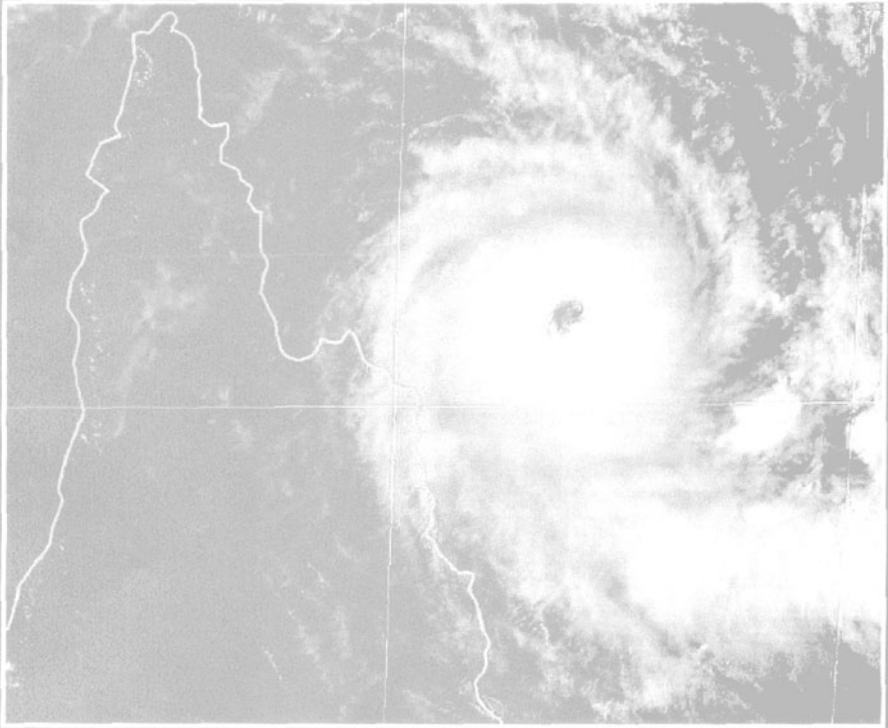
At 10:00 am EST SEVERE TROPICAL CYCLONE Ingrid, Category 4, with central pressure 935 hPa, was located near latitude 13.5 south longitude 145.5 east, which is about 140 km northeast of Cape Melville and 260 km east of Coen. The cyclone was moving westward at 11 km/h.

Severe Tropical Cyclone Ingrid poses a serious threat to the far north.

Queensland coast with very destructive wind gusts to 280 km/hr near the center. Gales are expected to develop between Cape Grenville and Cooktown

**Townsville**  
**Bulletin**

Established 1881 Tel: 4722 4400 Classifieds 4722 4466 Wednesday, March 9, 2005 \$1.10 Freight extra



# CATEGORY 5

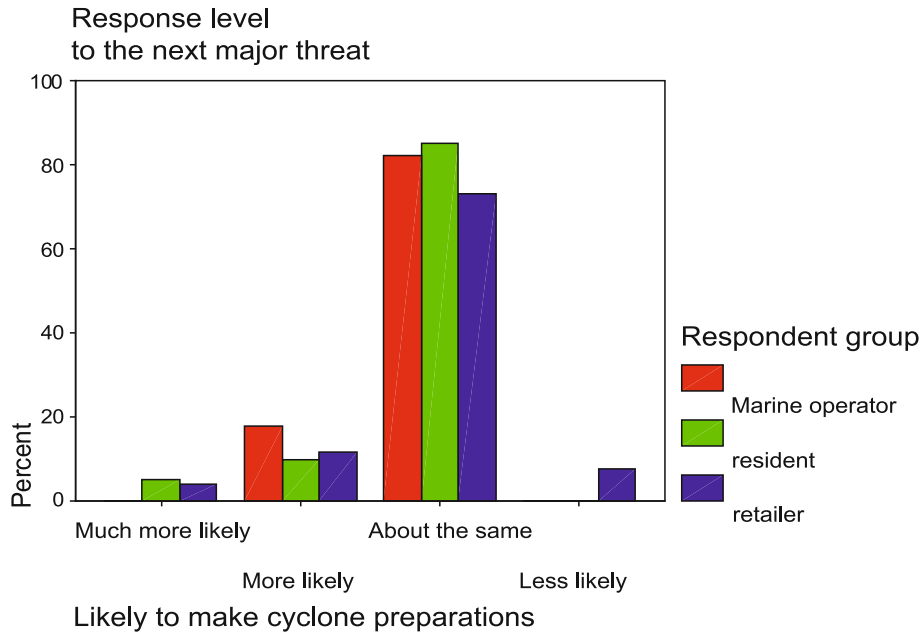
## 'A destructive core twice the size of Tracy'

AUSTRALIA'S worst cyclone in 30 years is bearing down on North Queensland as thousands of residents prepared for a double whammy of a massive king tide in the region. Category five Cyclone Ingrid was last night about 240km northeast of Cooktown, packing winds of up to 300km/h. Cyclone expert Jim Davidson said Ingrid was half the size of Cyclone Tracy, but had a core twice as big. "The destructive core is about twice the size of Tracy," he said. "Anyone in the path of this cyclone should get under cover. This is a one-off, the environmental conditions were perfect for this cyclone to form. That's why we have the intensity it is." Continued - Page 4

Evacuation as a Communication and Social Phenomenon, Figure 12  
Media coverage of threatening Cyclone Ingrid, 2005

during the afternoon. Destructive winds are expected between Coen and Cape Flattery overnight. The very destructive core of the cyclone is expected near the coast between Coen and Cape Melville on Thursday morning.

Coastal residents between Coen and Cape Flattery are specifically warned of the dangerous storm tide as the cyclone crosses the coast early Thursday. The sea is likely to rise steadily to a level significantly above the highest tides of the year with damaging



Evacuation as a Communication and Social Phenomenon, Figure 13  
 People of Port Douglas appreciated a precautionary evacuation

waves, strong currents and flooding of low-lying areas extending some way inland. People living in areas likely to be affected by this flooding should be prepared to evacuate if advised to do so.

Very heavy rain can be expected to develop on the coast and ranges north of Cooktown.

Research in Port Douglas, NQ was conducted immediately following Cyclone Ingrid, to help clarify the ‘Boy who cried wolf’ hypothesis about ‘concern fatigue’ over precautionary evacuations (Sect. “**Institutional Barriers to Greater Community Self-Help**”). Feedback from marine tourist operators and tourist businesses that removed about 60 large vessels to safe, up-creek moorings, or fully shuttered their premises and raised all stock. These preparations were arduous, and costly. The cyclone veered away from Port Douglas, with virtually no impact. Despite that, people were glad of the precautionary evacuation, as a practice (Fig. 13), reinforcing the findings of the Netherlands flood study [43]. People appreciate practice, as part of the new paradigm of a precautionary approach to hazards.

**From Those Impacted by Larry, 2006**

After Cyclone Larry, a powerful Category 4/5 severe cyclone which damaged the Innisfail region, Goudie lead

a social science research team into the impact area, interviewing households for 4 days (King et al. [53,54]), from the 150 householders survey, focused on risk communication, we learned: as with fire zone residents and those who experienced Cyclone Larry, that most people, once they have experienced a major disaster, maintain a healthy respect and inclination to act ahead of any future such threat. Hence the advocated merit of encouraging people with experience to ‘tell their story’; and community (Media and Message and Community Resilience of formula 1) to help make the threat real to others. The next section is, perhaps, at the leading edge of where disaster management will go: living community self-help. Such communities can provide pilot locations to ‘calibrate’ formula 1.

**Woodgate Beach and Community Safety Groups (CSGs)**

This section reports a collaborative process between the author and many others, mainly residents of Woodgate Beach, Queensland (Fig. 5), to develop a *Preparation and Evacuation Plan* for residents, formal response groups, the Local Government Disaster Management Group (LDMG) and Shire Council.

Through the 18 month consultation and development process, the plan needed to be realistic and achievable, relying mainly on locals working together under the LDMG

and local SES group, aiming to optimize each element of formula 1.

Community networking and self-responsibility for community safety is profoundly developed in the isolated, 700 strong, Central-coast Queensland settlement. The volunteer community safety groups were made up of dedicated people. Goudie led development of a community-based evacuation plan, aiming to: ensure maximum preparation and minimum impact on property and residents; optimizing formula 1.

There were 5 meetings, of up to 80 residents and representatives of all formal response groups, weather and earthquake experts. Meetings included researcher-led two-day 'table-top' evacuation exercise. Woodgate Beach did not develop an evacuation plan, but a preparations and evacuation plan:

### Preparations and Evacuation Self-Help Approach

To nurture aware, informed residents preparing to be safe through, and recover from natural disaster impacts.

### Overarching Evacuation Approach

- Identify vulnerable areas or houses
- Evacuate caravan park, visitors, people at risk unable to easily move themselves

### The Planning Process:

1. DEFINE THE THREATS – Fire, flood, wind, cyclone surge, earth movement, tsunami.
2. MOVE FROM WHERE TO WHERE? Fires can be fought; floods, storm surge and cyclonic winds, tsunami or earthquake cannot be. In all threats, make sure your property is as secure from impact as possible.
3. IDENTIFY THE VULNERABLE – Which buildings, infrastructure and people may be in the threat zone.

**Threats and Treatments** An all-of-community approach has developed an annual round of public education projects, press releases and pamphlets to inform residents and tourists of the annual cycle of dangers to be wary of, from cyclones to fire care and management. The Council newsletter will be a consistent source of information on threats and how to minimize risks (M & M), from the flying debris of a wind storm to being patient at a flood-swollen creek crossing.

A sequenced approach, where the aged and vulnerable are moved first from the highest risk areas will be taken, as a practice, as a precaution. To highlight the stages of

disaster and possible evacuation planning, the background preparations phase is included in this article:

### Background Preparations

1. Woodgate Beach residents recognize and act on the need for background preparations to minimize the impacts of all hazard impacts. This includes property maintenance and upkeep. This maximizes PPR of formula 1.
2. Provide newcomers with an information pack, including a copy of extracts of this Community preparations and evacuation plan (CPEP). All community members, including tourists, the elderly, infirm, and needy are incorporated into this CPEP. This enhances the KB of formula 1.
3. Provide dot points on evacuation for local residents in the *Disaster Preparedness Information Kit*, delivered to each household.
4. Expose tourists to the essence of local threats and what will be expected of them: leave early, unless they or their vehicles can actively help, under direction.
5. Define safe shelters – preferably with friends or compatible households. Organizing possible billets for any major impact on portions or all of Woodgate Beach can form a key function of the Community Safety Group. This is the CR of formula 1.
6. Go through whole plan, and address matters like the caravan park needing auxiliary power for fuel pumping before the cyclone season.

**The Community Safety Group Approach to Disaster Management** This approach is detailed to provide a guide for researchers to use as a framework for other communities:

### The Community Safety Group

**Purpose** encourage:

**1. Early Warning Alert** The CSG is an affiliation of existing community groups and neighborhood-level residents who make first contact with 'walking-distance' neighbors as soon as anyone hears of a warning that a natural disaster may be approaching their area.

**2. Final Preparations (Ramp-Up) Activation** The neighbor-level CSG will provide early local motivation for final safety preparations.

**Recovery** – Thinking Through to Recovery (TR) of formula 1.



Recovery is now seen as part of the preparations package, rather than just looking to minimizing impact. The developing approach is to see the whole threat event as one continuous process: from awareness and structural preparedness, through initial communication of threat, to final precautionary preparations and impact and rapid recovery to a fully functional community.

### International Snaps

The Center for Disaster Development within The Northumbria University, Newcastle on Tyne, coastal north east England specializes in recovery and response with an emphasis on development long term recovery and resilience-building. Embedded in this approach is to undertake mitigation. Interviewed by Goudie in May 2005, the Director reported “We use the approach that local knowledge is always drawn on, and that people involved should be in control”, KB and CR of formula (1) and agrees “all disaster management is under the umbrella of sustainable development”. Other interviewed staff support precautionary evacuation as a practice, and, independent of the media, an alert signal for people to tune in to the media.

### The Shetland Islands

Like Australia’s emerging approach to self-help communities, The Shetland Council’s Emergency Management

Planner (2005 interview) gave strong practical support to the approach of an informed, aware, self-activating community ready to act on reliable, clear, how-to emergency warnings, ranging from making safe if intrinsically out of the main impact zones, to early ‘self-evacuation’. This convergence from differently evolved and slightly different disaster management systems is grounds for optimism that the evacuation approach and formula expounded in this article has widely applicable merit.

Remote communities like the Shetlands have lessons for the mainstream in taking responsibility and perform being self reliant and oriented to robust self-help. Such isolated communities represent matured examples of the informed, aware communities, predisposed to precautionary action that mainstream populations now aspire to. Community-building is an international aspiration, achievable in urban settings.

### Hurricane Katrina

Hurricane Katrina, USA late August, 2005 (Fig. 14) has deliberately not been included in this evacuation analysis. With days of clear warning, general formal approaches to precautionary action which underpin the CST and the seven steps to safety and recovery were underplayed.

Figure 14 is included to remind all readers that hazards are real threats to real people in real hazard zones. If all the parameters to maximizing formula 1, such as Institutional



Evacuation as a Communication and Social Phenomenon, Figure 14

What no-one should have to stay through – Katrina '05 [http://news.bbc.co.uk/1/hi/in\\_pictures/4194032.stm](http://news.bbc.co.uk/1/hi/in_pictures/4194032.stm)

Barriers are not optimal to safety, great distress and loss can ensue.

This section has provided a cross-section of disaster threats, evolving reasons to empower communities, and some factors to include in modeling greater community safety. After discussion in the following section, some recommendations and future directions are provided.

Modeling not only the natural hazard but the various social and communication parameters will provide a good basis for not only simulating impact; say, of the extent of flood waters; but will also highlight which impact areas are the most and least likely to properly act in their own best safety.

## Discussion

### Risk Communication Theory and Residents in Hazard Zones

People need to know that an impact is possible before they will willingly evacuate. The concept of risk characterization [42,46,68,75,77,83] makes clear that people need a practical understanding (*the possibility of impact is real, I fully accept that*) to then illuminate practical choices. ‘Internalizing’ that *a threat is real and what you need to do to maximize your safety* may be the core goal of risk communication. This internalization and safety-oriented action may lead to a choice of precautionary evacuations. People need to make informed decisions, thus leading to decision-driven activity.

### Delivering Real-Time Warnings

The idea of subjective uncertainty [74] is well displayed by fire-zone residents who, despite all authority efforts to have them commit to an early decision to stay and defend or to leave their properties early, recurrently reported that they would decide on the day whether they would stay or go. There are tragic and recent international examples of resident’s decisions made on too little understanding or information.

The bushfire ‘stay or go’ decision point explored in 2007 research by the author with Australian fire-zone residents clarifies that decision impact on those under threat – packing up their valuables, children and pets and fleeing their house, almost certainly putting it at risk of burning down; as opposed to staying in their house and experiencing the terror of the developing bushfire, was a decision many preferred to delay. Further, many acknowledged that to leave early with all of the disruption, only to learn that the fire was not an actual threat to their property helped induces people to delay that decision point.

Subjective uncertainty is a psychological problem. The outcome of current research is to encourage the threat information gatherers and providers; the weather bureau and disaster managers, to refine information detail and use all currently available modes of dissemination to provide the threatened with the maximum amount of fine details to make that evacuation decision in an informed and timely manner. This is a clear example of where theory and the practical views of hazard zone residents converge.

The 2007 southeast Australian fire zone research shows that people who are more obviously at risk from bushfire are far more prepared than people who are at risk from an occasional but as potentially destructive bushfire. The literature [13,24,31] shows that the less frequent the event, the less prepared people are likely to be. Aligned to the goals of this publication, the formula explained in this article on the realities and complex elements of combining physical with human geography to the ‘social good’ of maximizing safety around disaster impacts is a great potential application of complexity and systems science. This complexity includes the Media (Media Support MS of formula 1).

### The Media

The media plays an ongoing role in socialization and the development of normative values. Dominick [20] argues that the media plays a key role in the cognitive development of individuals. Cognition is the act of coming to know something. Mass media can play a defining role in people’s awareness and responses to disaster threats. The CST embraces local media as an active agent in providing needed local information. Local media outlets can draw on and enhance Internet information to inform readers and listeners of ‘how to’ actions to maximize their safety. In this way the mass media can help clarify facts to help make contentious decisions (which may result in life or major property loss) by providing the fine details to reduce the uncertainty surrounding decisions that householders under threat need to make. The seven steps to community safety (Sect. “**Integrating Theory and Implementation**”) underline this core need for individuals in hazard zones to accept they are at risk and for the Internet and local media to act as providers of relevant, local, timely information for informed decision-making.

The media can play a powerful role in mobilizing communities [58], but they need to have accurate and timely information from disaster managers. The information which fire zone residents in southeastern Australia in 2007 asked for from the Internet is in full keeping with the-

ory, and a recognition that greater access to current facts will reduce uncertainty in decision-making.

### Triggering Action: Modeling and Simulating This Geographic, Communications and Psychological Complexity

The reason step one in the seven step process is so important is to counteract a psychological defense against accepting threat. We need to have a sense of personal invulnerability to get us through each day [75]. If we were scared of all possible problems we would cease to operate. The risk literature from Handmer, Goudie, Rohrmann, Salter [34,42,75,77] and many others underlines the importance of the clarity in the description of the threat and safety oriented action. Thus Acceptance they are In a Hazard Zone (AIHZ) of formula 1 is central to possible consequent safety-oriented behavior.

One reason given for risk communication failing is that it is not clear what should be done. The seven steps include a requirement that 'what to do' information be an integral part of any warning as a prelude to evacuation. The clear message is that the process of people accepting that a risk is real needs to well precede any actual impact. This is underlined by Kaspersen & Stallen [49], who also stress the importance of time perception and time horizons. Hence the importance of Knowledge Base of the hazard and safety-maximizing behavior (KB) of formula 1.

### Timely Preparation, Final Actions and Evacuations

Timeliness of warnings is paramount to the effectiveness of warnings [85]. In the early 1990s there was much psychology research on perceptions of risk probabilities, individual psychometric studies on perceptions of future time, time orientation, planning horizons, and prediction of future events. In 2007 Australian fire zone residents say they want fine, detailed information as a threat approaches, so the uncertainty of what they may face is minimized. Their planning horizons are generally well oriented to maximum safety, but the crucial 'stay or go' decision will be based on information as close to impact as allows them adequate time to act safely.

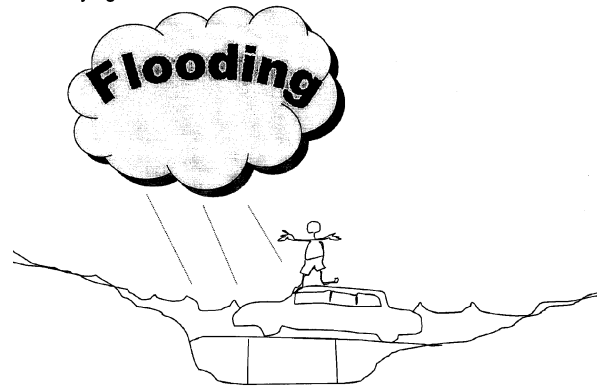
Svenson [85] and others make clear that it is not only what you may chose to do, but also when you do it. In the case of remote area warnings of flooding, for instance, when flood waters may take a week to block down-catchment roads, traveling earlier than planned may be the best way to avoid the fate of the truck shown in Fig. 15. This is an example of an 'Active warning image', along with the image of the person standing on the car roof in the flooded road crossing (Fig. 16). Images like these are of use to re-



Evacuation as a Communication and Social Phenomenon, Figure 15

Nearly crossing a flooded road – Tully NQ 2004, From: Townsville Bulletin, 28/4/4.

Conveying weather threat information - flood



Evacuation as a Communication and Social Phenomenon, Figure 16

An image dissuading people from driving in to flood waters

mind people to travel before or after expected flooding – not during the flood.

In 2007, in a two day threat and evacuation exercise in Woodgate Beach operated on the underlined imperative, from the outset, that evacuation ahead of a storm surge would need to be completed not as cyclonic winds struck, but six hours before landfall. Once the winds gusted above 100 km per hour, all peoples, including disaster managers, must be in safe shelter above cyclone surge height.

Disasters and evacuations are issues of people, information, time and space. People in threat zones are entitled to accurate and timely localized information, detailing the threat and likely time-linked movement. People also want to know, via the internet or battery operated radios, what authorities are doing. That information may influence their own actions. What they ask for is the enhancement of their 'risk decision landscape' [83].

In 1974 communication management of and responses to disasters in Darwin and Brisbane shocked Australians. Darwin's Cyclone Tracy killed 65 people, and the Brisbane floods killed 16 [10,11]. These catastrophes taught Australian emergency planners much about the importance of

effective warnings, sharing honest, complete and open information in a timely way to emergency managers, emergency workers, and those at risk to ensure sound preparation and responses. The Australian natural disasters of 1974, with major flooding elsewhere in Australia that year, also reinforced the importance of community and family ties to get people through the often profound emotional trauma allied with major natural disasters. The detailed documentation of Cyclone Tracey can help guide development of predictive models of disaster preparedness, communications and responses. With so few well documented cases, a Bayesian Logic approach could be used, some well documented disasters, communications and response cases to develop the model, others to test and refine it.

The more recent and current disaster impacts all point to the importance of communities being properly informed. Indeed, the literature on knowledge, risk and inclusion of residents in the bushfire planning process is well described by Goldammer [32]. Issues of increased encroachment on to bush edges along with climate change are causing increased international concern over fire threats, often to places without much collective wisdom over the reality of those fire threats, or the nature of the strategies needed to maximize community safety.

South Eastern Australia is seen as one of the most bushfire-prone environments in the world [92]. Boura [7], an Australian fire manager, describes the development of community fireguard groups in Victoria. These groups are managed and instructed by the Country Fire Authority, but they are residents within very localized neighborhoods. They have their counterparts in the ACT and some other Australian States. They draw on detailed, often technically advanced information from the weather bureau and fire agencies to increase local awareness and action. This is a real playing out of the theories of social amplification developed by psychologists and risk communication theorists [73]; advanced planning techniques [16] and aligned with the way many feel about the direction of emergency management in Australia [15].

### Linking the People with the Message

The foci on culture, community and social frameworks, either formal or informal neighborly links, is considered by Douglas [21]. Douglas argues that there is a need for people to understand that risk and danger exist where they live, despite a low probability that an impact may occur in any given hazard season. There is a large body of psychology which considers why people ignore clear messages which may maximize their safety [12]. Douglas speaks of artificially distorted world views. Douglas posits such bias

is rooted in over-simple views of heroic and bourgeois fiction. Changing such a normative world view has been discussed in this article in terms of a paradigm shift, well displayed by the Australian Government [13]. A model constructed from formula 1 needs to include all the subtle complex issues of social profiling.

Woodgate Beach (Sect. “Experiences and Lessons – Some Case Studies”), perhaps because of physical isolation, is fully embracing self-help, also displaying a deep culture of volunteerism that can be nurtured and emulated in ‘urban villages’. Sustainable urban planning concepts of nodes or activity centers are now well evolved as the hub of ‘urban villages’ [35]. What is happening in Woodgate Beach can be used as a model for any formula of developed social collective will to self-help, ultimately whether urban and surrounded by other neighborhoods, or more physically isolated.

Like the Ferny Creek (Victoria) community who agitated for a fixed bushfire siren after three of their neighbors were burned to death in the flash fire of 1997, community action needs a few individuals of vision and drive, residents of ‘neighborhoods’ can initiate and embrace safety-oriented behaviors and structures. The best thing the institutional systems can do is draw out, nurture and encourage these individuals to mobilize a focus on gaining threat knowledge and acting to increase community responses. Part of the community engagement in the fact they are in a threat zone, as step 1 to action, is shown in Fig. 17; a great road-side banners initiative by one section of the Victorian Country Fire Authority to make the fire threat and implications real for fire-prone residents. Figure 17 shows, in few words, that if you intend to evacuate, do it early. Figure 17 says that the support and creativity to maximize communication effectiveness (the opposite of Institutional Barriers) helps contribute to a positive Medium and Message (MM) of formula 1.

### Traditional Weather Warnings

Traditional weather warning signs include ant movement for looming flood, and general bird movement for strong winds, including cyclones [64,87,88]. On Palm Island, when the birds and animals go quiet, it signals that a major storm may be on the way [36,39]. The buildup to the summer monsoon rains is the universal experience of still, humid and hot conditions, continuing in intensity and cloud until the rains come. All cultures in all hazard zones have their traditional knowledge. Cultural Knowledge Bases (KB of formula 1) needs to be modeled into any mathematical prediction of likely propensity to respond to a natural disaster threat.

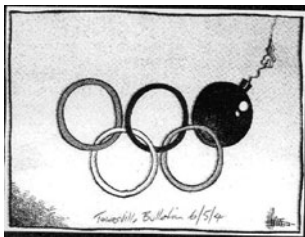


Evacuation as a Communication and Social Phenomenon, Figure 17  
Initiatives to empower residents to be prepared for fire

### Words and Images

Word and image use are critical to effective communication. A newspaper graphic ahead of the terrorist-threatened 2004 Olympic Games (Fig. 18) conveys much about the world we now occupy, and uses no words.

The need for clear messages, most likely to provoke a precautionary response, is supported by the risk communication literature of Sect. “**Effective Risk Communication**” and the communication and cognitive theory of Sect. “**Integrating Theory and Implementation**”. If this knowledge is melded to requests of remote Indigenous peoples and NESH, the safety goal of clear, plain words



Evacuation as a Communication and Social Phenomenon, Figure 18  
An image of terror in 2004 – Source: Townsville Bulletin

and images will become the risk communication norm. There are many examples of images conveying meaning, such as the internationally used figure running upstairs, with an arrow; signifying an evacuation route. The symbol for tsunami; a stylized figure racing up a steep slope with a very large wave following, threatening to engulf them is also without words, but conveys all we need to know about evacuating ahead of a tsunami: get up slope immediately.

### Images and Warning Maps

Risk communication is usually about attempts to prompt considered action by a person or community. Effective communication should make the future threat real in present thinking. Alternative responses should be outlined, along with likely consequences. To further prompt a considered and active response to the ‘action warning’, the consequences of inaction or a range of defensive actions should be made lucid. ‘Preferred’ behavior should seem reasonably attractive to ‘target individuals’ [85]. The theoretical overview of risk communication in Sect. “**Effective Risk Communication**” explains why the message must be clear to the target audience. It needs to have some cognitive content to get people thinking about how it may impact on them, and what the alternative outcomes for

them may be if the predicted impact strikes where they are. Stimulus and local risk simulations, as embedded in the Redlands shire flood map [102], should be saturation broadcast into all hazard zones. The CST should be attempted with at-risk groups; people in threat zones, encouraging them to properly think about the real threat and to indicate to people what inaction may bring. A range of safety-oriented actions should also be presented.

### Modeling Risk Communication into and Through Communities

Efforts have been made since the 1960s to see how well people understood the hazard. These studies continue [5]. The model of hearing, understanding, believing and feeling that the information is personalized so that those at risk will act has been well understood since the early 90s [81]. The general issues of credibility [18,69,70] still apply. Studies reported by Sorenson and Mileti [81] show increased knowledge as a result of risk communication efforts. People do become more aware of hazards and their personal place within the hazard threats. Unfortunately the link between knowledge and behavior remains tenuous [35].

Knowing that the normal first warning of major disruptive weather comes via the evening TV news in remote settlements, simulations will carry a high embedded likelihood of safety-oriented community response.

Sorenson and Mileti [81] showed that the believability of warnings increases as people get more warnings from officials with high credibility, and that women tend to believe emergency warnings more than men. They also showed that people higher up the socioeconomic ladder tend to believe warnings more than their counterparts. Minority groups have lower than average belief in warnings while people with a high knowledge of the hazard tend to find warnings more credible.

### New Cultures to Hazard Zones

The issues discussed in this article from remote Indigenous communities and from non-English-speaking households make clear the importance of using plain English. Work with representatives of both groups made clear that plain English and images were necessary to convey the concept of danger to people in hazard zones, concurring with the expositions of Douglas [21]. The simpler the language the better. Douglas argues that even the word 'risk' could be dropped and 'danger' used instead. Goudie's work with recent Somali arrivals in Australia showed that, like Japanese tourists and many other people from non-English-speaking backgrounds, they understood the word

and concept of 'danger', but did not, for instance, understand the word 'severe'. The importance of language clarity to the intended audiences in the MM expression of formula 1 needs emphasis.

For risk communication and evacuation advice to penetrate to all people in hazard zones, the language and images must be clear, simple, and compelling. If risk communicators construct campaigns of engagement which work for such 'marginalized' groups, the 'mainstream' can be easily included in that campaign.

### Playing the Odds

Renn and Rohrmann [74] make clear that the assumption that risk judgments and evaluations are universal processes independent of social status or national heritage are unreasonable. Goudie's Australian research with multicultural organizations shows most clearly that there are many populations in hazard zones and that each of those groups must be addressed in ways that make the hazard real to them. Findings from the Cyclone Larry research [53] also show that there are many groups other than cultural, such as 'social isolates' – the socially disenfranchised, who need special attention around disaster threats to ensure their safety.

Douglas [21] argues that risk perception and thus response is also an issue of moral and political issues. Westerners tend to consider the probability of an impact in terms of gambling whereas other cultures may need other triggers to internalize the threat and motivate them to action. The concept of a low probability event needing to be taken seriously is discussed by Douglas and others. Residents of Florida are well used to a near-annual evacuation ahead of frequent cyclones. Residents of the coastal communities of Queensland may be less likely to take the warning seriously, because they believe that there is no real chance that they will be personally affected by a cyclone. The whole issue of low probability and high impact events needs to be stressed to relevant impact zone residents.

The first and hardest step in an effective community evacuation remains convincing people in the projected impact zone that the looming threat is real. Renn and Rohrmann [74] suggest that to help make that threat real, project the expected number of fatalities or the catastrophic potential or where the threat may come from.

### Spreading the Warning

Handmer [43] recommends that the professional warning agencies should attempt to harness the "personal informational networks of individuals within formal communication systems, and by assuring that formal warning advice

is consistent with local norms and behavior” ([43] p 27). Shifting the normative values of recalcitrant disaster managers, and residents in ignorance or denial in urban risk zones becomes the key task of the *seven steps to community safety*, along with the amalgamating approach of the *community safety triangle*.

Part of the strength of the CST is that it taps into and informs an existing social predisposition for people to talk to each other, particularly in times of common threat (King and Goudie 2006). Using emergent technologies to provide real-time information to community media and households helps realize aware informed communities, predisposed to action.

With literature and research results discussed throughout this article, this discussion underlines the importance of modeling local norms, and interactive, safety-focused behavior. Emerging concepts which may be tested in further research are Community Safety groups (Sect. “**Experiences and Lessons – Some Case Studies**”) and Social Burnoffs or flood evacuation practices. The main thrust of this work is to encourage interested modelers to accurately simulate how communities understand and act on the need for safety-oriented action, where people are inclusive at the neighborhood level, acting as a self-preserving group, no matter what the ethnicity or state of surrounding populations. Neighborhood cohesion and empowerment is current social policy consistent with ESD philosophies and principles, enshrined in current planning laws such as IPA 1997 [46].

To illustrate the power of the internet, consider that any online reader globally can click on [103] and see the social aspects of Queensland’s sustainability planning law. Indeed, click on and draw from any of the web sites given. If we were not suffering from information overload and time paucity, the paradigm shift to sustainability implementation would be rapid, the modeling acted on by disaster managers and financial cost-benefit analysts. Until that time, models will help convince and guide people in testable and demonstrable ways to increase community safety. Our very busyness is perhaps the main barrier to reducing global warming and natural disaster impacts. An alternative way into a more sustainable future is centered on local needs-meeting, including nurturing more local community cohesion and sharing of reliable, safety-oriented information and safety-oriented actions.

### **Future Directions in Modeling Disaster Preparedness and Evacuations as a Social Issue**

As global warming and climate change intensifies, the need to model for preparedness and evacuations will in-

crease with more frequent and extreme weather events. The total and fine detail of all needed background preparations and ramp-up preparations are too numerous and arduous for formal disaster management organizations to implement alone, hence the increased need to promote and strongly support the role of community engagement, of community empowerment and nurturing self-help in maximizing effective natural disaster preparation, including evacuations.

Australia has matured disaster policy, law and evolving practice, all embedded in concepts of ecologically sustainable development. Researchers can access, model and test local applicability of some of the Australian experiences and culture of community self-help. Sustainability Implementation Research will become universal as the era of last-minute organizational flurries whilst goading an ignorant and passive population at threat is superseded by prepared and bonded communities who are primed to receive well-delivered warnings, often sourced from the web, and who move themselves to safety in plenty of time. Modeling this will help sell the approach to conservative disaster managers.

The Sustainability Implementation Research, including data-based simulations introduced in this article, is the logical next step to ‘action research’ of the social sciences and will become the norm in approaching all issues of sustainability where the policies are mature, but agencies are unsure how to implement the paradigm shift; the behavioral and technological shift, to agreed long-term goals.

Meaningful consultation is a defining requirement of sustainability planning and good modeling, so all future social research of merit will take the Human Geographer’s approach and “ask the local residents, involve the local residents.” Community empowerment, and its modeling means local residents are entitled to help mold their own future. With hazard management, the shared SIR goal is to empower hazard-zone communities to be aware of the threat, be basically prepared for any warning, and act, as a community, in a precautionary way to maximize safety, minimize loss, and speed recovery.

As we move forward, plain language(s), clear, widely displayed hazard maps and images of like hazard impacts will help energize hazard-zone residents to the threats and their own needed safety-oriented behavior.

Researchers and Complexity and Systems Scientists have a moral if onerous obligation to challenge any layer of government clearly exercising or imposing barriers to helping people internalize any threats, then be supported in getting safe and staying safe, recognizing that change may threaten some individuals or sections within bureaucracies. As disaster management moves from a militaristic

model of command and control to community empowerment and self help, the developing potential of the web as a key information source into hazard zones – ultimately web-to-air – will bear the fruit of greater community safety and minimized loss. Web-to-TV in disaster warning will be a fine conduit for showing residents-at-risk where the threat may be in relation to their home in a few hours. That immediacy of moving images is a compelling motivator for safety-oriented action.

In future, researchers, modelers and others involved in maximizing community safety will embrace some variations of the *communications safety triangle* and the *seven steps to community safety*, simply because they make sense; are highly cost-effective and easily web-refined from international to local conditions, populations and threats.

## Acknowledgment

I most thank my research guardian and mentor over 15 years, Prof. David King, Director of both the Australian Center for Disaster Studies and of the Center for Tropical Urban and Regional Planning. David allowed me freedom to develop as an ‘evidence-based’ scientist, to conceive core approaches to sustainability implementation research; productive in helping render positive change in both disaster management and sustainability planning.

Thanks to Australian Bureau of Meteorology staff for their interactive support to improve risk communications, listening to what real people in real hazard zones experience, how they hear warnings, and how the medium and the messages can be and are optimized. The Bureau embraces the core goal to motivate safety-oriented action by people in hazard zones. The Bureau listens and improves the message and the delivery.

The bushfire research of ’06 & 07 was funded by the Australian Bushfire Cooperative Research Center, supported by the University of Tasmania.

The 14 years of research reported in this article was not possible without the contributions of Authorities and more than 1000 Australians, old and new, who opened their organizations or doors to myself or research team members and shared their hazard experiences and specific warning needs. Thanks all.

## Bibliography

### Primary Literature

1. ACF (2004) Institutions for Sustainability. Australian Conservation Foundation ACF 1–37. [http://www.acfonline.org.au/uploads/res/res\\_tp007.pdf](http://www.acfonline.org.au/uploads/res/res_tp007.pdf). Accessed 2008
2. AMCORD (1995) AMCORD95, Australian Model Code of Residential Development. Department of Housing and Regional Development. Aust Govt Printing Service, Canberra
3. Australian Bushfire CRC (2005) Bushfire CRC/Research/Community Self Sufficiency for Fire Safety/Effective Risk Communication. <http://www.bushfirecrc.com/research/c41/c41.html>. Accessed 2008
4. Baram M (1991) Rights and duties concerning the availability of environmental risk information to the public. In: Kasperson RE, Stallen PJM (eds) Communicating Risks to the Public – International Perspectives. Kluwer, Dordrecht/Boston/London, p 481
5. Berry L, King D (1998) Tropical cyclone awareness and education issues for far north Queensland school students – Storm Watchers. Aust J Emerg Manag 13:6
6. Boughton GN (1992) Education on Natural Hazards, The Macedon Digest. Aust J Disaster Manag 7(2):4–7
7. Boura J (1998) Community Fireguard: Creating partnerships with the community to minimise the impact of bushfire. Aust J Emerg Manag 13:59–64
8. Brundtland GH (1988) Our Common Future. The World Commission for the Environment and Development. Alianza Publications. [http://tilastokeskus.fi/abo2004/foredrag/hoglund\\_pp.pdf](http://tilastokeskus.fi/abo2004/foredrag/hoglund_pp.pdf), <http://www.erf.es/eng/empresa/brundtland.html>, <http://web.uvic.ca/~stucraw/Lethbridge/MyArticles/Brundtland.htm>. Accessed 2008
9. Cairns City Council (2007) Storm surge maps. <http://www.cairns.qld.gov.au/cairns/files/StormTideMaps/index.pdf>. Accessed 2008
10. Chamberlain ER, Hartshorn AE, Muggleston H, Short P, Svensson H, Western JS (1981a) Queensland flood report Australia Day (1974). Australian Government Publishing Service, Canberra, p 38
11. Chamberlain ER, Doube L, Milne G, Rolls M, Western JS (1981b) The Experience of cyclone Tracy. Australian Government Publishing Service, Canberra, p 150
12. Cialdini R, Reno R, Kallgren C (1990) A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. J Pers Soc Psychol 58(6):1015–1026
13. COAG (2004) Natural Disasters in Australia, reforming mitigation, relief and recovery arrangements. [http://www.ema.gov.au/agd/EMA/rwpattach.nsf/VAP/\(756EDFD270AD704EF00C15CF396D6111\)~COAG+Report+on+Natural+Disasters+in+Australia+--+August+2002.pdf/\\$file/COAG+Report+on+Natural+Disasters+in+Australia+--+August+2002.pdf](http://www.ema.gov.au/agd/EMA/rwpattach.nsf/VAP/(756EDFD270AD704EF00C15CF396D6111)~COAG+Report+on+Natural+Disasters+in+Australia+--+August+2002.pdf/$file/COAG+Report+on+Natural+Disasters+in+Australia+--+August+2002.pdf). Accessed 2008, Council of Australian Governments Commonwealth of Australia, 201
14. Cohen E, White P, Hughes P (2006) Bushfire and the Media Reports 1–3. Latrobe University and BCRC, La Trobe University
15. Cronan K (1998) Foundations of emergency management. The Aust J Emerg Manag 1(13):20–23
16. Cuthill M (2004) Community well-being – the ultimate goal of democratic governance. Qld Plan 44(2):8–11
17. Cutter SL, Boruff BJ, Shirley WL (2003) Social Vulnerability to Environmental Hazards. Soc Sci Q 84(2):242–261
18. DIMIA (2004) Department of Immigration, Multicultural and Indigenous Affairs. [http://www.humanrights.gov.au/racial\\_discrimination/face\\_facts/mig.htm#q](http://www.humanrights.gov.au/racial_discrimination/face_facts/mig.htm#q). 2005
19. Dolphin RR, Richard R, Ying F (2000) Is Corporate Communications a Strategic Function. Manag Decis 38(2):99–106



20. Dominick JR (1993) *The dynamics of mass communication*. Mc Graw-Hill Inc, Columbus, p 616
21. Douglas M (1992) *Risk and Blame. Essays in cultural theory*. Routledge, London/New York
22. Drabek TE (1994) *Disaster evacuation and the tourist industry. Program on environment and behaviour monograph 57*. University of Colorado, Colorado
23. Dunlap RE, Van Liere KD (1978) The 'new environmental paradigm': a proposed measuring instrument and preliminary results. *J Environ Ed* 9(4):10–19
24. ECAP (1997) *Guidelines for disaster prevention and preparedness in tropical cyclone areas. Economic Commission for Asia and the Pacific, the World Meteorological Organisation and the Red Cross Societies, Geneva/Bangkok*
25. EMA (2000) *Emergency Risk Management Applications Guide. The Australian Emergency Manuals Series. Emergency Management Australia. Dickson ACT:Emergency Management Australia, p 4*
26. EMA (2002) *Research agenda for Emergency Management. March. EMA Research and development Strategy for "Safer Sustainable Communities". Dickson ACT:Emergency Management Australia*
27. EMA (2002) *Indigenous Communities and Emergency Management. Emergency Management Australia, Canberra, p 22*
28. EMA (2002) *Guidelines for Emergency Managers working with Culturally and Linguistically Diverse Communities. [http://www.ema.gov.au/ema/rwpattach.nsf/viewasattachmentpersonal/AFD7467016783EA8CA256CB30036EF42/\\$file/caldsept2002.pdf](http://www.ema.gov.au/ema/rwpattach.nsf/viewasattachmentpersonal/AFD7467016783EA8CA256CB30036EF42/$file/caldsept2002.pdf). Accessed 2008*
29. EMAI (1998) *Emergency Management Australia Information Service. Report of the strategic planning conference on the development of enhanced awareness education programs and materials for remote Aboriginal and Torres Strait Islander communities. Darwin, May 1997. Conference Proceedings, EMA, Mt Macedon Victoria*
30. Fien J (1993) *Education for the environment – critical curriculum theorising and environmental education. Deakin University, Geelong, Victoria Australia*
31. Finnis K, Johnston D, Paton D (2004) *Volcanic Hazard Risk Perceptions in New Zealand. Tephra, Earth and atmospheric sciences University of AlbertaEdmonton, Alberta, Canada, pp 60–64. [http://www.civildefence.govt.nz/memwebsite.nsf/Files/Tephra%20v21%20chapters/\\$file/volcanichazardrisk.pdf](http://www.civildefence.govt.nz/memwebsite.nsf/Files/Tephra%20v21%20chapters/$file/volcanichazardrisk.pdf), 2008*
32. Fowler KL, Kling ND, Larson MD (2007) *Eigenvalues. Bus Soc* 46:1. March. 88–103. Sage Publications
33. Goldammer J (2005) *Wildland fire – rising threats and vulnerabilities at the interface with society. In: Know risk. United Nations Tudor Rose, T. Jeggle, United Nations, ed; UN International Strategy for Disaster Reduction (UN-ISDR), Geneva, 376 p, pp 322–3*
34. Goudie D, King D (1999) *Cyclone surge and community preparedness. Austn J Emerg Manag* 13:1:454–60. [http://www.ema.gov.au/agd/EMA/rwpattach.nsf/viewasattachmentpersonal/\(85FE07930A2BB4482E194CD03685A8EB\)~Cyclone\\_surge\\_and\\_community\\_preparedness.pdf/\\$file/Cyclone\\_surge\\_and\\_community\\_preparedness.pdf](http://www.ema.gov.au/agd/EMA/rwpattach.nsf/viewasattachmentpersonal/(85FE07930A2BB4482E194CD03685A8EB)~Cyclone_surge_and_community_preparedness.pdf/$file/Cyclone_surge_and_community_preparedness.pdf), 2008
35. Goudie D (2001) *Toward Sustainable Urban Travel. Ph D thesis. James Cook University. <http://eprints.jcu.edu.au/967/>, 2008*
36. Goudie D (2004) *Disruptive weather warnings and weather knowledge in remote Australian Indigenous communities. Web-based report [http://www.tesag.jcu.edu.au/CDS/Pages/reports/Gou\\_iwwrpt/index.shtml?Id=23](http://www.tesag.jcu.edu.au/CDS/Pages/reports/Gou_iwwrpt/index.shtml?Id=23), 2008*
37. Goudie D (2005) *Sustainability planning: pushing against institutional barriers. Ecosystems and Sustainable Development V. WIT Press. WIT Transactions on Ecology and the Environment* 81(5):215–224, [www.witpress.com](http://www.witpress.com), 2008
38. Goudie D (2007) *Transport and Evacuation Planning. In: King D, Cottrell A (eds) Communities Living With Hazards. Centre for Disaster Studies, James Cook University with Queensland Department of Emergency Services, 293, James Cook University, North Queensland Australia, pp 48–62*
39. Goudie D (2007) *Oral histories about weather hazards in northern Australia. In: King D, Cottrell A (eds) Communities Living With Hazards. Centre for Disaster Studies, James Cook University with Queensland Department of Emergency Services, 293, James Cook University, North Queensland Australia, pp 102–125*
40. Gurmankin AD, Baron J, Armstrong K (2004) *Intended message versus received message in hypothetical physician risk communication: exploring the gap. Risk Analysis* 24(5):1337–1347
41. Handmer J (1992) *Can we have too much warning time? A study of Rockhampton, Australia. The Macedon Digest. Aust J Disaster Manag* 7:2 p 8–10
42. Handmer J (2000) *Are Flood Warnings Futile? Risk Commun emergencies* 2(e):1–14. <http://www.massey.ac.nz/~trauma/issues/2000-2/handmer.htm>, 2008
43. Handmer J (2001) *Improving flood warnings in Europe: a research and policy agenda. Environ Hazards* 3(2001):19–28
44. Heller K, Alexander DB, Gatz M, Knight BG, Rose T (2005) *Social And Personal Factors As Predictors Of Earthquake Preparation: The Role Of Support Provision, Network Discussion, Negative Affect, Age, and Education. J Appl Soc Psychol* 35(2):399–422
45. Hoeting JA, Madigan D, Raftery AE, Volinsky CT (1999) *Bayesian Model Averaging: A Tutorial. Stat Sci* 14(4):382–417
46. IPA (1997) *The Integrated Planning Act. Queensland Government, <http://www.legislation.qld.gov.au/LEGISLTN/CURRENT/I/integplana97.pdf>, 2008*
47. ISR (2007) *Australian Policyonline. Institute for Social Research, Swinburne University of Technology, [http://www.apo.org.au/linkboard/results.shtml?filename\\_num=117732](http://www.apo.org.au/linkboard/results.shtml?filename_num=117732), 2008*
48. Jarach M (1989) *Overview of the literature on barriers to the diffusion of renewable energy sources in agriculture. Appl En* 32(2):117–131
49. Kasperson RE, Stallen PJM (1991) *Communicating Risks to the Public International Perspectives. Kluwer, Dordrecht/Boston/London*
50. Kim YC, Ball-Rokeach SJ (2006) *Civic engagement from a communication infrastructure perspective. Commun Theory* 16:173–197
51. Kitchin RM (1996) *Increasing the integrity of cognitive mapping research: appraising conceptual schemata of environment-behaviour interaction. Progress Hum Geogr* 20(1):56–84

52. King D, Goudie D (1998) Breaking through the disbelief – the March 1997 floods at Cloncurry. Even the duck swam away. *Aust J Emerg Manag* 4:12 29–33
53. King D, Goudie D (2006) Cyclone Larry, March 2006 Post Disaster Residents Survey. Centre for Disaster Studies, James Cook University, with the Australian Bureau of Meteorology P77 [http://www.tesag.jcu.edu.au/CDS/Pages/reports/Larry\\_mainreport.pdf](http://www.tesag.jcu.edu.au/CDS/Pages/reports/Larry_mainreport.pdf), 2008
54. King D, Goudie D, Dominey-Howes D (2006) Cyclone knowledge and household preparation – some insights from Cyclone Larry Report on how well Innisfail prepared for Cyclone Larry. *The Aust J Emerg Manag* 21:3 52–59 [http://www.ema.gov.au/agd/EMA/rwpattach.nsf/VAP/\(A80860EC13A61F5BA8C1121176F6CC3C\)~AJEM\\_EMA\\_Larry\\_Aug2006.pdf/\\$file/AJEM\\_EMA\\_Larry\\_Aug2006.pdf](http://www.ema.gov.au/agd/EMA/rwpattach.nsf/VAP/(A80860EC13A61F5BA8C1121176F6CC3C)~AJEM_EMA_Larry_Aug2006.pdf/$file/AJEM_EMA_Larry_Aug2006.pdf). Accessed August 2006
55. Kobb P (2000) Emergency Risk Management Applications Guide. *Emerg Manag Aust*, Dickson, A.C.T.: Emergency Management Australia. Australian Emergency Manuals Series; 05 [http://www.ema.gov.au/agd/EMA/rwpattach.nsf/VAP/\(383B7EDC29CDE21FBA276BBCE12CDC0\)~Manual+05a.pdf/\\$file/Manual+05a.pdf](http://www.ema.gov.au/agd/EMA/rwpattach.nsf/VAP/(383B7EDC29CDE21FBA276BBCE12CDC0)~Manual+05a.pdf/$file/Manual+05a.pdf), 2008
56. Leibovitz J (2003) Institutional barriers to associative city-region governance: the politics of institution-building and economic governance in 'Canada's Technology Triangle.' *Urban Studies*: 40(13):2613–2642
57. Lewis C (2006) Risk Management and Prevention Strategies. *Aust J Emerg Manag* 21(3):47–51
58. Lichtenberg J, Maclean D (1991) The role of the media in risk communication. In: Kasperson RE, Stallen PJM (ed) *Communicating Risks to the Public – International Perspectives*. Kluwer, Dordrecht/Boston/London, p 481
59. Lidstone J (2006) Blazer to the Rescue! The role of puppetry in enhancing fire prevention and preparedness for young children. *Aust J Emerg Manag* 21(2):17–28
60. Loudness RS (1977) Tropical Cyclones in the Australian Region July 1909 to June 1975. AGPS, Canberra
61. McKenna F (1993) It won't happen to me: Unrealistic optimism or illusion of control. *Brit J Psych* 84:39–50
62. Munro DA (1995) Ecologically sustainable development – is it possible? How will we recognise it? In: Sivakumar M, Messer J (eds) *Protecting the future – ESD in action*. Futureworld, Wollongong
63. NDSI Working Group (2000) Effective Disaster Warnings. Working Group on Natural Disaster Information Systems. Subcommittee on Natural Disaster Reduction National Science and Technology Council Committee on Environment and Natural Resources. Executive Office of the President of the United States of America, 56, [http://www.incident.com/cap/docs/NDIS\\_rev\\_Oct27.pdf](http://www.incident.com/cap/docs/NDIS_rev_Oct27.pdf), 2008
64. Napurrurlarlu NO, Jakamarlarlu NP (1988) Ngawarra-Kurlu. Yuendumu B.R.D.U., Darwin, p 19
65. O'Neill P (2004) Why don't they listen – Developing a risk communication model to promote community safety behaviour. The International Emergency Management Society, 11th Annual Conference Proceedings, Melbourne, Victoria, Australia, May 18–21 2004, pp 160–169
66. Young J, O'Neill P (1999) A social marketing framework for the development of effective public awareness programs. [http://www.ses.nsw.gov.au/multiattachments/2740/documentname/A\\_Social\\_Marketing\\_Framework\\_for\\_the\\_Development\\_of\\_Effective\\_Public\\_Awareness\\_Programs.pdf](http://www.ses.nsw.gov.au/multiattachments/2740/documentname/A_Social_Marketing_Framework_for_the_Development_of_Effective_Public_Awareness_Programs.pdf), 2008. NSW Australia
67. Paton D (2003) Stress in Disaster Response: A risk management approach. *Disaster Prev Manag* 12(3):203–209
68. Paton D, Smith L, Johnston D (2005) When good intentions turn bad: promoting natural Hazard preparedness. *Aust J Emerg Manag* 20:25–30
69. Phillips R (1994) Long Range Planning. *Lond* 27(4):143–145
70. QG & QES (2003) State Planning Policy. Mitigating the adverse impacts of flood, bushfire and landslide. State planning policy 1/03. Dept Local Government and planning, & Dept of Emergency Services, <http://www.emergency.qld.gov.au/publications/spp/pdf/spp.pdf>, 2008
71. Quarantelli EL (2002) The role of the Mass Communication system in natural and technological disasters and possible extrapolation to terrorism situations. *Risk Manag Int J* 4(4):7–21
72. Raggatt P, Butterworth E, Morrissey S (1993) Issues in Natural Disaster Management: Community Response to the Threat of Tropical Cyclones in Australia. *Disaster Prev Manag* 2(3):12–21
73. Renn O, Levine D (1991) Credibility and trust in risk communication. In: Kasperson RE, Stallen PJM (ed) *Communicating Risks to the Public – International Perspectives*. Kluwer, Dordrecht/Boston/London, p 481
74. Renn O, Rohrman B (2000) Cross-cultural risk perception, A survey of empirical studies. Kluwer, Dordrecht, p 240
75. Rohrman B (2000) A socio-psychological model for analysing risk communication processes. *Australas J Disaster Trauma Stud* 2000(2) <http://www.massey.ac.nz/~trauma/issues/2000-2/rohrmann.htm>, 2008
76. Rounsefell V (1992) Unified human settlement ecology. In: Birkeland J (ed) *Design for sustainability: A sourcebook of integrated eco-logical solutions*. Earthscan Publications, London, 54.2:78–83
77. Salter J (1992) The Nature of the disaster – more than just the meanings of words: Some reflections on definitions, doctrine and concepts. *The Macedon Digest*. *Aust J Disaster Manag* 7(2):1–3
78. Salter J, Bally J, Elliott J, Packham D (1993) Natural disasters: protecting vulnerable communities. In: Merriman PA, Browitt CW (eds) *Conference Proceedings*. London, 13–15 October 1993, Royal Society (Great Britain)
79. Sheppard E (1986) Modelling and predicting aggregate flows. In: Hanson S (ed) *The geography of urban transportation*. Guilford Press, New York/London, pp 91–110
80. Skertchly A, Skertchly K (2000) Message Sticks – Hazard Mitigation Visual Language. EMA, ACT, Dickson, Australian Capital Territory
81. Sorenson J, Mileti D (1991) Risk communication in emergencies. In: Kasperson RE, Stallen PJM (ed) *Communicating Risks to the Public – International Perspectives*. Kluwer, Dordrecht/Boston/London, p 481
82. Stern P (1992) Psychological dimensions of global environmental change. *Ann Rev Psychol* 43:269–302
83. Stern PC, Fineberg HV (1996) *Understanding Risk, Informing decisions in a democratic society*. National Academy Press, Washington DC, pp 249
84. Sullivan M (2003) Communities and their experience of emergencies. *Aust J Emerg Manag* 18(1):19–26

85. Svenson O (1991) The time dimension in perception and communication of risk. In: Kasperson RE, Stallen PJM (ed) *Communicating Risks to the Public – International Perspectives*. Kluwer, Dordrecht/Boston/London, p 481
86. Thompson KM (2002) Variability and uncertainty meet risk management and risk communication. *Risk Analysis* 22(3):647–654
87. Utemorrhah D, Clendon M (2000) Dumbi the owl. In: Kimberley Language Resource Centre (ed) *Worrorra Lalai, Worrorra Dreamtime Stories*. KLRC, Halls Creek WA, pp 113
88. Utemorrhah D (1980) How The People Were All Drown(ed). In: Mowanjum (ed) *Visions of Mowanjum: Aboriginal writings from the Kimberley*. Rigby, Adelaide
89. Wakefield SE, Elliot SJ (2003) Constructing the news: the role of local newspapers in environmental risk communication. *Prof Geogr* 55(2):216–266
90. Wall M (2006) The case study method and management learning: making the most of a strong story telling tradition in emergency services management education. *Aust J Emerg Manag* 21(2):11–16
91. Walmsley DJ (1988) *Urban living, the individual in the city*. Longman scientific and technical, Longman, London, p 104
92. Woods F, Gabriel P (2005) Individual responsibility and state-wide strategies: bushfire in Victoria, Australia. In: Know risk. United Nations Tudor Rose, UK, Jeggle T (ed) Leicester, LE1 5RA, UK 376. pp 326–8
93. Yates J (1992) Assisting the community to plan: A pilot program in Western Australia. *The Macedon Digest*. *Aust J Disaster Manag* 7(2):12
94. Yates J (1997) Federalism and disaster mitigation in remote Aboriginal communities in Western Australia. Spring, AJEM, 25–32, *Emergency Management Australia*, Mt Macedon Victoria Australia. Publisher: Grey Worldwide Canberra Australia
95. Yeo S (2002) Natural Hazards. Flooding in Australia: A review of events in 1998. 25:177–191. Department of Physical Geography, Macquarie University, NSW, <http://www.springerlink.com/content/n160g0121800n742/> Natural Hazards 25(2):177–191, 2002. Kluwer Academic Publishers. Printed in the Netherlands. <http://www.springerlink.com/content/n160g0121800n742/fulltext.pdf>
96. Zamecka A, Buchanan G (2000) *Disaster risk management*. Queensland Department of Emergency Services, Brisbane, p 115
97. Kikuchi T, Nakamori Y (2007) Agent model analysis to explore effects of interaction and environment on individual performance. *J Syst Sci Complexity* 2007 20:1–17. Springer Science + Business Media, Inc
98. <http://news.bbc.co.uk/1/hi/world/europe/6963373.stm> BBC (2005) Access to 2008, In pictures: Hurricane onslaught
99. <http://news.bbc.co.uk/2/hi/americas/7055721.stm> BBC (2007) Californians flee as fires rage
100. Mornington Peninsula Shire Council (2006) Fire wise fire management. <http://www.mornpen.vic.gov.au/Files/FireWiseFireManagementBooklet.pdf>
101. Cairns City Council (2006) Storm tide maps. <http://www.hazardsaustralia.info/Mapping.html>
102. <http://maps.redland.qld.gov.au/website/redemapexternal%5Fv2%5F03/Default.aspx>, Redlands Shire Accessed 2008 Flood Maps
103. Queensland Government (1997) Integrated planning act. <http://www.legislation.qld.gov.au/LEGISLTN/CURRENT/I/integplana97.pdf>

### Books and Reviews

- Bushnell S, Cottrell A, Spillman M, Lowe D (2006) Thuringowa bushfire case study. Understanding Communities. Project BCRC Program C Community self-sufficiency for fire safety. Bushfire CRC, JCU CDS, Townsville, Australia
- Granger KJ, Smith DI (1995) Storm tide impact and consequence modelling: some preliminary observations. *Math Comput Model* 21:9:15–21
- Roth W (1897) *Ethnological Studies among the North West Central Queensland Aborigines*. Queensland government, Brisbane/London
- FEMA (2008) Prepare for a Wildfire. Federal Emergency Management Agency. U.S. Department of Homeland Security Washington, DC [http://www.fema.gov/hazard/wildfire/wf\\_prepare.shtm](http://www.fema.gov/hazard/wildfire/wf_prepare.shtm)
- Australian Bureau of Meteorology (2008) Protecting yourself and your home. Bushfire weather. BoM. Melbourne Australia [http://www.bom.gov.au/inside/services\\_policy/fire\\_ag/bushfire/protect.htm](http://www.bom.gov.au/inside/services_policy/fire_ag/bushfire/protect.htm)
- Emergency Management Australia (2003). Community safety Bushfire action guide. EMA. Dickson, Australian Capital Territory <http://www.ema.gov.au/agd/ema/emainternet.nsf/Page/RWP07C6046B98D07DB8CA256C5A00230553>
- ABC (2006) Bushfire summer. Australian Broadcasting Commission. Melbourne, [www.abc.net.au/bushfire](http://www.abc.net.au/bushfire)
- Geoscience Australia (2005) Sentinel. Commonwealth of Australia. Canberra, ACT, Australia, <http://sentinel.ga.gov.au/acres/sentinel/index.shtml>
- Rural Fire Service (2008) Fire Safety Information. New South Wales Rural Fire Service. NSW Government, Sydney, [http://www.rfs.nsw.gov.au/dsp\\_content.cfm?CAT\\_ID=515](http://www.rfs.nsw.gov.au/dsp_content.cfm?CAT_ID=515) NSW Rural Fire Service, with information in 27 languages
- CFA (2008) Country Fire Authority. Melbourne, Victoria, Australia, <http://www.cfa.vic.gov.au>
- ESA (2008) Community Education. Emergency Services Agency, Australian Capital Territory, [http://www.esa.act.gov.au/esawebsite/content\\_esa/community\\_education/community\\_education.html](http://www.esa.act.gov.au/esawebsite/content_esa/community_education/community_education.html)

## Evacuation Dynamics: Empirical Results, Modeling and Applications

ANDREAS SCHADSCHNEIDER<sup>1,2</sup>, WOLFRAM KLINGSCH<sup>3</sup>,  
HUBERT KLÜPFEL<sup>4</sup>, TOBIAS KRETZ<sup>5</sup>,

CHRISTIAN ROGSCHE<sup>3</sup>, ARMIN SEYFRIED<sup>6</sup>

<sup>1</sup> Institut für Theoretische Physik, Universität zu Köln,  
Köln, Germany

<sup>2</sup> Interdisziplinäres Zentrum für Komplexe Systeme,  
Bonn, Germany

<sup>3</sup> Institute for Building Material Technology  
and Fire Safety Science, University of Wuppertal,  
Wuppertal, Germany

<sup>4</sup> TraffGo HT GmbH, Duisburg, Germany

<sup>5</sup> PTV Planung Transport Verkehr AG, Karlsruhe,  
Germany

<sup>6</sup> Jülich Supercomputing Centre, Research Centre Jülich,  
Jülich, Germany

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Empirical Results](#)

[Modeling](#)

[Applications](#)

[Future Directions](#)

[Acknowledgments](#)

[Bibliography](#)

### Glossary

**Pedestrian** A person traveling on foot. In this article, other characterizations are used depending on the context, e. g., agent or particle.

**Crowd** A large group of pedestrians moving in the same area, but not necessarily in the same direction.

**Evacuation** The movement of persons from a dangerous place due to the threat or occurrence of a disastrous event. In normal situations this is called “egress” instead.

**Flow** The flow or current  $J$  is defined as the number of persons passing a specified cross-section per unit time. The common unit of flow is “persons per second”. Specific flow is the flow per unit cross-section. The maximal flow supported by a facility (or a part of it) is called “capacity”.

**Fundamental diagram** In traffic engineering (and physics): density-dependence of the flow:  $J(\rho)$ . Due to the hydrodynamic relation  $J = \rho v b$  equivalent rep-

resentations used frequently are  $v = v(\rho)$  or  $v = v(J)$ . The fundamental diagram is probably the most important quantitative characterization of traffic systems.

**Lane formation** In bidirectional flows, lanes are often dynamically formed in which all pedestrians move in the same direction.

**Bottleneck** A limited resource for pedestrian flows, for example a door, a narrowing in a corridor, or stairs, i. e., a location of reduced capacity. At bottlenecks jamming occurs if the inflow is larger than the capacity. Other phenomena that can be observed are the formation of lanes and the zipper-effect.

**Microscopic models** Models which represent each pedestrian separately with individual properties like walking velocity or route choice behavior and the interactions between them. Typical models that belong to this class are cellular automata and the social-force model.

**Macroscopic models** Models which do not distinguish individuals. The description is based on aggregate quantities, e. g., appropriate densities. Typical models belonging to this class are fluid-dynamic approaches. Hand calculation methods which are based on related ideas and are often used in the field of (fire-safety) engineering belong to this class as well.

**Crowd disaster** An accident in which the specific behavior of the crowd is a relevant factor, e. g., through competitive and non-adaptive behavior. In the media, it is often called “panic” which is a controversial concept in crowd dynamics and should thus be avoided.

### Definition of the Subject

Today, there are many occasions on which a large number of people gathers in a rather small area. Office buildings and apartment houses grow larger and more complex. Very large events related to sports, entertainment or cultural and religious events are held all over the world on a regular basis. This brings about serious safety issues for the participants and for the organizers who must be prepared for any case of emergency or critical situation. Usually in such cases the participants must be guided away from the dangerous area as quickly as possible. Therefore the understanding of the dynamics of large groups of people is very important.

In general, evacuation is egress from an area, a building or a vessel due to a potential or actual threat. In the cases described above, the dynamics of the evacuation processes are quite complex due to the large number of people and their interaction, external factors such as fire, complex building geometries, etc. Evacuation dynamics must be described and understood on different levels: physical,

physiological, psychological, and social. Accordingly, the scientific investigation of evacuation dynamics involves many research areas and disciplines. The system “evacuation process” (i. e., the population and the environment) can be modeled on many different levels of detail, ranging from hydro-dynamic models to artificial intelligence and multi-agent systems. There are at least three aspects of evacuation dynamics that motivate its scientific investigation:

- 1) As in most many-particle systems several interesting collective phenomena can be observed that need to be explained;
- 2) Models need to be developed that are able to reproduce pedestrian dynamics in a realistic way, and
- 3) Pedestrian dynamics must be applied to facility design and to emergency preparation and management.

The investigation of evacuation dynamics is a difficult problem that requires close collaboration between different fields. The origin of the apparent complexity lies in the fact that one is concerned with a many-‘particle’ system with complex interactions that are not fully understood. Typically the systems are far from equilibrium and so are, e. g., sensitive to boundary conditions. Motion and behavior are influenced by several external factors and often crowds can be rather inhomogeneous.

In this article we want to deal with these problems from different perspectives and will not only review the theoretical background, but will also discuss some concrete applications.

## Introduction

The awareness that emergency exits are one of the most important factors to ensure the safety of persons in buildings can be traced back more than 100 years. Disasters due to the fires in the Ring theater in Vienna and the urban theater in Nizza in 1881 resulted in several hundred fatalities and led to a rethinking of the safety in buildings [24]. First, attempts were made to improve safety by using non-flammable building materials. However, the disaster at the Troquois Theater in Chicago with more than 500 fatalities, where only the decorations burned, demonstrated the need for more effective measures. This was a starting point for studying the influences of emergency exits and thus the dynamics of pedestrian streams [24,32].

In recent years there have been two major evacuation incidents which gained immense global attention. First, there was the capsizing of the Baltic Sea ferry MV Estonia (September 28, 1994, 852 casualties) [100] and, of course, the terrorist attacks of 9/11 (2,749 casualties).

Other prominent examples of the possible tragic outcomes of the dynamics of pedestrian crowds are the Hillsborough stadium disaster in Sheffield (April 15, 1989, 96 casualties) [182], the accident at Bergisel (December 4, 1999, 5 casualties) [189], the stampede in Baghdad (August 30, 2005, 1,011 casualties), the tragedy at the concert of “The Who” (December 3, 1979, 11 casualties) [73] and – very early – the events at the crowning ceremony of Tsar Nicholas II in St. Petersburg in May 1896 with 1,300 to 3,000 fatalities (sources vary considerably) [168]. In the past, tragic accidents have happened frequently in Mecca during the Hajj (1990: 1,426, 1994: 270, 1997: 343, 1998: 107, 2001: 35, 2003: 14, 2004: 244, and 2006: 364 casualties). What stands out is that the initiating events are very diverse and range from external human aggression (terrorism) to external physical dangers (fire) and rumors to various shades of greedy behavior in absence of any external danger.

Many authors have pointed out that the results of experts’ investigations and the way the media typically reports about an accident very often differ strongly [17,77,109,155,156,178]. Public discussion has a much greater tendency to identify “panic” as the cause of a disaster, while expert commissions often conclude that there either was no panic at all, or panic was merely a result of some other preceding phenomenon.

This article first discusses the empirical basis of pedestrian dynamics in Sect. “Empirical Results”. Here we introduce the basic observables and describe the main qualitative and quantitative results, focusing on collective phenomena and the fundamental diagram. It is emphasized that even for the most basic quantities, no consensus about basic behavior has been reached.

In Sect. “Modeling” various model approaches that have been applied to the description of pedestrian dynamics are reviewed.

Section “Applications” discusses more practical issues and gives a few examples for applications to safety analysis. In this regard, prediction of evacuation times is an important problem as legal regulations must often be fulfilled. Here, commercial software tools are available. A comparison shows that the results must be interpreted with care.

## Empirical Results

### Overview

Pedestrians are three-dimensional objects and a complete description of their highly developed and complicated motion sequence is rather difficult. Therefore, in pedestrian and evacuation dynamics, pedestrian motion is usu-

ally treated as two-dimensional by considering the vertical projection of the body.

In the following sections we review the present knowledge of empirical results. These are relevant not only as a basis for the development of models, but also for applications such as safety studies and legal regulations.

We start with the phenomenological description of collective effects. Some of these are known from everyday experience and will serve as benchmark tests for any kind of modeling approach. Any model that does not reproduce these effects is missing some essential part of the dynamics. Next, the foundations of a quantitative description are laid by introducing the fundamental observables of pedestrian dynamics. Difficulties arise from different conventions and definitions. Then pedestrian dynamics in several simple scenarios (corridor, stairs etc.) are discussed. Surprisingly, even for these simple cases no consensus about the basic quantitative properties exists. Finally, more complex scenarios are discussed which are combinations of the simpler elements. Investigations of scenarios such as evacuations of large buildings or ships suffer even more from lack of reliable quantitative and sometimes even qualitative results.

### Collective Effects

One of the reasons why the investigation of pedestrian dynamics is attractive for physicists is the large variety of interesting collective effects and self-organization phenomena that can be observed. These macroscopic effects reflect the individuals' microscopic interactions and thus give important information for any modeling approach.

**Jamming** Jamming and clogging typically occur for high densities at locations where the inflow exceeds capacity. Locations with reduced capacity are called *bottlenecks*. Typical examples are exits (Fig. 1) or narrowings. This kind of jamming phenomenon does not depend strongly on the microscopic dynamics of the particles. Rather it is a consequence of an exclusion principle: space occupied by one particle is not available for others.

This clogging effect is typical for a bottleneck situation. It is important for practical applications, especially evacuation simulations.

Other types of jamming occur in the case of counterflow where two groups of pedestrians mutually block each other. This happens typically at high densities and when it is not possible to turn around and move back, e. g., when the flow of people is large.

**Density waves** Density waves in pedestrian crowds can be generally characterized as quasi-periodic density varia-



Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 1  
Clogging near a bottleneck. The shape of the clog is discussed in more detail in Subsect. "Theoretical Results"

tions in space and time. A typical example is the movement in a densely crowded corridor (e. g., in subway-stations close to the density that causes a complete halt of motion) where phenomena similar to stop-and-go vehicular traffic can be observed, e. g., density fluctuations in a longitudinal direction that move backwards (opposite to the movement direction of the crowd) through the corridor. More specifically, for the situation on the Jamarat Bridge in Makkah (during the Hajj pilgrimage 2006), stop-and-go waves have been reported. At densities of 7 persons per  $\text{m}^2$  upstream, moving stop-and-go waves of period 45 s have been observed that lasted for 20 minutes [59]. Fruin reports, that "at occupancies of about 7 persons per square meter the crowd becomes almost a fluid mass. Shock waves can be propagated through the mass sufficient to lift people off their feet and propel them distances of 3 m (10 ft) or more." [36].

**Lane formation** In counterflow, i. e., two groups of people moving in opposite directions, (dynamically varying) lanes are formed where people move in just one direction [135,139,197]. In this way, strong in-

interactions with oncoming pedestrians are reduced which is more comfortable and allows higher walking speeds.

The occurrence of lane formation does not require a preference of moving on one side. It also occurs in situations without left- or right-preference. However, cultural differences for the preferred side have been observed. Although this preference is not essential for the phenomenon itself, it has an influence on the kind of lanes formed and their order.

Several quantities for the quantitative characterization of lane formation have been proposed. Yamori [197] has introduced a band index which is basically the ratio of pedestrians in lanes to their total number. In [13] a characterization of lane formation through the (transversal) velocity profiles at fixed positions has been proposed. Lane formation has also been predicted to occur in colloidal mixtures driven by an external field [15,28,158]. Here, an order parameter  $\phi = \frac{1}{N} \langle \sum_{j=1}^N \phi_j \rangle$  has been introduced where  $\phi_j = 1$  if the lateral distance to all other particles of the other type is larger than a typical density-dependent length scale, and  $\phi_j = 0$  otherwise.

The number of lanes can vary considerably with the total width of the flow. Figure 2 shows a street in the city center of Cologne during World Youth Day in Cologne (August 2005) where two comparatively large lanes have been formed.

The number of lanes usually is not constant and might change in time, even if there are relatively small changes in density. The number of lanes in opposite di-

rections is not always identical. This can be interpreted as a sort of spontaneous symmetry breaking.

Quantitative empirical studies of lane formation are rare. Experimental results have been reported in [94] where two groups with varying relative sizes had to pass each other in a corridor with a width of 2 m. On one hand, similar to [197] a variety of different lane patterns were observed, ranging from 2 to 4 lanes. On the other hand, in spite of this complexity, surprisingly large flows could be measured: the sum of (specific) flow and counterflow was between 1.8 and 2.8 persons per meter per second and exceeded the specific flow for one-directional motion ( $\approx 1.4$  P/ms).

**Oscillations** In counterflow at bottlenecks, e. g., doors, one can sometimes observe oscillatory changes of the direction of motion. Once a pedestrian is able to pass the bottleneck it becomes easier for others to follow in the same direction until somebody is able to pass the bottleneck (e. g., through a fluctuation) in the opposite direction.

**Patterns at intersections** At intersections, various collective patterns of motion can be formed. A typical example is short-lived roundabouts which make motion more efficient. Even if these are connected with small detours, the formation of these patterns can be favorable since they allow for “smoother” motion.

**Emergency situations, “panic”** In emergency situations various collective phenomena have been reported that have sometimes misleadingly been attributed to *panic behavior*. However, there is strong evidence that this is not the case. Although a precise accepted definition of *panic* is missing, usually certain aspects are associated with this concept [77]. Typically “panic” is assumed to occur in situations where people compete for scarce or dwindling resources (e. g., safe space or access to an exit) which leads to selfish, asocial or even completely irrational behavior and contagion that affects large groups. A closer investigation of many crowd disasters has revealed that most of the above characteristics have played almost no role and most of the time have not been observed at all (see e. g. [73]). Often the reason for these accidents is much simpler, e. g., in several cases the capacity of the facilities was too small for the actual pedestrian traffic, e. g., Luschniki Stadium Moskau (October 20, 1982), Bergisel (December 4, 1999), pedestrian bridge Kobe (Akashi) (July 21, 2001) [186]. Therefore the term “panic” should be avoided, *crowd disaster* being a more appropriate characterization. Also it should be kept in mind that in dangerous situations it is *not* irrational to fight for resources (or for your own life), if everybody else does



Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 2

The “Hohe Straße” in Cologne during World Youth Day 2005. The yellow line is the border of the two walking directions

this [18,113]. Only from the outside is this behavior perceived as irrational since it might lead to a catastrophe [178]. The latter aspect is therefore better described as *non-adaptive behavior*. We will discuss these issues in more detail in Subsect. “**Evacuations: Empirical Results**”.

## Observables

Before we review experimental studies in this section, the commonly used observables are introduced.

The flow  $J$  of a pedestrian stream gives the number of pedestrians crossing a fixed location of a facility per unit of time. Usually it is taken as a scalar quantity since only the flow normal to some cross-section is considered. There are various methods to measure flow. The most natural approach is to determine the times  $t_i$  at which pedestrians pass a fixed measurement location. The time gaps  $\Delta t_i = t_{i+1} - t_i$  between two consecutive pedestrians  $i$  and  $i + 1$  are directly related to the flow

$$J = \frac{1}{\langle \Delta t_i \rangle} \quad \text{with} \quad \langle \Delta t_i \rangle = \frac{1}{N} \sum_{i=1}^N (t_{i+1} - t_i) = \frac{t_{N+1} - t_1}{N}. \quad (1)$$

Another possibility for measuring the flow of a pedestrian stream is borrowed from fluid dynamics. The flow through a facility of width  $b$  is determined by the average density  $\rho$  and the average speed  $v$  of a pedestrian stream as

$$J = \rho v b = J_s b. \quad (2)$$

where the *specific flow*<sup>1</sup>

$$J_s = \rho v \quad (3)$$

gives the flow per unit-width. This relation is also known as *hydrodynamic relation*.

There are several problems concerning the way in which velocities, densities or time gaps are measured and the conformance of the two definitions of flow. The flow according to Eq. (1) is usually measured as a mean value over time at a certain location, while the measurement of the density in Eq. (2) is connected with an instantaneous mean value over space. This can lead to a bias caused by underestimation of fast moving pedestrians at the average over space compared to the mean value of the flow over time at a single measurement line (see the discussion for vehicular traffic e. g., in [51,81,102]). Furthermore,

<sup>1</sup>In strictly one-dimensional motion often a line density (dimension: 1/length) is used. Then the flow is given by  $J = \rho v$ .

most experimental studies measuring the flow according to Eq. (2) combine for technical reasons an *average* velocity of a single pedestrian over time with an *instantaneous* density. To ensure a correspondence of the mean values the average velocity of all pedestrians contributing to the density at a certain instant has to be considered. However this procedure is very time consuming and not realized in practice up to now. Moreover, the fact that the dimension of the test section has usually the same order of magnitude as the extent of the pedestrians can influence the averages over space. These all are possible factors why different measurements can differ in a large way, see discussion in Subsect. “**Fundamental Diagram**”.

Another way to quantify the pedestrian load of facilities has been proposed by Fruin [35]. The “pedestrian area module” is given by the reciprocal of the density. Thompson and Marchant [184] introduced the so-called “inter-person distance”  $d$ , which is measured between center coordinates of the assessing and obstructing persons. According to the “pedestrian area module” Thompson and Marchant call  $\sqrt{1/\rho}$  the “average inter-person distance” for a pedestrian stream of evenly spaced persons [184]. An alternative definition is introduced in [58] where the local density is obtained by averaging over a circular region of radius  $R$ ,

$$\rho(\mathbf{r}, t) = \sum_j f(\mathbf{r}_j(t) - \mathbf{r}), \quad (4)$$

where  $\mathbf{r}_j(t)$  are the positions of the pedestrians  $j$  encompassed by  $\mathbf{r}$  and  $f(\dots)$  is a Gaussian, distance-dependent weight function.

In contrast to the density definitions above, Predtechenskii and Milinskii [151] consider the ratio of the sum of the projection area  $f_j$  of the bodies and the total area of the pedestrian stream  $A$ , defining the (dimensionless) density  $\tilde{\rho}$  as

$$\tilde{\rho} = \frac{\sum_j f_j}{A}, \quad (5)$$

a quantity known as *occupancy* in the context of vehicular traffic. Since the projection area  $f_j$  depends strongly on the type of person (e. g., it is much smaller for a child than for an adult), the densities for different pedestrian streams consisting of the same number of persons and the same stream area can be quite different.

Beside technical problems due to camera distortions and camera perspective there are several conceptual problems, such as the association of averaged with instantaneous quantities, the need to choose an observation area in the same order of magnitude as the extent of a pedestrian together with the definition of the density of objects



with nonzero extent and much more. A detailed analysis of the ways in which measurement influences the relations is necessary but still lacking.

### Fundamental Diagram

The fundamental diagram describes the empirical relation between density  $\rho$  and flow  $J$ . The name indicates its importance and naturally it has been the subject of many investigations. Due to the hydrodynamic relation (3) there are three equivalent forms:  $J_s(\rho)$ ,  $v(\rho)$  and  $v(J_s)$ . In applications the relation is a basic input for engineering methods developed for the design and dimensioning of pedestrian facilities [35,136,150]. Furthermore, it is a quantitative benchmark for models of pedestrian dynamics [21,86,112,175].

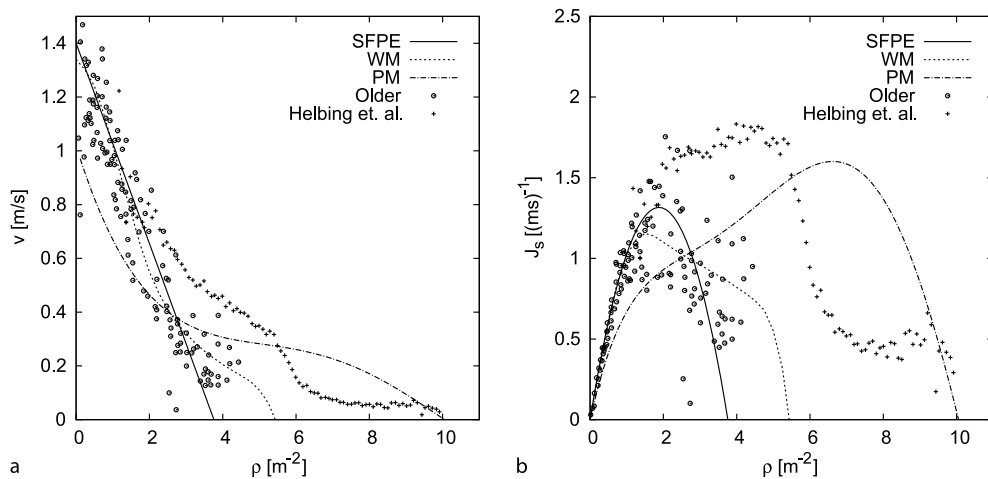
In this section we will concentrate on planar facilities such as sidewalks, corridors and halls. For various facilities such as floors, stairs or ramps, the shape of the diagrams differ, but in general it is assumed that the fundamental diagrams for the same type of facilities but having different widths merge into one diagram for the specific flow  $J_s$ . In first order this is confirmed by measurements on different widths [49,135,139,142]. However, Navin and Wheeler observed in narrow sidewalks more orderly movement leading to slightly higher specific flows than for wider sidewalks [135]. A natural lower bound for the independence of the specific flow from the width is given by the body size and the asymmetry in movement possibilities of the hu-

man body. Surprisingly, Kretz et al. found an increase of the specific flow for bottlenecks with  $b \leq 0.7$  m [93]. This will be discussed in more detail later. For the following discussion we assume facility widths larger than  $b = 0.6$  m and use the most common representations  $J_s(\rho)$  and  $v(\rho)$ .

Figure 3 shows various fundamental diagrams used in planning guidelines and measurements of two selected empirical studies representing the overall range of the data. The comparison reveals that specifications and measurements disagree considerably. In particular, the maximum of the function giving the capacity  $J_{s,\max}$  ranges from  $1.2$  ( $\text{ms}^{-1}$ ) to  $1.8$  ( $\text{ms}^{-1}$ ), the density value where the maximum flow is reached ( $\rho_c$ ) ranges from  $1.75$   $\text{m}^{-2}$  to  $7$   $\text{m}^{-2}$  and, most notably, the density  $\rho_0$ , where the velocity approaches zero due to overcrowding, ranges from  $3.8$   $\text{m}^{-2}$  to  $10$   $\text{m}^{-2}$ .

Several explanations for these deviations have been suggested, including cultural and population differences [58,116], differences between uni- and multidirectional flow [99,135,154], short-ranged fluctuations [154], influence of psychological factors given by the incentive of the movement [150] and, partially related to the latter, the type of traffic (commuters, shoppers) [139].

It seems that the most elaborate fundamental diagram is given by Weidmann who collected 25 data sets. An examination of the data which were included in Weidmann's analysis shows that most measurements with densities larger than  $\rho = 1.8$   $\text{m}^{-2}$  are performed on multidirectional streams [135,139,140,142,148]. But data gained



Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 3

Fundamental diagrams for pedestrian movement in planar facilities. The lines refer to specifications according to planning guidelines (SFPE Handbook [136]), Predtechenskii and Milinskii (PM) [150], Weidmann (WM) [192]). Data points give the range of experimental measurements (Older [142] and Helbing [58])

by measurements on strictly unidirectional streams has also been considered [35,49,188]. Thus Weidmann neglected differences between uni- and multidirectional flow in accordance with Fruin, who states in his often cited book [35] that the fundamental diagrams of multidirectional and unidirectional flow differ only slightly. This disagrees with results of Navin and Wheeler [135] and Lam et al. [99] who found a reduction of the flow in dependence of directional imbalances. Here lane formation in bidirectional flow has to be considered. Bidirectional pedestrian flow includes unordered streams as well as lane-separated and thus quasi-unidirectional streams in opposite directions. A more detailed discussion and data can be found in [99,135,154]. A surprising finding is that the sum of flow and counterflow in corridors is larger than the unidirectional flow and for equally distributed loads it can be twice the unidirectional flow [94].

Another explanation is given by Helbing et al. [58] who argue that cultural and population differences are responsible for the deviations between Weidmann and their data. In contrast to this interpretation the data of Hanking and Wright [49] gained by measurements in the London subway (UK) are in good agreement with the data of Mori and Tsukaguchi [115] measured in the central business district of Osaka (Japan), both on strictly uni-directional streams. This brief discussion clearly shows that up to now there is no consensus about the origin of the discrepancies between different fundamental diagrams and how one can explain the shape of the function.

However, all diagrams agree in one characteristic: velocity decreases with increasing density. As the discussion above indicates there are many possible reasons and causes for the velocity reduction. For the movement of pedestrians along a line, a linear relation between speed and the inverse of the density was measured in [174]. The speed for walking pedestrians depends also linearly on the step size [192] and the inverse of the density can be regarded as the required length for one pedestrian to move. Thus it seems that smaller step sizes caused by a reduction of the available space with increasing density is, at least for a certain density region, one cause for the decrease of speed. However, this is only a starting point for a more elaborated modeling of the fundamental diagram.

### Bottleneck Flow

The flow of pedestrians through bottlenecks shows a rich variety of phenomena, e. g., the formation of lanes at the entrance to the bottleneck [64,66,93,176], clogging and blockages at narrow bottlenecks [24,57,93,121,122,150] or some special features of bidirectional bottleneck flow [57].

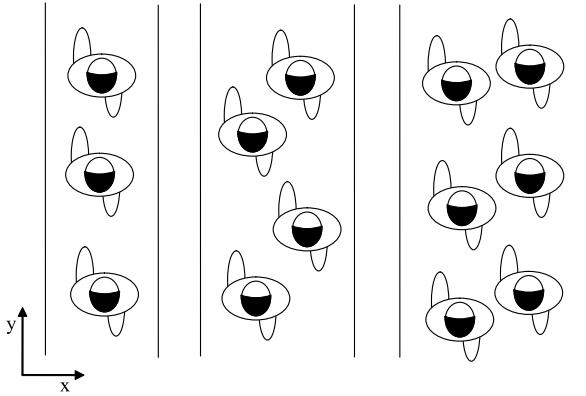
Moreover, the estimation of bottleneck capacities by the maxima of fundamental diagrams is an important tool for the design and dimensioning of pedestrian facilities.

**Capacity and Bottleneck Width** One of the most important practical questions is how the capacity of a bottleneck rises with increasing width. Studies of this dependence can be traced back to the beginning of the last century [24,32] and, up to now, have been discussed controversially. As already mentioned in the context of the fundamental diagram there are multiple possible influences on pedestrian flow and thus on the capacity. In the following, the major findings are outlined, demonstrating the complexity of the system and documenting a controversial discussion over one hundred years.

At first sight, a stepwise increase of capacity with the width appears to be natural if lanes are formed. For independent lanes, where pedestrians in one lane are not influenced by those in others, the capacity increases only if an additional lane can be formed. This is reflected in the stepwise enlargement of exit width, which has been a requirement of several building codes and design recommendations. See e. g., the discussion in [146] for the USA and GB and [130] for Germany. e. g.; the German building code requires an exit width (e. g., for a door) to be at least 90 cm plus 60 cm for every 200 persons. Independently from this simple lane model, Hoogendoorn and Daamen [64,66] measured by a laboratory experiment the trajectories of pedestrians passing a bottleneck. The trajectories show that inside a bottleneck the formation of lanes occurs, resulting from the zipper effect occurring on entry to the bottleneck. Due to the zipper effect, a self-organization phenomenon leading to an optimization of the available space and velocity; the lanes are not independent and thus do not allow passing (Fig. 4). The empirical results of [64,66] indicate a distance between lanes of  $d \approx 0.45$  m, independent of the bottleneck width  $b$ , implying a stepwise increase of capacity. However, the investigation was restricted to two values ( $b = 1.0$  m and  $b = 2.0$  m) of the width.

In contrast, the study [176] considered more values of the width and found that the lane distance increases continuously as illustrated in Fig. 4. Moreover it was shown that a continuous increase of the lane distance leads to a very weak dependence on its width of the density and velocity inside the bottleneck. Thus in reference to Eq. (2) the flow does not necessarily depend on the number of lanes. This is consistent with common guidelines and handbooks<sup>2</sup> which assume that the capacity is a linear function

<sup>2</sup>One exception is the German MVStättV [130], see above.



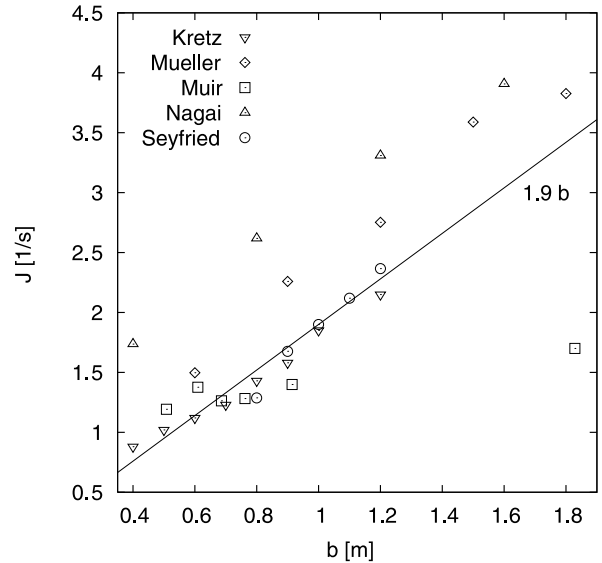
Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 4

A sketch of the zipper effect with continuously increasing lane distances in  $x$ : The distance in the walking direction decreases with increasing lateral distance. Density and velocities are the same in all cases, but the flow increases continuously with the width of the section

of the width [35,136,150,192]. It is given by the maximum of the fundamental diagram and in reference to the specific flow concept introduced in Subsect. “Observables”, Eqs. (2), (3), the maximum grows linearly with the facility width. To find a conclusive judgment on the question if the capacity grows continuously with the width the results of different laboratory experiments [93,121,122,132,176] are compared in [176].

In the following we discuss the data of flow measurement collected in Fig. 5. The corresponding setups are sketched in Fig. 6. First, note that all presented data are taken under laboratory conditions where the test persons are advised to move normally. The data by Muir et al. [121], who studied the evacuation of airplanes (see Fig. 6b), seem to support the stepwise increase of flow with the width. They show constant flow values for  $b > 0.6$  m. But the independence of flow over the large range from  $b = 0.6$  m to  $b = 1.8$  m indicates that in this special setup the flow is not restricted by the bottleneck width. Moreover, it was shown in [176] by determination of the trajectories that the distance between lanes changes continuously, invalidating the basic assumption leading to a stepwise increasing flow. Thus all collected data for flow measurements in Fig. 5 are compatible with a continuous and almost linear increase with the bottleneck width for  $b > 0.6$  m.

The data in Fig. 5 differ considerably in values of bottleneck capacity. In particular, the flow values of Nagai [132] and Müller [122] are much higher than the maxima of empirical fundamental diagrams (see Sub-



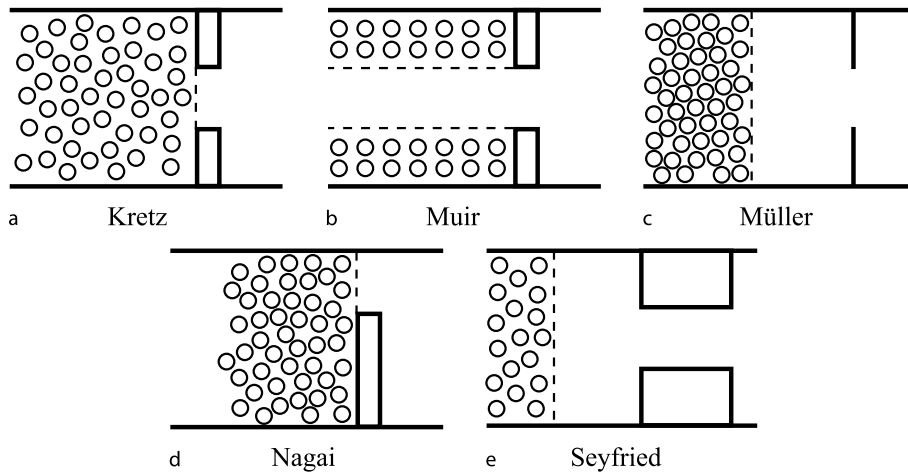
Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 5

Influence of the width of a bottleneck on the flow. Experimental data [121,122,132,176] of different types of bottlenecks and initial conditions. All data are taken under laboratory conditions where the test persons are advised to move normally

sect. “Fundamental Diagram”). The influence of “panic” or pushing can be excluded since in all experiments the participants were instructed to move normally. The comparison of the different experimental setups (Fig. 6) shows that the exact geometry of the bottleneck is of only minor influence on the flow, while a high initial density in front of the bottleneck can increase the resulting flow values. This is confirmed by the study of Nagai et al., see Figure 6 in [132]. There it is shown that for  $b = 1.2$  m the flow grows from  $J = 1.04 \text{ s}^{-1}$  to  $3.31 \text{ s}^{-1}$  when the initial density is increased from  $0.4 \text{ m}^{-2}$  to  $5 \text{ m}^{-2}$ .

The linear dependence of the flow on the width has a natural limitation due to the nonzero body-size and the asymmetry given by the sequence of movement in steps. Movement of pedestrians through bottlenecks smaller than shoulder width requires a rotation of the body. Kretz et al. found in their experiment [93] that the specific flow  $J_s$  increases if the width decreases from  $b = 0.7$  m to  $b = 0.4$  m.

**Connection Between Bottleneck Flow and Fundamental Diagrams** An interesting question is how the bottleneck flow is connected to the fundamental diagram. General results for driven diffusive systems [149] show that boundary conditions only *select* between the states of the



Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 6  
 Outlines of the experimental arrangements under which the data shown in Fig. 5 were taken

undisturbed system instead of creating completely different ones. Therefore it is surprising that the measured maximal flow at bottlenecks can exceed the maximum of the empirical fundamental diagram. These questions are related to the common jamming criterion. Generally, it is assumed that a jam occurs if the incoming flow exceeds the capacity of the bottleneck. In this case one expects the flow through the bottleneck to continue with the capacity (or lower values).

The data presented in [176] show a more complicated picture. While the density in front of the bottleneck amounts to  $\rho \approx 5.0(\pm 1) \text{ m}^{-2}$ , the density inside the bottleneck tunes around  $\rho \approx 1.8 \text{ m}^{-2}$ . The observation that the density inside the bottleneck is lower than in front of the bottleneck is consistent with measurements of Daamen and Hoogendoorn [20] and the description given by Predtechenskii and Milinskii in [150]. The latter assumes that in the case of a jam the flow through the bottleneck is determined by the flow in front of the bottleneck. The density inside the jam will be higher than the density associated with the capacity. Thus the reduced flow in front of the bottleneck causes a flow through the bottleneck smaller than the bottleneck capacity. Correspondingly the associated density is also smaller than that at capacity. But the discussion above cannot explain why the capacities measured at bottlenecks are significantly higher than the maxima of empirical fundamental diagrams and cast doubts on the common jamming criterion. Possible unconsidered influences are stochastic flow fluctuations, non-stationarity of the flow, flow interferences due to the necessity of local organization or changes of the incentive during the access into the bottleneck.

**Blockages in Competitive Situations** As stated above all data collected in Fig. 5 are gained by runs where the test persons were instructed to move normally. By definition a bottleneck is a limited resource and it is possible that under competitive situations pedestrian flow through bottlenecks is different from the flow in normal situations. One qualitative difference to normal situations is the occurrence of blockages. Regarding the term ‘panic’ one has to bear in mind that for the occurrence of blockages some kind of reward is essential, while the emotional state of the test persons is not. This was a result of a very interesting and often cited study by Mintz [113]. First experiments with real pedestrians have been performed by Dieckmann [24] in 1911 as a reaction to many fatalities in theater fires at the end of the 19th century. In these small scale experiments test persons were instructed to go through great trouble to pass the door as fast as possible. Even in the first run he observed a stable “wedging”. In [150] it is described how these obstruction occurs due to the formation of arches in front of the door under high pressure. This is very similar to the well-known phenomenon of *arching* occurring in the flow of granular materials through narrow openings [194].

Systematic studies including the influence of the shape and width of the bottleneck and comparisons with flow values under normal situations have been performed by Müller and Muir et al. [121,122]. Müller found that funnel-like geometries support the formation of arches and thus blockages. For further discussion, one must distinguish between temporary blockages and stable blockages leading to a zero flow. For the setup sketched in Fig. 6c Müller found that temporary blockages occur only for

$b < 1.8$  m. For  $b \leq 1.2$  m the flow shows strong pulsing due to unstable blockages. Temporal disruptions of the flow appear for  $b \leq 1.0$  m. In comparison to normal situations the flow is higher, and in general the occurrence of blockages decreases with width. However a surprising result is that for narrow bottlenecks, increasing the width can be counterproductive since it also increases the probability of blockages. Muir et al. for example note that in their setup (Fig. 6b) the enlargement of the width from  $b = 0.5$  m to  $b = 0.6$  m leads to an increase of temporary blockages. The authors explain this by differences in the perception of the situation by the test persons. While the smaller width is clearly passable only for one person, the wider width may lead to the perception that the bottleneck is sufficiently wide to allow two persons to pass through. How many people have direct access to the bottleneck is clearly influenced by the width of the corridor in front of the bottleneck. Also, Müller found hints that flow under competitive situations did not increase in general with the bottleneck width. He notes an optimal ratio of 0.75:1 between the bottleneck width and the width of the corridor in front of the bottleneck.

To reduce the occurrence of blockages, and thus evacuation times, Helbing et al. [54,55,83] suggested putting a column (asymmetrically) in front of a bottleneck. It should be emphasized that this theoretical prediction was made under the assumption that the system parameters, i. e., the basic behavior of the pedestrians, does not change in the presence of the column. This is highly questionable in real situations where a column can be perceived as an additional obstacle or can make it difficult to find the exit. In experiments [57] an increase of the flow of about 30% has been observed for a door with  $b = 0.82$  m. But this experiment was performed only for one width and the discussion above indicates the strong influence of the specific setup used. Independent of this uncertainty this concept is limited, as the occurrence of stable arches, to narrow bottlenecks. In practice narrow bottlenecks are not suitable for a large number of people and an opening in a room has other important functionalities, which would be restricted by a column.

Another finding is the observation that the total flow at bottlenecks with bidirectional movement is higher than it is for unidirectional flows [57].

## Stairs

In most evacuation scenarios stairs are important elements that are a major determinant for the evacuation time. Due to their physical dimension, which is often smaller than other parts of a building, or due to a reduced walking

speed, stairs generally must be considered as bottlenecks for the flow of evacuees. For the movement on stairs, just as for the movement on flat terrain, the fundamental diagram is of central interest. Compared to the latter there are more degrees of freedom, which influence the fundamental diagram:

- One has to distinguish between upward and downward movement.
- The influence of riser height and tread width (which determine the incline) has to be taken into account.
- For upward motion exhaustion effects lead to a strong time dependence of the free speed.

It is probably a consequence of the existence of a continuum of fundamental diagrams in dependence of the incline that there are no generally accepted fundamental diagrams for movement on stairs. However, there are studies on various details—mostly the free speed—of motion on stairs in dependence of the incline [35,38,39,46], conditions (comfortable, normal, dangerous) [151], age and sex [35], tread width [33], and the length of a stair [95]; and in consideration of various disabilities [11].

In addition there are some compilations or “meta studies”: Graat [46] compiled a list of capacity measurements and Weidmann [192] built an average of 58 single studies and found an average for the horizontal upstairs speed—the speed when the motion is projected to the horizontal level—of 0.610 m/s.

Depending on various parameters, the aforesaid studies report horizontal upward walking speeds varying over a wide range from 0.391 to 1.16 m/s. Interestingly, on one and the same short stairs it could be observed [95] that people on average walked faster up- than downwards.

There is also a model where the upstairs speed is calculated from the stair geometry (riser and tread) [183] and an empirical investigation of the collision avoidance behavior on stairs [37].

On stairs (up- as well as downward) people like to put their hand on the handrail, i. e., they tend to walk close to walls, even if there is no counterflow. This is in contrast to movement on flat terrain, where at least in situations of low density there is a tendency to keep some distance from walls.

The movement on stairs is typically associated with a reduction of the walking speed. For upward motion this follows from the increased physical effort required. This has two aspects: first, there is the physical potential energy that a pedestrian has to supply if he wants to rise in height; second, the motion process itself is more exertive – the leg has to be lifted higher – than during motion on a level, even if this motion process is executed only on the spot.

Concerning the potential energy there is no comparable effect for people going downstairs. But still one can observe jams forming at the upper end of downstairs streams. These are due to the slight hesitation that occurs when pedestrians synchronize their steps with the geometry of the (down-)stairs ahead. Therefore the bottleneck character of downstairs is less a consequence of the speed on the stairs itself and more of the transition from planar to downward movement, at least as long as the steps are not overly steep.

### Evacuations: Empirical Results

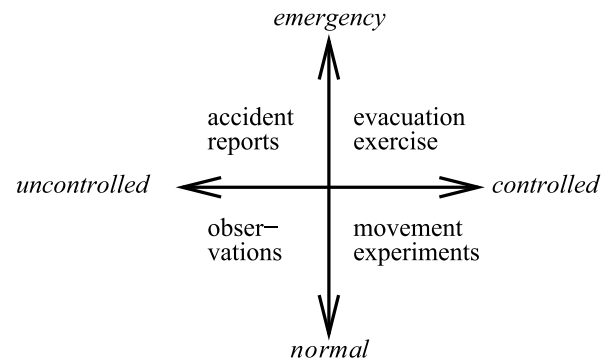
Up to now this section has focused on empirical results for pedestrian motion in rather simple scenarios. As we have seen there are many open questions where no consensus has been reached, sometimes even about the qualitative aspects. This becomes even more relevant for full-scale descriptions of evacuations from large buildings or cruise ships. These are typically a combination of many of the simpler elements, so a lack of reliable information is not surprising. In the following we will discuss several complex scenarios in more detail.

**Evacuation Experiments** In the case of an emergency, the movement of a crowd usually is more straightforward than in the general case. Commuters in a railway station, for example, or visitors of a building might have complex itineraries which are usually represented by origin-destination matrices. In the case of an evacuation, however, the aims and routes are known and usually the same, i. e., the exits and the egress routes. This is the reason why an evacuation process is rather strictly limited in space and time, i. e., its beginning and end are well-defined: the sound of the alarm, initial position of all persons, safe areas (final position of all persons), and the time at which the last person reaches the safe area. When all people have left a building or vessel and reached a safe area (or the lifeboats or life rafts), then the evacuation is finished. Therefore, it is also possible to perform evacuation trials and measure overall evacuation times. Before we go into details, we will clarify three different aspects of data on evacuation processes:

- (1) The definition and parts of evacuation time,
- (2) The different sources of data, and
- (3) The application of these data.

Concerning the evacuation time, five different phases can be distinguished [48,118,153]:

- (1) Detection time,
- (2) Awareness time,



Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 7

**Empirical data can be roughly classified according to controlled/uncontrolled and emergency/normal situations**

- (3) Decision time,
- (4) Reaction time, and
- (5) Movement time.

In IMO's regulations [118,119], the first four are grouped together into *response time*. Usually, this time is called *pre-movement time*

One possible scheme for the classification of data on evacuation processes is shown in the following Fig. 7.

Please note that not only data obtained from uncontrolled or emergency situations can be used in the context of evacuation assessment. Knowledge about bottleneck capacities (i. e., flows through doors and on stairs) is especially important when assessing the layout of a building with respect to evacuation. The purpose of empirical data in the context of evacuation processes (and modeling in general) is threefold [43,71]:

- (1) Identify parameters (factors that influence the evacuation process, e. g., bottleneck widths and capacities),
- (2) Quantify (calibrate) those parameters, e. g., flow through a bottleneck in persons per meter per second, and
- (3) Validate simulation results, e. g., compare the overall evacuation time measured in an evacuation with simulation or calculation results.

The validation is usually based on data from the evacuation of complete buildings, aircraft, trains or ships. These are available from two different sources:

- (1) Full scale evacuation trials and
- (2) Real evacuations.

Evacuation trials are usually observed and videotaped. Reports of real evacuation processes are obtained from eye-

witness records and a posteriori incident investigations. Since the setting of a complete evacuation is not experimental, it is hardly possible to measure microscopic features of the crowd motion. Therefore, calibration of parameters is usually not the main purpose in evacuation trials; rather, they are carried out to gain knowledge about the overall evacuation process, the behavior of the persons, to identify the governing influences/parameters and to validate simulation results.

One major concern in evacuation exercises is the well-being of the participants. Due to practical, financial, and ethical constraints, an evacuation trial cannot be, by nature, realistic. Therefore, an evacuation exercise does not convey the increased stress of a real evacuation. To draw conclusions on the evacuation process, the walking speed observed in an exercise should not be assumed to be higher in a real evacuation [145]. Along the same lines of argument, a simplified evacuation analysis based on, e.g., a hydro-dynamic model can predict an evacuation exercise, and the same constraints apply for its results concerning the prediction of evacuation times and the evacuation process. If population parameters (such as gender, age, walking speed, etc.) are explicitly stated in the model, increased stress can be simulated by adapting these parameters.

In summary, evacuation exercises are just too expensive, time consuming, and dangerous to be a standard measure for evacuation analysis. An evacuation exercise organized by the UK Marine Coastguard Agency on the Ro-Ro ferry “Stena Invicta” held in Dover Harbor in 1996 cost more than 10,000 GBP [117]. This is one major argument for the use of evacuation simulations based on hydro-dynamic models and calculations.

**Panic, Herding, and Similar Conjectured Collective Phenomena** As already mentioned earlier in Subsect. “Collective Effects”, the concept of “panic” and its relevance for crowd disasters is rather controversial. It is usually used to describe irrational and unsocial behavior. In the context of evacuations, empirical evidence shows that this type of behavior is rare [3,17,77,178]. On the other hand there are indications that fear might be “contagious” [22]. Related concepts like “herding” and “stampede” imply a certain similarity between the behavior of human crowds and animal behavior. This terminology is quite often used in the public media. *Herding* has been described in animal experiments [166] and is difficult to measure in human crowds. However, it seems to be natural that herding exists in certain situations, e.g., limited visibility due to failing lights or strong smoke when exits are hard to find.

*Panic* As stated earlier, “panic” behavior is usually characterized by selfish and anti-social behavior which through contagion affects large groups and even leads to completely irrational actions. Often it is assumed, especially in the media, to occur in situations where people compete for scarce or dwindling resources, which in the case of emergencies are safe space or access to an exit. However, this point of view does not stand close scrutiny and it has turned out that this behavior has played no role at all in many tragic events [73,77]. For these incidents *crowd disaster* is a much more appropriate characterization.

Furthermore, lack of social behavior seems to be more frequent during so called “acquisitive panics” or “crazes” [179] than during “flight panics”. That is, social behavior seems to be less stable if there is something to gain than if there is some external danger which threatens all members of a group. Examples of crazes (acquisitive panics) include the Victoria Hall Disaster (1883) [150], the crowning ceremony of Tsar Nicholas II (1896) [168], a governmental Christmas celebration in Aracaju (2001), the distribution of free Saris in Uttar Pradesh (2004), and the opening of an IKEA store in Jeddah (2004). Crowd accidents which occur at rock concerts and religious events as well bear more similarities with crazes than with panics.

However, it is not the case that altruism and cooperation increase with danger. The events during the capsizing of the MV Estonia (see Sect. 16.6 of [100]) show some behavioral threshold: faced with immediate life-threatening danger, most people struggle for their own survival or that of close relatives.

*Herding* Herding in a broad context means “go with the flow” or “follow the crowd”. Like “panic”, the term “herding” is often used in the context of stock market crashes, i. e., causing an avalanche effect. Like “panic” the term is usually not well defined and is used in an allegoric way. Therefore, it is advisable to avoid the term in a scientific context (apart from zoology, of course). Furthermore, “herding”, “stampede”, and “panic” have a strong connotation of “deindividuation”. The conjecture of an automatic deindividuation caused by large crowds [101] has been replaced by a social attachment theory (“the typical response to a variety of threats and disasters is not to flee but to seek the proximity of familiar persons and places”) [109].

*Stampede* Stampede is – like herding – a term from zoology where herds of large mammals, such as buffalo, collectively run in one direction and might overrun any obstacles. This is dangerous for human observers if they cannot get out of the way. The term “stampede” is sometimes used

for crowd accidents [73], too. It is furthermore assumed to be highly correlated with panic. When arguing along those lines, a stampede might be the result of “crowd panic” or vice versa.

**Shock or Density Waves** Shock waves are reported for rock concerts [180] and religious events [2,58]. They might result in people standing close to each other falling down. Pressures in dense crowds of up to 4, 450 N/m<sup>2</sup> have been reported.

Although empirical data on crowd disasters exist, e. g., in the form of reports from survivors or even video footage, it is almost impossible to derive quantitative results from them. Models that aim at describing such scenarios make predictions for certain counter-intuitive phenomena that should occur. In the faster-is-slower effect [54] a higher desired velocity leads to a slower movement of a large crowd. In the freezing-by-heating effect [53] increasing the fluctuations can lead to a more ordered state. For a thorough discussion we refer to [54,55] and references therein. However, from a statistical point of view there is insufficient data to decide the relevance of these effects in real emergency situations, not least because it is almost impossible to perform “realistic” experiments.

### Sources of Empirical Data on Evacuation Processes

The evacuation of a building can either be an isolated process (due to fire restricted to this building, a bomb threat, etc.) or it can be part of the evacuation of a complete area. We will focus on the single building evacuation, here. For the evacuation of complete areas, e. g., because of flooding or hurricanes, cf. [157] and references therein.

For passenger ships, a distinction between High Speed Craft (HSC), Ro-Ro passenger ferries, and other passenger vessels (cruise ships) is made. High Speed Craft do not have cabins and the seating arrangement is similar to aircraft. Therefore, there is a separate guideline for HSC [119]. A performance-based evacuation analysis at an early stage of design is required for HSC and Ro-Pax. There is currently no such requirement for cruise ships. For an overview of IMO’s requirements and the historical development up to 2001 cf. [27]. In addition to the five components for the overall evacuation time listed above, there are three more specific to ships:

- (6) Preparation time (for the life-saving appliances, i. e., lifeboats, life-rafts, davits, chutes),
- (7) Embarkation time, and
- (8) Launching time.

Therefore, the evacuation procedure on ships is more complex than for buildings. Additionally, SAR (Search

And Rescue) is an integral part of ship evacuation.

For High Speed Craft, the time limit is 17 minutes for evacuation [70], for Ro-Ro passenger ships it is 60 minutes [118], and for all other passenger ships (e. g., cruise ships) it is 60 minutes if the number of main vertical zones is less or equal to five and 80 minutes otherwise [118]. For HSC, no distinction is made between assembly and embarkation phases.

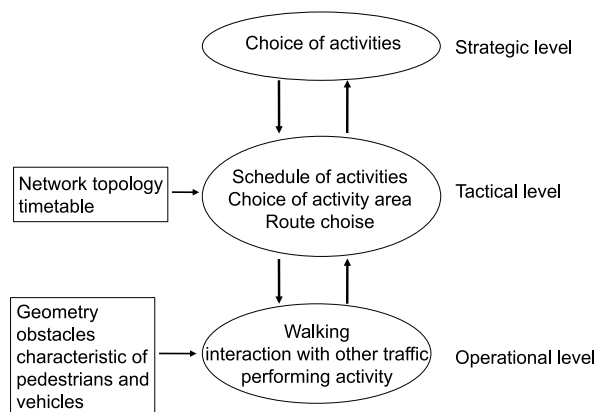
For aircraft, the approach can be compared to that of HSC. First, an evacuation test is mandatory and there is a time limit of 90 seconds that has to be complied to in the test [31].

In many countries there is no strict criterion for the maximum evacuation time of buildings. The requirements are based on minimum exit widths and maximum escape path lengths.

A number of real evacuations has been investigated and reports are publicly available. Among the most recent ones are: Beverly Hills Club [12], MGM Grand Hotel, [12], retail store [4], department store [1], World Trade Center [47] and [www.wtc.nist.gov](http://www.wtc.nist.gov), high-rise buildings [144, 173], theater [191] for buildings, High Speed Craft “Sleipner” [138] for HSC, an overview up to 1998 [143], exit width variation [121], double deck aircraft [74], another overview for aircraft [120], and for trains [43,169].

### Modeling

A comprehensive theory of pedestrian dynamics has to take into account three different levels of behavior (Fig. 8). At the *strategic level*, pedestrians decide which activities they like to perform and the order of these activities. With



Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 8

The different levels of modeling pedestrian behavior (after [19, 65])



the choices made at the strategic level, the *tactical level* concerns the short-term decisions made by the pedestrians, e. g., choosing the precise route taking into account obstacles, density of pedestrians etc. Finally, the *operational level* describes the actual walking behavior of pedestrians, e. g., their immediate decisions necessary to avoid collisions etc.

Processes at the strategic and tactical level are usually considered to be exogenous to pedestrian simulation. Here information from other disciplines (sociology, psychology etc.) is required. In the following we will mostly be concerned with the operational level, although some of the models that we are going to describe allow us to take into account certain elements of behavior at the tactical level as well.

Modeling on the operational level is usually based on variations of models from physics. Indeed the motion of pedestrian crowds shares certain similarities with fluids and the flow of granular materials. The goal is to find models which are as simple as possible, but at the same time can reproduce “realistic” behavior in the sense that the empirical observations are reproduced. Therefore, based on the experience from physics, pedestrians are often modeled as simple “particles” that interact with each other.

There are several characteristics which can be used to classify the modeling approaches:

**Microscopic vs. macroscopic** In microscopic models each individual is represented separately. Such an approach allows us to introduce different types of pedestrians with individual properties as well as issues such as route choice. In contrast, in macroscopic models, individuals cannot be distinguished. Instead the state of the system is described by densities, usually a mass density derived from the positions of the persons and a corresponding locally averaged velocity.

**Discrete vs. continuous** Each of the three basic variables for a description of a system of pedestrians, namely space, time and state variable (e. g., velocities), can be either discrete (i. e., an integer number) or continuous (i. e., a real number). Here all combinations are possible. In a cellular automaton approach all variables are by definition discrete, whereas in hydrodynamic models all are continuous. These are the most common choices, but other combinations are used as well. Sometimes for a cellular automata approach a continuous time variable is also allowed. In computer simulation this is realized through a *random-sequential update* where at each step the particle or site to be updated (moved) is chosen randomly (from *all* particles or sites, respectively). A discrete time is usually real-

ized through a *parallel* or *synchronous update* where all particles or sites are moved at the same time. This introduces a timescale. In so-called coupled map lattices time is discrete, whereas space and state variables are continuous.

**Deterministic vs. stochastic** The dynamics of pedestrians can either be deterministic or stochastic. In the first case the behavior at a certain time is completely determined by the present state. In stochastic models, behavior is controlled by certain probabilities such that the agents can react differently in the same situation. This is one of the lessons learnt from the theory of complex systems where it has been shown for many examples that through introduction of stochasticity into rather simple systems very complex behavior can be generated. On the other hand, the stochasticity in the models reflects our lack of knowledge of the underlying physical processes that, e. g., determine the decision-making of the pedestrians. Through stochastic behavioral rules it often becomes possible to generate a rather realistic representation of complex systems such as pedestrian crowds.

This “intrinsic” stochasticity should be distinguished from “noise”. Sometimes external noise terms are added to the *macroscopic* observables, such as position or velocity. Often the main effect of these terms is to avoid certain special configurations which are considered to be unrealistic, like completely blocked states. Otherwise the behavior is very similar to the deterministic case. For true stochasticity, on the other hand, the deterministic limit usually has very different properties from the generic case.

**Rule-based vs. force-based** Interactions between the agents can be implemented in at least two different ways: In a rule-based approach agents make “decisions” based on their current situation, the nature of their neighborhood as well as their goals, etc. It focuses on the *intrinsic properties* of the agents and thus the rules are often justified from psychology. In force-based models, agents “feel” a force exerted by others and the infrastructure. They therefore emphasize *extrinsic properties* and their relevance for the motion of the agents. This is a physical approach based on the observation that the presence of others leads to deviations from a straight motion. In analogy to Newtonian mechanics a force is made responsible for these accelerations.

Cellular automata are typically rule-based models, whereas, e. g., the social-force model belongs to the force-based approaches. However, sometimes a clear

distinction cannot be made; many models combine aspects of both approaches.

**High vs. low fidelity** *Fidelity* here refers to the apparent realism of the modeling approach. High fidelity models try to capture the complexity of decision making, actions, etc. that constitute pedestrian motion in a realistic way. In contrast, in the simplest models pedestrians are represented by particles without any intelligence. Usually the behavior of these particles is determined by “forces”. This approach can be extended, e. g., by allowing different “internal” states of the particles so that they react differently to the same force depending on the internal state. This can be interpreted as some kind of “intelligence” and leads to more complex approaches, like multi-agent models. Roughly speaking, the number of parameters in a model is a good measure for fidelity in the sense introduced here, but note that higher fidelity does not necessarily mean that empirical observations are reproduced better!

It should be mentioned that a clear classification according to the characteristics outlined here is not always possible. In the following we will describe some model classes in more detail.

### Fluid-dynamic and Gas kinetic Models

Pedestrian dynamics has some obvious similarities with fluids. For example, the motion around obstacles appears to follow “streamlines”. Motion at intermediate densities is restricted (short-ranged correlations). Therefore it is not surprising that, very much like for vehicular dynamics, the earliest models of pedestrian dynamics took inspiration from hydrodynamics or gas-kinetic theory [50,61,68,69]. Typically these macroscopic models are deterministic, force-based and of low fidelity.

Henderson [60,61] has tried to establish an analogy of large crowds with a classical gas. From measurements of motion in different crowds in a low density (“gaseous”) phase he found good agreement of the velocity distribution functions with Maxwell-Boltzmann distribution [60].

Motivated by this observation, he later developed a fluid-dynamic theory of pedestrian flow [61]. Describing the interactions between the pedestrians as a collision process where the particles exchange momenta and energy, a homogeneous crowd can be described by the well-known kinetic theory of gases. However, the interpretation of the quantities is not entirely clear, e. g., what the analogues of pressure and temperature are in the context of pedestrian motion. Temperature could be identified with the velocity

variance, which is related to the distribution of desired velocities, whereas the pressure expresses the desire to move against a force in a certain direction.

The applicability of classical hydrodynamic models is based on several conservation laws. The conservation of mass, corresponding to conservation of the total number of pedestrians, is expressed through a continuity equation of the form

$$\frac{\partial \rho(\mathbf{r}, t)}{\partial t} + \nabla \cdot \mathbf{J}(\mathbf{r}, t) = 0, \quad (6)$$

which connects the local density  $\rho(\mathbf{r}, t)$  with the current  $\mathbf{J}(\mathbf{r}, t)$ . This equation can be generalized to include source and sink terms. However, the assumption of conservation of energy and momentum is not true for interactions between pedestrians which in general do not even satisfy Newton’s Third Law (“actio = reaction”). In [50] several other differences to normal fluids were pointed out, e. g., the anisotropy of interactions or the fact that pedestrians usually have an individual preferred direction of motion.

In [50] a better founded fluid-dynamical description was derived on the basis of a gas kinetic model which describes the system in terms of a density function  $f(\mathbf{r}, \mathbf{v}, t)$ . The dynamics of this function are determined by Boltzmann’s transport equation that describes its change for a given state as difference of inflow and outflow due to binary collisions.

An important new aspect in pedestrian dynamics is the existence of desired directions of motion which allows us to distinguish different groups  $\mu$  of particles. The corresponding densities  $f_\mu$  change in time due to four different effects:

1. A relaxation term with characteristic time  $\tau$  describes tendency of pedestrians to approach their intended velocities.
2. The interaction between pedestrians is modeled by a Stosszahlansatz as in the Boltzmann equation. Here, pair interactions between types  $\mu$  and  $\nu$  occur with a total rate that is proportional to the densities  $f_\mu$  and  $f_\nu$ .
3. Pedestrians are allowed to change from type  $\mu$  to  $\nu$  which, e. g., accounts for turning left or right at a crossing.
4. Additional gain and loss terms allow us to model entrances and exits where pedestrian can enter or leave the system.

The resulting fluid-dynamic equations derived from this gas kinetic approach are similar to that of ordinary fluids. However, due to the different types of pedestrians, corre-

sponding to individuals who have approximately the same desired velocity, one actually obtains a set of coupled equations describing several interacting fluids. These equations contain additional characteristic terms describing the approach to the intended velocity and the change of fluid-type due to interactions in avoidance maneuvers.

Equilibrium is approached through the tendency to walk with the intended velocity, not through interactions as in ordinary fluids. Momentum and energy are not conserved in pedestrian motion, but the relaxation towards the intended velocity describes a tendency to restore these quantities.

Unsurprisingly for a macroscopic approach, the gasekinetic models have problems at low densities. For a discussion, see e. g. [50].

**Hand Calculation method** For practical applications effective engineering tools have been developed from the hydrodynamical description. In engineering these are often called *hand calculation methods*. One could also classify some of them as queuing models since the central idea is to describe pedestrian dynamics as flow on a network with links of limited capacities. These methods allow us to calculate evacuation times in a relatively simple way that does not require any simulations. Parameters entering in the calculations can be adapted to the situation that is studied. Often they are based on empirical results, e. g., evacuation trials. Details about this kind of model can be found in Subsect. “**Calculation of Evacuation Times**”.

### Social-Force Models

The social-force model [52] is a deterministic continuum model in which interactions between pedestrians are implemented by using the concept of a *social force* or *social field* [103]. It is based on the idea that changes in behavior can be understood in terms of fields or forces. Applied to pedestrian dynamics, the social force  $\mathbf{F}_j^{(\text{soc})}$  represents the influence of the environment (other pedestrians, infrastructure) and changes the velocity  $\mathbf{v}_j$  of pedestrian  $j$ . Thus it is responsible for acceleration which justifies the interpretation as a force. The basic equation of motion for a pedestrian of mass  $m_j$  is then of the general form

$$\frac{d\mathbf{v}_j}{dt} = \mathbf{f}_j^{(\text{pers})} + \mathbf{f}_j^{(\text{soc})} + \mathbf{f}_j^{(\text{phys})} \quad (7)$$

where  $\mathbf{f}_j^{(\text{soc})} = \frac{1}{m_j} \mathbf{F}_j^{(\text{soc})} = \sum_{l \neq j} \mathbf{f}_{jl}^{(\text{soc})}$  is the total (specific) force due to other pedestrians.  $\mathbf{f}_j^{(\text{pers})}$  denotes a “personal” force which makes a pedestrian attempt to move with his or her own preferred velocity  $\mathbf{v}_j^{(0)}$  and thus acts as a driving

term. It is given by

$$\mathbf{f}_j^{(\text{pers})} = \frac{\mathbf{v}_j^{(0)} - \mathbf{v}_j}{\tau_j} \quad (8)$$

where  $\tau_j$  is reaction or acceleration time. In high density situations, physical forces  $\mathbf{f}_{jl}^{(\text{phys})}$  also become important, e. g., friction and compression when pedestrians make contact.

The most important contribution to the social force  $\mathbf{f}_j^{(\text{soc})}$  comes from the territorial effect, i. e., the private sphere. Pedestrians feel uncomfortable if they get too close to others, which effectively leads to a repulsive force between them. Similar effects are observed for the environment, e. g., people prefer not to walk too close to walls.

Since social forces are difficult to determine empirically, some assumptions must be made. Usually an exponential form is assumed. Describing the pedestrians as disks of radius  $R_j$  and position (of the center of mass)  $\mathbf{r}_j$ , the typical structure of the force between the pedestrians is described by [54]

$$\mathbf{f}_{jl}^{(\text{soc})} = A_j \exp \left[ \frac{R_{jl} - \Delta r_{jl}}{\xi_j} \right] \mathbf{n}_{jl} \quad (9)$$

with  $R_{jl} = R_j + R_l$ , the sum of the disk radii,  $\Delta r_{jl} = |\mathbf{r}_j - \mathbf{r}_l|$ , the distance between the centers of mass,  $\mathbf{n}_{jl} = \mathbf{r}_j - \mathbf{r}_l / \Delta r_{jl}$ , the normalized vector pointing from pedestrian  $l$  to  $j$ .  $A_j$  can be interpreted as strength,  $\xi_j$  as the range of the interactions.

The appeal of the social-force model is given mainly by analogy to Newtonian dynamics. For the solution of the equations of motion of Newtonian many-particle systems, well-founded molecular dynamics techniques exist. However, in most studies so far the distinctions between pedestrian and Newtonian dynamics are not discussed in detail. A straightforward implementation of the equations of motion neglecting these distinctions can lead to unrealistic movement of single pedestrians. For example, negative velocities in the main moving direction cannot be excluded in general even if asymmetric interactions (violating Newton’s Third Law) between the pedestrians are chosen. Another effect is the occurrence of velocities higher than the preferred velocity  $v_j^{(0)}$  due to the forces on pedestrians in the moving direction. To prevent this effect, additional restrictions for the degrees of freedom must be introduced, see for example [52], or the superposition of forces has to be discarded [175]. A general discussion of the limited analogy between Newtonian dynamics and the social-force model as well as the consequences for model implementations is still missing.

Apart from the ad hoc introduction of interactions, the structure of the social-force model can also be derived from an extremal principle [62,63]. It follows under the assumption that pedestrian behavior is determined by the desire to minimize a certain cost function which takes into account not only kinematic aspects and walking comfort, but also deviations from a planned route.

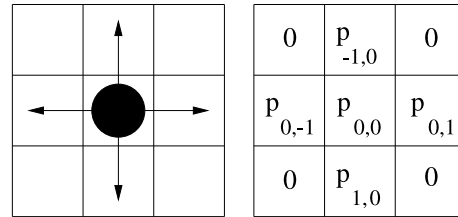
### Cellular Automata

Cellular automata (CA) are rule-based dynamical models that are discrete in space, time and state variable which in the case of traffic usually corresponds to velocity. Discreteness in time means that the positions of the agents are updated in well defined steps. In computer simulations this is realized through a *parallel* or *synchronous* update where all pedestrians move at the same time. The time step corresponds to a natural timescale  $\Delta t$  which could be identified, e. g., with some reaction time. This can be used for the calibration of the model which is essential for making quantitative predictions. A natural space discretization can be derived from the maximal densities observed in dense crowds which gives the minimal space requirement of one person. Usually each cell in the CA can be occupied by only one particle (exclusion principle) so that this space requirement can be identified with the cell size. In this way, a maximal density of  $6.25 \text{ P/m}^2$  [192] leads to a cell size of  $40 \times 40 \text{ cm}^2$ . Sometimes finer discretizations are more appropriate (see Subsect. “Theoretical Results”). In this case pedestrians correspond to extended particles that occupy more than one cell (e. g., four cells). The exclusion principle and the modeling of humans as non-compressible particles mimics short-range repulsive interactions, i. e., the “private-sphere”.

The dynamics are usually defined by rules which specify transition probabilities for motion to one of the neighboring cells (Fig. 9). The models differ in the specification of these probabilities as well in that of the “neighborhood”. For deterministic models, all but one are of probability zero.

The first cellular automata (CA) models [7,41,89,129] for pedestrian dynamics can be considered two-dimensional variants of the asymmetric simple exclusion process (ASEP) (for reviews, see [9,23,172]) or models for city or highway traffic [6,16,133] based on it. Most of these models represent pedestrians by particles without any internal degrees of freedom. They can move to one of the neighboring cells based on certain transition probabilities which are determined by three factors:

(1) The desired direction of motion, e. g., to find the shortest connection,



Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 9

A particle, its possible directions of motion and the corresponding transition probabilities  $p_{ij}$  for the case of a von Neumann neighborhood

- (2) Interactions with other pedestrians, and
- (3) Interactions with the infrastructure (walls, doors, etc.).

**Fukui–Ishibashi Model** One of the first CA models for pedestrian dynamics was proposed by Fukui and Ishibashi [40,41] and is based on a two-dimensional variant of the ASEP. They studied bidirectional motion in a long corridor where particles moving in opposite directions were updated alternatingly. Particles move deterministically in their desired direction; only if the desired cell is occupied by an oppositely moving particle do they make a random sidestep.

Various extensions and variations of the model have been proposed, e. g., an asymmetric variant [129] where walkers prefer lane changes to the right, different update types [193], simultaneous (exchange) motion of pedestrians standing “face-to-face” [72], or the possibility of backstepping [107]. The influence of the shape of the particles has been investigated in [131]. Also other geometries [128, 181] and extensions to full 2-dimensional motion have been studied in various modifications [106,107,127]

**Blue–Adler Model** The model of Blue and Adler [7,8] is based on a variant of the Nagel–Schreckenberg model [133] of highway traffic. Pedestrian motion is considered in analogy to a multi-lane highway. The structure of the rules is similar to the basic two-lane rules suggested in [159]. The update is performed in four steps which are applied to all pedestrians in parallel. In the first step each pedestrian chooses a preferred lane. In the second step the lane changes are performed. In the third step the velocities are determined based on the available gap in the new lanes. Finally, in the fourth step the pedestrians move forward according to the velocities determined in the previous step.

In counterflow situations head-on-conflicts occur. These are resolved stochastically and with some probability opposing pedestrians are allowed to exchange positions within one time step. Note that the motion of a single pedestrian (not interacting with others) is deterministic otherwise.

Unlike the Fukui–Ishibashi model, motion is not restricted to nearest-neighbor sites. Instead, pedestrians can have different velocities  $v_{\max}$  which correspond to the maximal number of cells they are allowed to move forward. In contrast to vehicular traffic, acceleration to  $v_{\max}$  can be assumed to be instantaneous in pedestrian motion.

In order to study the effects of inhomogeneities, the pedestrians are assigned different maximal velocities  $v_{\max}$ . Fast walkers have  $v_{\max} = 4$ , standard walkers  $v_{\max} = 3$  and slow walkers  $v_{\max} = 2$ . The cell size is assumed to be  $50 \text{ cm} \times 50 \text{ cm}$ . The best agreement with empirical observations has been achieved with 5% slow and 5% fast walkers [8]. Furthermore the fundamental diagram in more complex situations, such as bi- or four-directional flows, has been investigated.

**Gipps–Marksjös Model** A more sophisticated discrete model was suggested by Gipps and Marksjös [45] in 1985. One motivation for developing a discrete model was the limited computer power at that time. Therefore a discrete model, which reproduces the properties of pedestrian motion realistically, was in many respects a real improvement over the existing continuum approaches.

Interactions between pedestrians are assumed to be repulsive, anticipating the idea of social forces (see Subsect. “**Social-Force Models**”). The pedestrians move on a grid of rectangular cells of size  $0.5 \times 0.5 \text{ m}$ . To each cell a score is assigned based on its proximity to other pedestrians. This score represents the repulsive interactions and actual motion is then determined by the competition between these repulsions and the gain of approaching the destination. Applying this procedure to all pedestrians, a potential value is assigned to each cell which is the sum of the individual contributions. The pedestrian then selects the cell of its nine neighbors (Moore neighborhood) which leads to the maximum benefit. This benefit is defined as the difference between the gain of moving closer to the destination and the cost of moving closer to other pedestrians as represented by the potential. This requires a suitable chosen gain function  $P$ .

The updating is done sequentially to avoid conflicts of several pedestrians trying to move to the same position. In order to model different velocities, faster pedestrians are updated more frequently. Note that the model dynamics are deterministic.

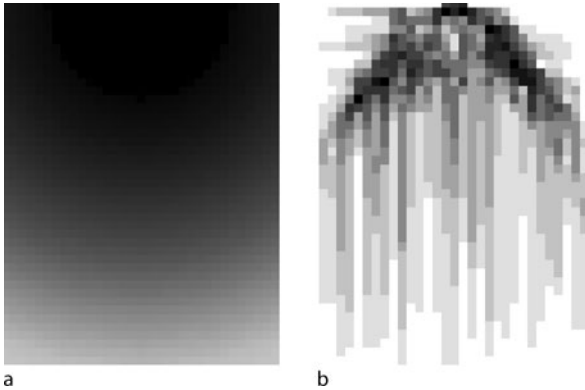
**Floor Field CA** Floor field CA [13,14,83,167] can also be considered as an extension of the ASEP. However, the transition probabilities to neighboring cells are no longer fixed but vary dynamically. This is motivated by the process of chemotaxis (see [5] for a review) used by some insects (e. g., ants) for communication. They create a chemical trace to guide other individuals to food sources. In this way a complex trail system is formed that has many similarities with human transport networks.

In the approach of [13] the pedestrians also create a trace. In contrast to chemotaxis, however, this trace is only virtual, although one could assume that it corresponds to some abstract representation of the path in the mind of the pedestrians. Although this is mainly a technical trick which reduces interactions to local ones that allow efficient simulations in arbitrary geometries, one could also think of the trail as representation of the paths in the mind of a pedestrian. The locality becomes important in complex geometries as no algorithm is required to check whether the interaction between particles is screened by walls, etc. The number of interaction terms always grows linearly with the number of particles.

The translation into local interactions is achieved by the introduction of so-called *floor fields*. The transition probabilities for all pedestrians depend on the strength of the floor fields in their neighborhood in such a way that transitions in the direction of larger fields are preferred. The *dynamic floor field*  $D_{ij}$  corresponds to a virtual trace which is created by the motion of the pedestrians and in turn influences the motion of other individuals. Furthermore it has its own dynamics, namely through diffusion and decay, which leads to a dilution and finally the vanishing of the trace after some time. The *static floor field*  $S_{ij}$  does not change with time since it only takes into account the effects of the surroundings. Therefore it exists even without any pedestrians present. It allows us to model, e. g., preferred areas, walls and other obstacles. Figure 10 shows the static floor field used for the simulation of evacuations from a room with a single door. Its strength decreases with increasing distance from the door. Since the pedestrians prefer motion into the direction of larger fields, this is already sufficient to find the door.

Coupling constants control the relative influence of both fields. For a strong coupling to the static field pedestrians will choose the shortest path to the exit. This corresponds to a ‘normal’ situation. A strong coupling to the dynamic field implies a strong herding behavior where pedestrians try to follow the lead of others. This often happens in emergency situations.

The model uses a fully parallel update. Therefore conflicts can occur where different particles choose the same

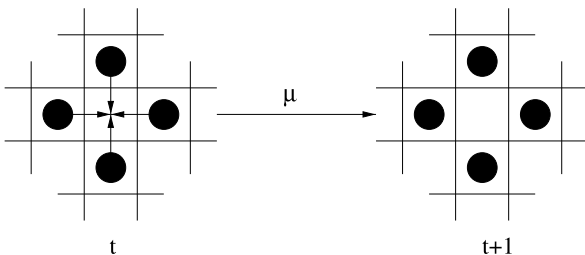


Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 10

**Left:** Static floor field for the simulation of an evacuation from a large room with a single door. The door is located in the middle of the upper boundary and the field strength increases with increasing intensity. **Right:** Snapshot of the dynamical floor field created by people leaving the room

destination cell. This is relevant for high density situations and happens in all models with parallel update if motion in different directions is allowed. Conflicts have been considered a technical problem for a long time and usually the dynamics have been modified in order to avoid them. The simplest method is to update pedestrians sequentially instead of using fully parallel dynamics. However, this leads to other problems, e. g., the identification of the relevant timescale. Therefore it has been suggested in [84,85] to take these conflicts seriously as an important part of the dynamics.

For the floor field model it has been shown in [85] that the behavior becomes more realistic if not all conflicts are resolved by allowing one pedestrian to move while the others stay at their positions. Instead with probability  $\mu \in [0, 1]$ , which is called the friction parameter, the movement of *all* involved pedestrians is denied [85] (see Fig. 11).



Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 11

Refused movement due to the friction parameter  $\mu$  (for  $m = 4$ )

This allows one to describe clogging effects between the pedestrians in a much more detailed way [85].  $\mu$  works as some kind of local pressure between the pedestrians. If  $\mu$  is high, the pedestrians handicap each other trying to reach their desired target sites. This local effect can have enormous influence on macroscopic quantities like flow and evacuation time [85]. Note that the kind of friction introduced here only influences interacting particles, not the average velocity of a freely moving pedestrian.

Surprisingly, the qualitative behaviors of the floor field model and the social-force models are very similar despite the fact that the interactions are very different. In the floor field model interactions are attractive, whereas in the social-force model they are repulsive. However, in the latter interactions are between particle densities. In contrast, in the floor field model the particle density interacts with the velocity density.

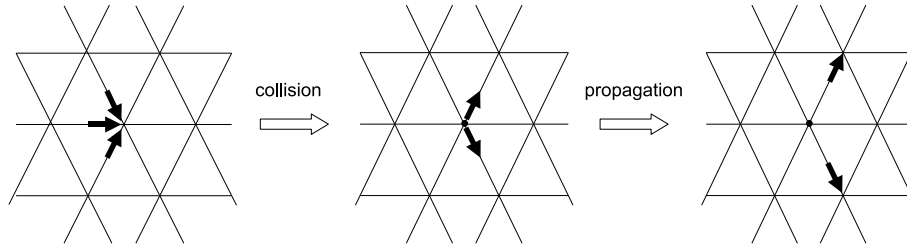
### Other Approaches

**Lattice-gas models** In 1986, Frisch, Hasslacher, and Pomeau [34] showed that one does not have to take into account detailed molecular motion within fluids in order to obtain a realistic picture of (2d) fluid dynamics. They proposed a lattice gas model [164,165] on a triangular lattice with hexagonal symmetry, which is similar in spirit to CA models, but the exclusion principle is relaxed: particles with different velocities are allowed to occupy the same site. Note that the allowed velocities differ only in the direction, not absolute value. The dynamics are based on a succession of collision and propagation that can be chosen in such a way that the coarse-grained averages of this microscopic dynamic is asymptotically equivalent to the Navier–Stokes equations of incompressible fluids.

In [108] a kind of mesoscopic approach inspired by these lattice gas models has been suggested as a model for pedestrian dynamics. In analogy with the description of transport phenomena in fluids (e. g., the Boltzmann equation) the dynamics are based on a succession of collision and propagation.

Pedestrians are modeled as particles, moving on a triangular lattice, which have a preferred direction of motion  $\mathbf{c}_F$ . However, the particles do not strictly follow this direction but also have a tendency to move with the flow. Furthermore, at high densities the crowd motion is influenced by a kind of friction which slows down the pedestrians. This is achieved by reducing the number of individuals allowed to move to neighboring sites.

As in a lattice gas model [165], the dynamics now consists of two steps. In the *propagation step* each pedestrian moves to the neighbor site in the direction of its velocity



Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 12

The dynamics of lattice gas models proceed in two steps. Pedestrians coming from neighboring sites interact in the collision step where velocities are redistributed. In the propagation step the pedestrians move to neighbor sites in directions determined by the collision step

vector. In the *collision step* the particles interact and new velocities (directions) are determined. In contrast to physical systems, momentum, etc., does not need to be conserved during the collision step. These considerations lead to a collision step that takes into account the favorite direction  $\mathbf{c}_F$ , the local density (the number of pedestrians at the collision site), and a quantity called mobility at all neighbor sites which is a normalized measure of the local flow after the collision.

**Optimal-Velocity Model** The optimal velocity (OV) model originally introduced for the description of highway traffic can be generalized to higher dimensions [134] which allows its application to pedestrian dynamics.

In the two-dimensional extension of the OV model the equation of motion for particle  $i$  is given by

$$\frac{d^2}{dt^2}\mathbf{x}_i(t) = a \left\{ \mathbf{v}_0 + \sum_j \mathbf{v}(\mathbf{x}_j(t) - \mathbf{x}_i(t)) - \frac{d}{dt}\mathbf{x}_i(t) \right\}, \quad (10)$$

where  $\mathbf{x}_i = (x_i, y_i)$  is the position of particle  $i$ . It can be considered as a special case of the general social-force model (7) without physical forces. The optimal-velocity function

$$\mathbf{V}(\mathbf{x}_j - \mathbf{x}_i) = f(r_{ij})(1 + \cos \varphi)\mathbf{n}_{ij}, \quad (11)$$

$$f(r_{ij}) = \alpha \{ \tanh \beta(r_{ij} - b) + c \}, \quad (12)$$

where  $r_{ij} = |\mathbf{x}_j - \mathbf{x}_i|$ ,  $\cos \varphi = (x_j - x_i)/r_{ij}$  and  $\mathbf{n}_{ij} = (\mathbf{x}_j - \mathbf{x}_i)/r_{ij}$  is determined by interactions with other pedestrians.  $\mathbf{v}_0$  is a constant vector that represents a ‘desired velocity’ at which an isolated pedestrian would move. The strength of the interaction depends on the distance  $r_{ij}$  between the  $i$ th and  $j$ th particles, and on the angle  $\varphi$  between the directions of  $\mathbf{x}_j - \mathbf{x}_i$  and the current velocity  $d\mathbf{x}_i/dt$ . Due to the term  $(1 + \cos \varphi)$ , a particle reacts more sensitively to particles in front than to those behind.

Now two cases can be distinguished: repulsive and attractive interactions. The former is relevant for pedestrian dynamics whereas the latter is more suitable for biological motion. Therefore, for pedestrian motion one chooses  $c = 1$  which implies  $f < 0$ .

A detailed analysis [134] shows that the model exhibits a rich phase diagram including the formation of various patterns.

**Other Models** We briefly mention a few other model approaches that have been suggested. In [10] a discretized version of the social-force model has been introduced and shown to reproduce qualitatively the observed collective phenomena.

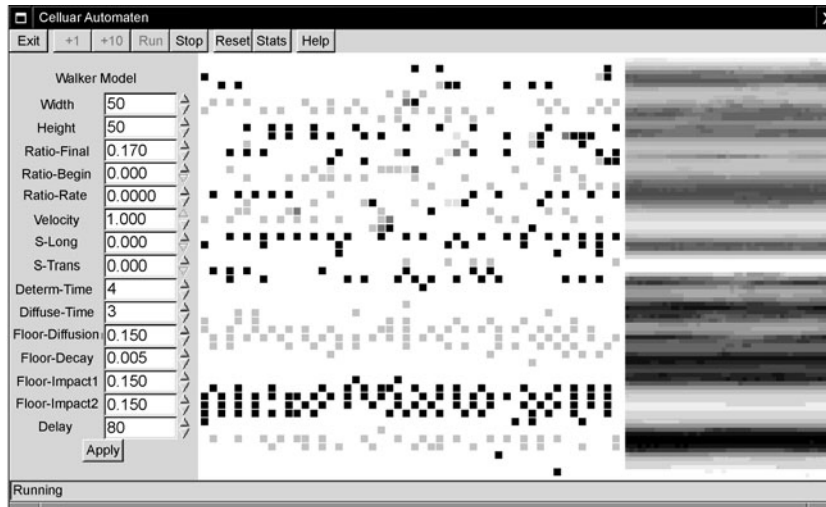
In [141] a magnetic force model has been proposed where pedestrians and their goals are treated as magnetic poles of opposite sign.

Another class of models is based on ideas from queuing theory. In principle, some hand calculation methods can be considered as macroscopic queuing models. Typically, rooms are represented as nodes in the queuing network and links correspond to doors. In microscopic approaches, in the movement process each agent chooses a new node, e. g., according to some probability [105].

## Theoretical Results

As emphasized in Subject. “**Collective Effects**”, the collective effects observed in the motion of pedestrian crowds are a direct consequence of microscopic dynamics. These effects are reproduced quite well by some models, e. g., the social-force and floor-field model, at least on a qualitative level. As mentioned before, the qualitative behavior of the two models is rather similar despite the very different implementation of the interactions. This indicates a certain robustness of the collective phenomena observed.

As an example we discuss the formation of lanes in counterflow formation. Empirically one observes a strong



Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 13

Lane formation in the floor-field model. The central window is the corridor and the light and dark squares are right- and left-moving pedestrians, respectively. In the bottom part well-separated lanes can be observed whereas in the top part the motion is still disordered. The right part of the figure shows the floor fields for the right-movers (*upper half*) and left-movers (*lower half*)

tendency to follow immediately in the “wake” of another person heading in the same direction. Such lane formation was reproduced in the social-force model [52,53] as well as in the floor-field model [13,76] (see Fig. 13). While the formation of lanes in general is essential to avoid deadlocks and thus keep the chance of reproducing realistic fluxes, the number of direction changes per meter cross section is a parameter which in reality crucially depends on the situation [76]: the longer a counterflow situation is assumed to persist, the fewer lanes per meter cross section can be found. The correct reproduction of counterflow is an issue for an accommodating animation, but more or less unimportant for macroscopic observables. This is probably the main reason why there seems to have been little effort put into the attempt to reproduce different “kinds” of lane formation in a controlled, situation-dependent manner.

On the quantitative side, the fundamental diagram is the first and most serious test for any model. Since most quantitative results rely on the fundamental diagram, it can be considered the most important characteristic of pedestrian dynamics. It is not only relevant for movement in a corridor or through a bottleneck, but also as an important determinant of evacuation times. However, as emphasized earlier, there is currently no consensus on the empirical form of the fundamental diagram. Therefore, a calibration of the model parameters is currently difficult.

Most cellular automata models are based on the asymmetric simple exclusion process. This strictly one-dimensional stochastic process has a fundamental diagram which is symmetric around density  $\rho = 1/2$ . Lane changes in two-dimensional extensions lead to only a small shift towards smaller densities. Despite the discrepancies in the empirical results, an almost symmetric fundamental diagram can be excluded.

Based on the experience with modeling of highway traffic [16,133], models with higher speeds have been introduced which naturally lead to an asymmetric fundamental diagram. Typically this is implemented by allowing the agents to move more than one cell per update step [82, 86,87,92,195,196]. These model variants have been shown to be flexible enough to reproduce, e. g., Weidmann’s fundamental diagram for the flow in a corridor [192] with high precision. Usually in the simulations a homogeneous population is assumed. However, in reality, different pedestrians have different properties such as walking speed, motivation, etc. This is easily taken into account in every microscopic model. There are many parameters that could potentially have an influence on the fundamental diagram. However, the current empirical situation does not allow to decide this question.

Another problem occurring in CA models has its origin in the discreteness of space. Through the choice of the lattice discretization, space is no longer isotropic. Motion in directions not parallel to the main axis of the lattice are



difficult to realize and can only be approximated by a sequence of steps parallel to the main directions.

Higher velocities also require the extension of the neighborhood of a particle which is no longer identical to the cells adjacent to the current position. A natural definition of “neighborhood” corresponds to those cells that could be reached within one time step. In this way the introduction of higher velocities also reduces the problem of space isotropy as the neighborhoods become more isotropic for larger velocities.

Other solutions to this problem have been proposed. One way is to count the number of diagonal steps and let the agent suspend from moving following certain rules which depend on the number of diagonal steps [171]. A similar idea is to sum up the real distance that an agent has moved during one round: a diagonal step counts  $\sqrt{2}$  and a horizontal or vertical step counts 1. An agent has to finish its round as soon as this sum is bigger than its speed [87]. A third possibility – which works for arbitrary speeds – is to assign selection probabilities to each of the four lattice positions adjacent to the exact final position [195,196]. Naturally these probabilities are inversely proportional to the square area between the exact final position and the lattice point, as in this case the probabilities are normalized by construction if one has a square lattice with points on all integer number combinations. However, one also could think of other methods to calculate the probability.

For the social-force model, the specification of the repulsive interaction (with and without hard core, exponential or reciprocal with distance) as well as the parameter sets for the forces changes in different publications [52,53,54,114]. In [55] the authors state that “most observed self-organization phenomena are quite insensitive to the specification of interaction forces”. However, at least for the fundamental diagram, a relation connected with all phenomena in pedestrian dynamics, this statement is questionable. As remarked in [56] the reproduction of the fundamental diagram “requires a less simple specification of the repulsive interaction forces”. Indeed in [175] it was shown that the choice of hard-core forces or repulsive soft interactions as well as the particular parameter set can strongly influence the resulting fundamental diagram regarding qualitative as well as quantitative effects.

Also a more realistic behavior at higher densities requires a modification of the basic model. Here the use of density-dependent desired velocities leads to a reduction of the otherwise unrealistically large number of collisions [10].

The particular specification of forces and the previously mentioned problem with Newton’s Third law can

lead in principle to some unwanted effects, such as momentary velocities larger than the preferred velocity [52] or the penetration of pedestrians into each other or into walls [98]. It is possible that these effects can be suppressed for certain parameter sets by contact or friction forces, but the general appearance is not excluded. Only in the first publication [52] are restrictions for the velocity explicitly formulated to prevent velocities larger than the intended speed; other authors tried to improve the model by introducing more parameters [98]. But additional parameter and artificial restrictions of variables diminish the simplicity and thus the attractiveness of the model. A general discussion of how to deal with these problems of the social-force model and a verification that the observed phenomena are not limited to a certain specification of the interaction and a special parameter set is up to now still missing.

While realistic reproduction within the empirical range of these macroscopic observables, especially the fundamental diagram, is absolutely essential to guarantee safety standards in evacuation simulations, and while a user should always be distrustful of models where no fundamental diagram has ever been published, it is by no means sufficient to exclusively check for the realism of macroscopic observables. On the microscopic level there are a large number of phenomena which need to be reproduced realistically, be it just to make a simulation animation look realistic or because microscopic effects can often easily influence macroscopic observables.

If one compares simulations of bottleneck flows with real events, one observes that in simulations the form of the queue in front of bottlenecks is often a half-circle, while in reality it is drop- or wedge-shaped. In most cases this discrepancy probably does not have an influence on the simulated evacuation time, but it is interesting to note where it originates from. Most simulation models implicitly or explicitly use some kind of utility maximization to steer the pedestrians – with the utility being foremost inversely proportional to the distance from the nearest exit. This obviously leads to half-circle-shaped queues in front of bottlenecks. So wherever one observes queues different than half-circles, people have exchanged their normal “utility function based on the distance” with something else. One such alternative utility function could be that people are just curious about what is inside or behind the bottleneck, so they seek a position where they can look into it. A more probable explanation would be that in any case it is the time distance not the spatial distance which is sought to be minimized. As anyone knows what the inescapable loss in time a bottleneck means for the whole waiting group, the precise waiting spot is not that impor-

tant. However, in societies with a strong feeling for equality, people would strongly wish to equally distribute the waiting time and keep a first-in-first-out principle, which can best be accomplished and controlled when the queue is more or less one-dimensional, respectively just as wide as the bottleneck itself.

Finally it should be mentioned that theoretical investigations based on simulations of models for pedestrian dynamics have led to the prediction of some surprising and counter-intuitive collective phenomena, such as the reduction of evacuation times through additional columns near exits (see Subsect. “**Bottleneck Flow**”) or the faster-is-slower [54] and freezing-by-heating effect [53]. However, so far the empirical evidence for the relevance or even the occurrence of these effects in real situations is rather scarce.

## Applications

In the following section we discuss more practical aspects of based on the modeling concepts presented in Sect. “**Modeling**”. Tools of different sophistication have been developed that are nowadays routinely used in safety analysis. The latter becomes more and more relevant since many public facilities must fulfill certain legal standards. As an example we mention aircrafts which must be evacuated within 90 seconds. The simulations etc. are already used in the planning stages because changes of the design at a later stage are difficult and expensive.

For this kind of safety analysis tools of different sophistication have been developed. Some of them mainly are able to predict just evacuation times whereas others are based on microscopic simulations which allow also to study various external influences (fire, smoke, ...) in much detail.

## Calculation of Evacuation Times

The basic idea of hand calculation methods has already briefly been described at the end of Subsect. “**Fluid-Dynamic and Gas Kinetic Models**”. Here we want to discuss its practical aspects in more detail.

The approach has been developed since the middle of the 1950s [185]. The basic idea of these methods is the assumption that people can be modeled to behave like fluids. Knowledge of the flow (see Eq. 1) and the technical data of the facility are then sufficient to evaluate evacuation times, etc.

Hand calculation method can be divided into two major approaches: methods with “dynamic” flow [35,42,78,79,80,136,151,152,163,192] and methods with “fixed” flow [110,123,124,125,126,137,145,173,185]. As methods

with “dynamic” flow we cite methods where the pedestrian flow is dependent on the density of the pedestrian stream (see Subsect. “**Observables**”) in the selected facility, thus the flow can be obtained from fundamental diagrams (see Subsect. “**Fundamental Diagram**”) or it is explicitly prescribed in the chosen method. This flow can change during movement through the building, e. g., by using stairs, thus the pedestrian stream has a “dynamic” flow. Methods with “fixed” flow do not use this concept of relationship between density and flow. In these methods selected facilities (e. g., stairs or doors) have a fixed flow which is independent from the density, which is usually not used in these methods. The “fixed” flow is usually based upon empirical and measured data of flow, which are specified for a special type of building, such as high-rise buildings or railway stations, for example. Because of much simplification, in these “fixed” flow methods a calculation can always be done very quickly.

Methods with “dynamic” flow allow one to describe the condition of the pedestrian flow in every part of a selected building or environment, because they are mostly based upon the continuity equation, thus it is possible to calculate different kinds of buildings. This allows the user to calculate transitions from wide to narrow, floor to door, floor to stair, etc. The disadvantage is that some these methods are very elaborate and time-intensive. But in general, a method with “dynamic” flow is not complicated to calculate, thus we want to divide hand calculation methods in simple [35,42,110,123,124,125,126,136,137,145,152,163,173,185,192] and complex [78,79,80,151] for evacuation calculation. All of these hand calculation methods are able to predict total evacuation times for a selected building, but differences between different methods still exist. Thus the user has to ensure that he is familiar with assumptions made by each method to ensure that a result is interpreted in a correct way [161].

## Simulation of Evacuation Processes

Before we go into the details of evacuation simulation, let us briefly clarify its scope and limitations and contrast it to other methods used in evacuation analysis. When analyzing evacuation processes, three different approaches can be identified:

- (1) Risk assessment,
- (2) Optimization, and
- (3) Simulation.

The aim and result of risk-assessment is a list of events and their consequences (e. g., damage, financial loss, loss of life), i. e., usually an event tree with probabilities and

expectation values for financial loss. Optimization aims at, roughly speaking, minimizing the evacuation time and reducing the area and duration of congestion. And finally, simulation describes a system with respect to its function and behavior by investigating a model of the system. This model is usually non-analytic, so does not provide explicit equations for the calculation of, e. g., evacuation time. Of course, simulations are used for “optimization” in a more general sense, too, i. e., they can be part of an optimization. This holds for risk assessment, too, if simulations are used to determine the outcomes of the different scenarios in the event tree.

In evacuation analysis the system is, generally speaking, a group of persons in an environment. More specifically, four components (sub-systems/sub-models) of the system *evacuation process* can be identified:

- (1) Geometry,
- (2) Environment,
- (3) Population, and
- (4) Hazards [43].

Any evacuation simulation must at least take into account (1) and (3). The behavior of the persons (which can be described on the strategic, tactical, and operational level—see Sect. “**Modeling**”) is part of the population sub-model. An alternative way of describing behavior is according to its algorithmic representation: no behavior modeling – functional analogy – implicit representation (equation) – rule based – artificial intelligence [43].

In the context of evacuation, hazards are first of all fire and smoke, which then require a toxicity sub-model, e. g., the fractional effective dose model (FED), to assess the physiological effect of toxic gases and temperature [25]. Further hazards to take into account might be earthquakes, flooding, or in the case of ships, list, heel, or roll motion. The sub-model environment comprises all other influences that affect the evacuation process, e. g., exit signs, surface texture, public address system, etc.

In summary, aims of an evacuation analysis and simulation are to provide feedback and hints for improvement at an early stage of design, information for safer and more rigorous regulations, improvement of emergency preparedness, training of staff, and accident investigation [43]. They usually do not provide direct results on the probability of a scenario or a systematic search for optimal geometries.

**Calculation of Overall Evacuation Time, Identification of Congestion, and Corrective Actions** The scope of this section is to show general results that can be obtained by evacuation simulations. They are general in the sense

that they can basically be obtained by any stochastic and microscopic model, i. e., apart from these two requirements, the results are not model specific. In detail, five different results of evacuation simulations can be distinguished:

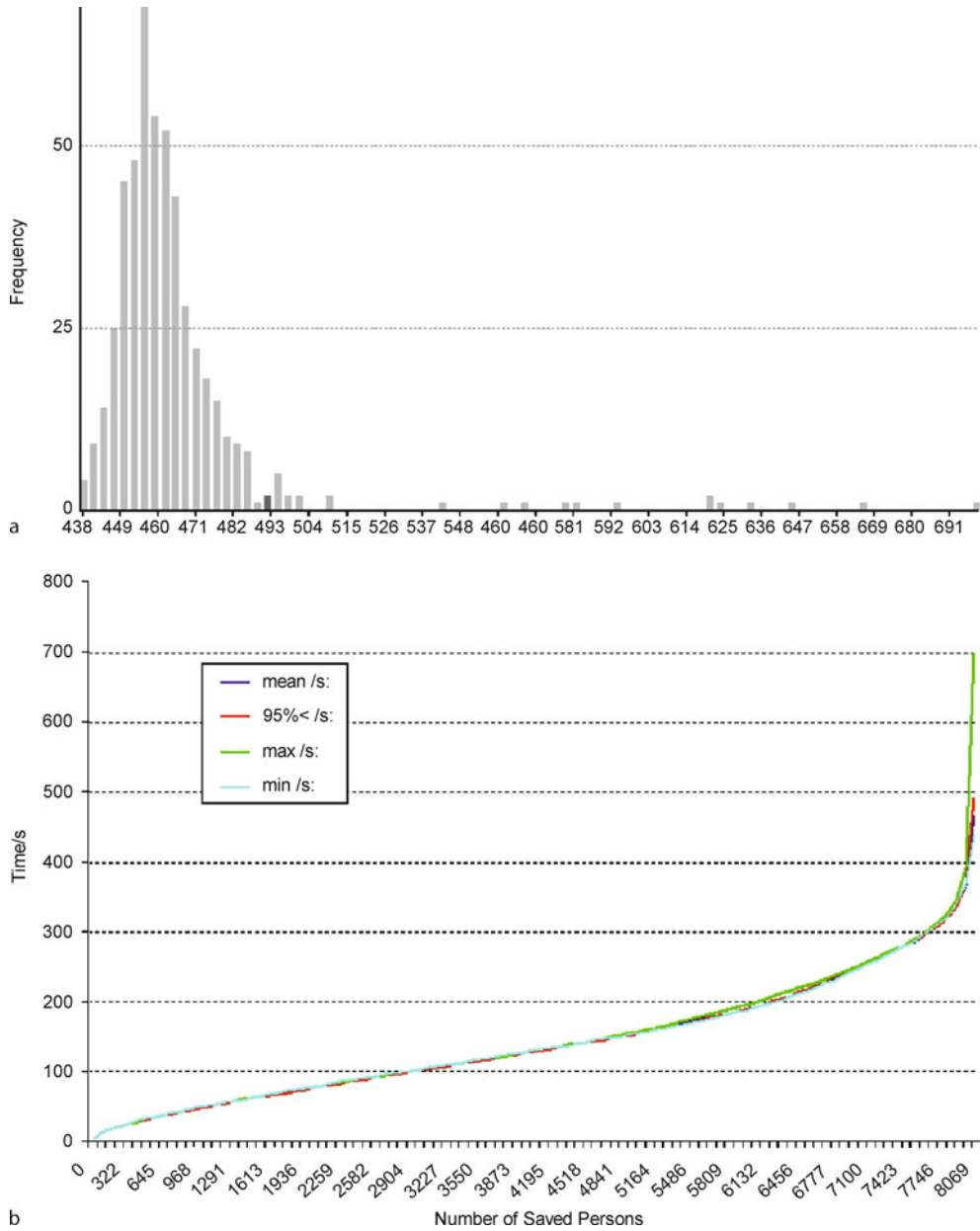
- (1) Distribution of evacuation times,
- (2) Evacuation curve (number of persons evacuated vs. time),
- (3) Sequence of the evacuation (e. g., snapshots/screenshots at specific times, e. g., every minute), and
- (4) Identification of congestion, usually based on density and time.

The last point (4), in particular, needs some more explanation: congestion is defined based on density. Notwithstanding the difficulties of measuring density, we suggest density as the most suitable criterion for the identification of congestion. In addition to the mere occurrence of densities exceeding a certain threshold (say 3.5 persons per square meter), the time this threshold is exceeded is another necessary condition for a sensible definition of congestion. In the case presented here, 10% of the overall evacuation time is used. Both criteria are in accordance with the IMO regulations [118].

Based on these results, evacuation time and areas of congestion, corrective actions can be taken. The most straightforward measure would be a change of geometry, i. e., shorter or wider escape paths (floors, stairs, doors). This can be directly put into the geometry sub-model, the simulation can be re-run, and the result checked. Secondly, the signage, and therefore the orientation capability, can be improved. This is not as straightforward as geometrical changes. It does depend more heavily on the model characteristics as to how these changes influence the evacuation sequence.

We will not go into these details in the following two sections but rather show two typical examples for evacuation simulations and the results obtained. We will also not discuss the results in detail, since they are of an illustrative nature in the context of this article. The following examples are based on investigations that have been performed using a cellular automaton model which is described along with the simulation program in [90,111].

**Simulation Example 1 – Hotel** The first example we show is a hotel with 8069 persons. In Fig. 15 only the ground floor is shown. There are nine floors altogether. The upper floors influence the ground floor only via the stair landings and the exits adjacent to them. Most of the 8069 persons are initially located in the ground floor, since the theater and conference area is located there. The upper



Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 14  
 Frequency distribution for the overall evacuation time (a) and evacuation curve (b)

floors are mainly covering bedrooms and some small conference areas.

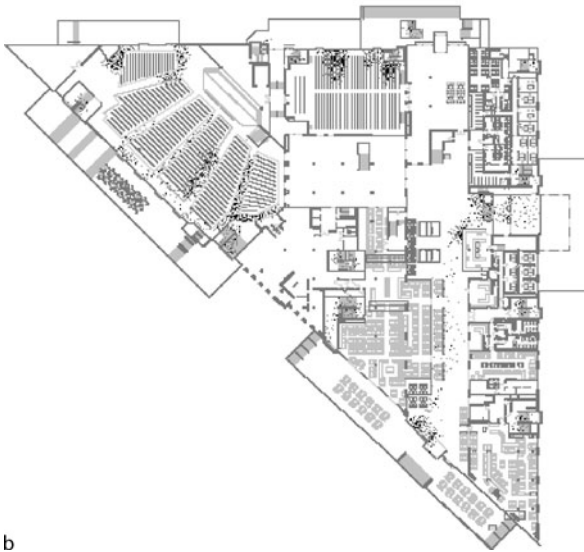
The first step in our example (which might well be a useful recipe for evacuation analyses in general and is again in accordance with [118]) is to perform a statistical analysis. To this end, 500 samples are simulated. The evacuation time of a single run is the time it takes for all persons to get out. In this context, no fire or smoke are

taken into account. Since there are stochastic influences in the model used, the significant overall evacuation time is taken to be the 95-percentile (cf. Fig. 14). Finally, the maximum, minimum, mean, and significant values for the evacuation curve (number of persons evacuated vs. time) are also shown in Fig. 14.

The next figure (Fig. 16) shows the cumulated density. The thresholds (red areas) are 3.5 persons per square me-



a



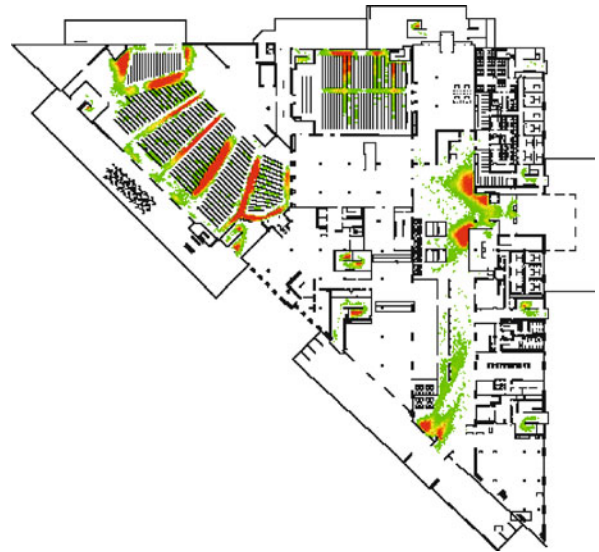
b

Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 15

Initial population distribution and situation after two minutes

ter and 10% of the overall evacuation time (in this case 49 seconds). The overall evacuation time is 8:13 minutes (493 seconds). This value is obtained by taking the 95-percentile of the frequency distribution for the overall evacuation times (cf. Fig. 14).

Of course, a distribution of overall evacuation times (for one scenario, i.e., the same initial parameters) can only be obtained by a stochastic model. In a deterministic model only one single value is calculated for the overall evacuation time. The variance of the overall evacuation



Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 16

Density plot, i.e., cumulated person density exceeding 3.5 persons per square meter and 10% of overall evacuation time

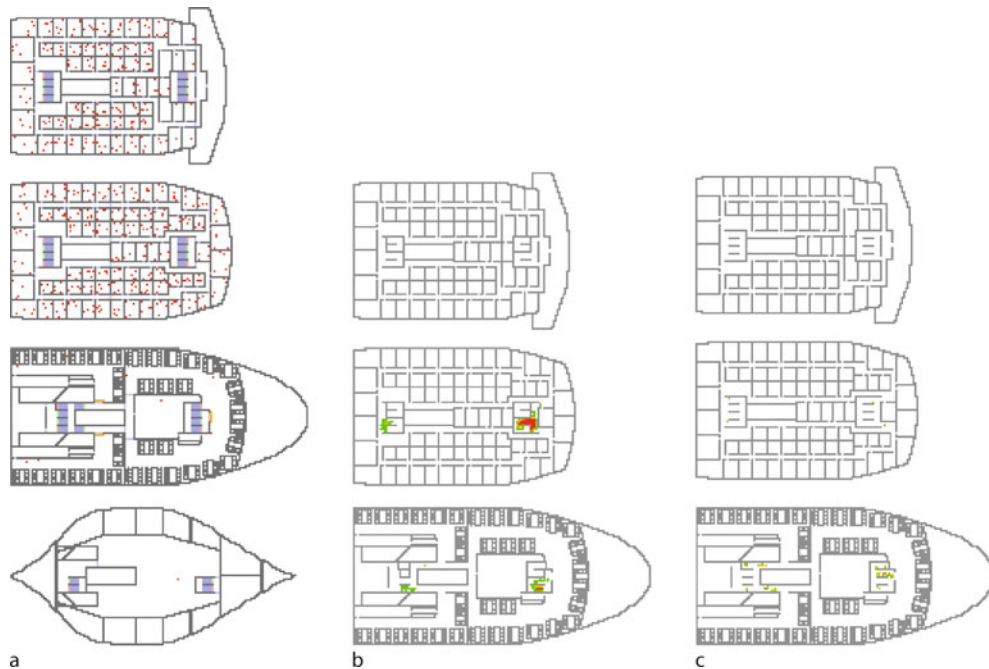
times is due to two effects in the model used here: the initial position of the persons is determined anew at the beginning of each simulation run since only the statistical properties of the overall population are set and the motion of the persons is governed by partially stochastic rules (e.g., probabilistic parameters).

**Simulation Example 2 – Passenger Ship** The second example we will show is a ship. The major difference from the previous example is the addition of (1) the assembly phase and (2) embarkation and launching.

$$\begin{aligned} T &= A + \frac{2}{3}(E + L) \\ &= f_{\text{safety}} \cdot (t_{\text{react}} + t_{\text{walk}}) + \frac{2}{3}(E + L) \\ &\leq 60 \text{ minutes} . \end{aligned}$$

Embarkation and launching time ( $E + L$ ) are required to be less than 30 minutes. For the sake of the evacuation analysis at an early design stage, the sum of embarkation and launching time can be assumed to be 30 minutes. Therefore, the requirement for  $A$  is 40 minutes. Alternatively, the embarkation and launching time can be determined by an evacuation trial.

Figure 17 shows the layout, initial population distribution (night case), density plot for the day case, and density plot for the night case. The reaction times are different for the day and the night case: 3 to 7 minutes (equally distributed) in the one and 7 to 13 minutes in the other. The



Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 17

Initial distribution for the night case, density plot for the day case, and density plot for the night case for the "AENEAS steamliner"

longer reaction time in the night case results in less congestion (cf. Fig. 17). Both cases must be done in the analysis according to [118]. Additionally, a secondary night and day case are required (making up four cases altogether). In these secondary cases the main vertical zone (MVZ) leading to the longest overall individual assembly time is identified, and then either half of the stairway capacity in this zone is assumed to be not available, or 50% of the persons initially located in this zone must be led via one neighboring zone to the assembly station.

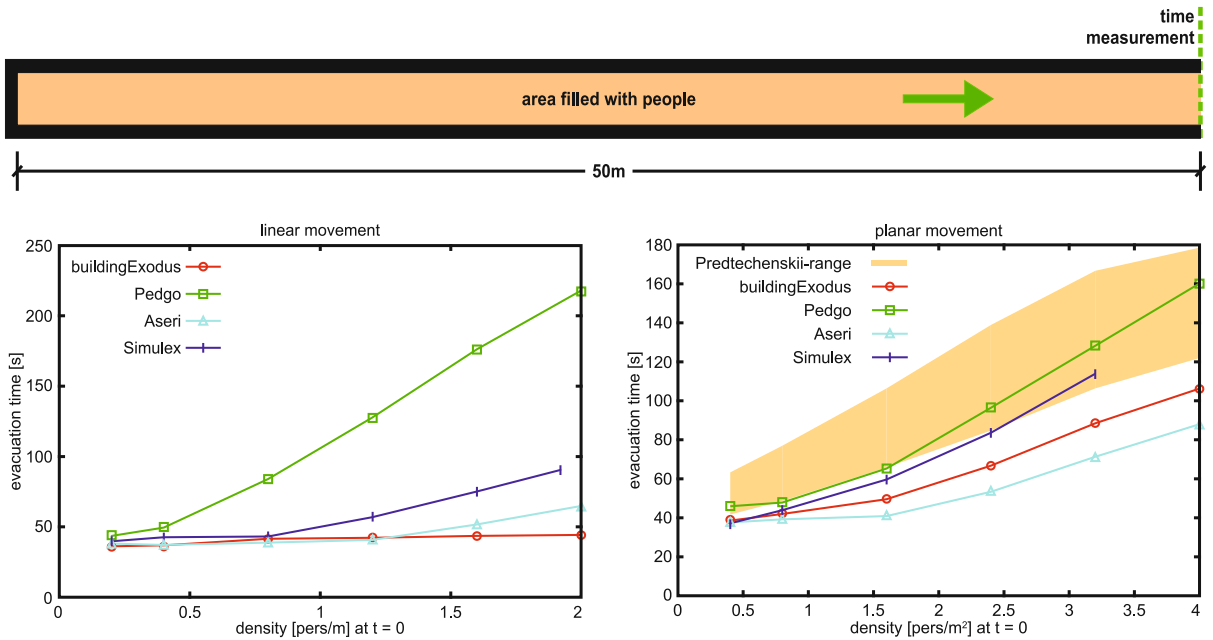
In the same way as shown for the two examples, simulations can be performed for other types of buildings and vessels. This technique has been applied to various passenger ships [112] to football stadiums [88] and the World Youth Day 2005 [88], the Jamarat Bridge in Makkah [88], a movie theater and schools (mainly for calibration and validation) [90] and airports [171]. Of course, many examples of applications based on various models can be found in the literature. For an overview, the proceedings of the PED conference series are an excellent starting point [44, 170, 190].

### Comparison of Commercial Software Tools

From a practical point of view, application of models for pedestrian dynamics and evacuation processes becomes

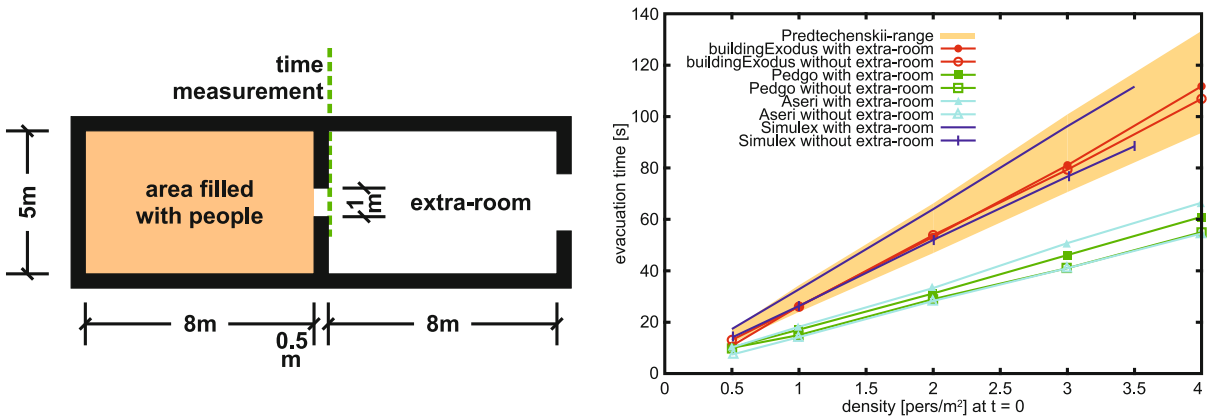
more and more relevant in safety analysis. This has led to the development of a number of software tools that, with different sophistication, help us study many aspects without risking the health of test persons in evacuation trials.

There are commercial, as well as non-commercial software tools. All tools might be based on different models [97, 187]. They have become very popular since the middle of the 1990s. A first comparison of different commercial software tools can be found in [191], where they were said to produce "reasonable results". Further comparisons of real evacuation data with software tools or hand calculation methods can be found in [29, 67, 91, 96, 104, 160, 161, 177]. But results predicted by different commercial software tools can differ by up to 40% for the same building [96]. Results may differ, too, when calculating with different assumptions, e.g., different reaction times, use of more or less detailed stair models, or when calculating with a real occupant load in contrast to an uncertainty analysis [96, 104]. Contrary to these results, another study [161] shows that calculations with different software tools are able to predict total evacuation times for high-rise buildings and there are no large differences as shown in [96]. In [161] the results of an evacuation trial and simulations with different commercial software tools differed for selected floors of a highrise building. The den-



Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 18

Comparison of different software tools by simulating linear (*left*, narrow floor) and planar (*right*, 2 m wide floor) movement [162]



Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 19

Comparison of different software tools by simulating a simple room geometry [162]

sities were very low in this instance. In this case human behavior has a very large influence on the evacuation time. By contrast, evacuations at medium or high densities, human behavior has a smaller influence on the evacuation time of selected areas because congestion appears and continues larger than in low density situations – thus people reach the exit while congestion is still a factor [162]. In low density situations congestions are very rare, thus people move narrowly with free walking velocity through the building [162].

But the results presented in [161] also show that commercial software tools sometimes have problems with the empirical relationship of density and walking speed (see Fig. 18). Furthermore, it is very important how boundary conditions are implemented in these tools (see Fig. 19), and the investigation of a simple scenario of a single room using different software tools shows results differing by about a factor of two (see Fig. 19) [161]. In this case all software tools predict a congestion at the exit. Furthermore it is possible that the implemented algorithm fails [161].

Thus for the user it is hard to know which algorithms are implemented in closed-source tools so that such a tool must be considered as “black box” [147]. It is also quite difficult to compare results about density and appearing congestions calculated by different software tools [162] and so it is questionable how these results should be interpreted. But, as pointed out earlier, reliable empirical data are often missing so that a validation of software tools or models is quite difficult [162].

### Future Directions

The discussion has shown that the problem of crowd dynamics and evacuation processes is far from being well understood. One big problem is experimental basis. As in many human systems, it is difficult to perform controlled experiments on a sufficiently large scale. This would be necessary since data from actual emergency situations is usually not available, at least in sufficient quality. Progress should be possible by using modern video and computer technology which should allow us, in principle, to extract precise data even for the trajectories of individuals.

The full understanding of the complex dynamics of evacuation processes requires collaboration between engineering, physics, computer science, psychology, etc. Engineering in cooperation with computer science will lead to an improved empirical basis. Methods from physics allow us to develop simple but realistic models that capture the main aspects of the dynamics. Psychology is then needed to understand the interactions between individuals in sufficient detail to get a reliable set of ‘interaction’ parameters for the physical models.

In the end, we hope these joint efforts will lead to realistic models for evacuation processes that not only allow us to study these in the planning stages of facilities, but even allow for dynamical real-time evacuation control in case an emergency occurs.

### Acknowledgments

The authors would like to acknowledge the contribution of Tim Meyer-König (the developer of PedGo) and Michael Schreckenberg, Ansgar Kirchner, Bernhard Steffen for many fruitful discussions and valuable hints.

### Bibliography

#### Primary Literature

1. Abe K (1986) *The Science of Human Panic*. Brain, Tokyo (in Japanese)
2. AlGadhi SAH, Mahmassani HS, Herman R (2002) A speed-concentration relation for bi-directional crowd movements with strong interaction. In: Schreckenberg M, Sharma SD (eds) *Pedestrian and Evacuation Dynamics*. Springer, Berlin, pp 3–20
3. American Sociological Association (2002) In disasters, panic is rare, altruism dominates. Technical report, American Sociological Association
4. Ashe B, Shields TJ (1999) Analysis and modelling of the unannounced evacuation of a large retail store. *Fire Mater* 23: 333–336
5. Ben-Jacob E (1997) From snowflake formation to growth of bacterial colonies, Part II. Cooperative formation of complex colonial patterns. *Contemp Phys* 38:205
6. Biham O, Middleton AA, Levine D (1992) Self-organization and a dynamical transition in traffic-flow models. *Phys Rev A* 46:R6124
7. Blue VJ, Adler JL (2000) Cellular automata microsimulation of bi-directional pedestrian flows. *J Trans Res Board* 1678: 135–141
8. Blue VJ, Adler JL (2002) Flow capacities from cellular automata modeling of proportional spills of pedestrians by direction. In: Schreckenberg M, Sharma SD (eds) *Pedestrian and Evacuation Dynamics*. Springer, Berlin, pp 115–121
9. Blythe RA et al (2007) Nonequilibrium steady states of matrix product form: a solver’s guide. *Math Theor* 40:R333–R441, doi:10.1088/1751-8113/40/46/R01
10. Bolay K (1998) Nichtlineare Phänomene in einem fluid-dynamischen Verkehrsmodell. Diploma Thesis, Stuttgart University
11. Boyce KE, Shields TJ, Silcock GWH (1999) Toward the Characterization of Building Occupancies for Fire Safety Engineering: Capabilities of Disabled People Moving Horizontally and on an Incline. *Fire Technol* 35:51–67
12. Bryan JL (1995) Behavioral response to fire and smoke. In: DiNenno PJ, Beyler CL, Custer RLP, Walton WD, Watts JM, Drysdale D, Hall JR (eds) *SFPE Handbook of Fire Protection Engineering*, 2nd edn. National Fire Protection Association, Quincy, p 263
13. Burstedde C, Klauck K, Schadschneider A, Zittartz J (2001) Simulation of pedestrian dynamics using a two-dimensional cellular automaton. *Physica A* 295:507–525
14. Burstedde C, Kirchner A, Klauck K, Schadschneider A, Zittartz J (2002) Cellular automaton approach to pedestrian dynamics – applications. In: Schreckenberg M, Sharma SD (eds) *Pedestrian and Evacuation Dynamics*. Springer, Berlin, pp 87–98
15. Chakrabarti J, Dzubiella J, Löwen H (2004) Reentrance effect in the lane formation of driven colloids. *Phys Rev E* 70:012401
16. Chowdhury D, Santen L, Schadschneider A (2000) Statistical physics of vehicular traffic and some related systems. *Phys Rep* 329(4–6):199–329
17. Clarke L (2002) Panic: Myth or reality? *Contexts* 1(3):21–26
18. Coleman JS (1990) *Foundation of Social Theory*. Belknap, Cambridge, Chap 9
19. Daamen W (2004) *Modelling Passenger Flows in Public Transport Facilities*. Ph.D. thesis, Technical University of Delft
20. Daamen W, Hoogendoorn SP (2006) Flow-density relations for pedestrian traffic. In: Schadschneider A, Pöschel T, Kühne R, Schreckenberg M, Wolf DE (eds) *Traffic and Granular Flow 05*. Springer, Berlin, pp 315–322
21. Daamen W, Bovy PHL, Hoogendoorn SP (2002) Modelling pedestrians in transfer stations. In: Schreckenberg M, Sharma SD (eds) *Pedestrian and Evacuation Dynamics*. Springer, Berlin, pp 3–20



- SD (eds) *Pedestrian and Evacuation Dynamics*. Springer, Berlin, pp 59–73
22. de Gelder B, Snyder J, Greve D, Gerard G, Hadjikhani N (2004) Fear fosters flight: A mechanism for fear contagion when perceiving emotion expressed by a whole body. *Proc Natl Acad Sci* 101(47):16701–16706
  23. Derrida B (1998) An exactly soluble non-equilibrium system: The asymmetric simple exclusion process. *Phys Rep* 301:65
  24. Dieckmann D (1911) *Die Feuersicherheit in Theatern*. Jung, München (in German)
  25. DiNenno PJ (ed) (2002) *SFPE Handbook of Fire Protection Engineering*, 3rd edn. National Fire Protection Association, Bethesda
  26. DiNenno PJ, Beyler CL, Custer RLP, Walton WD, Watts JM, Drysdale D, Hall JR (eds) (1995) *SFPE Handbook of Fire Protection Engineering*, 2nd edn. National Fire Protection Association, Quincy
  27. Dogliani M (2002) An overview of present and under-development IMO's requirements concerning evacuation from ships. In: Schreckenberg M, Sharma SD (eds) *Pedestrian and Evacuation Dynamics*. Springer, Berlin, pp 339–354
  28. Dzubiella J, Hoffmann GP, Löwen H (2002) Lane formation in colloidal mixtures driven by an external field. *Phys Rev E* 65:021402
  29. Ehm M, Linxweiler J (2004) Berechnungen von Evakuierungszeiten bei Sonderbauten mit dem Programm buildingExodus. Technical report, TU Braunschweig
  30. El Yacoubi S, Chopard B, Bandini S (eds) (2006) *Cellular Automata – 7th International Conference on Cellular Automata for Research and Industry, ACRI 2006, Perpignan*. Springer, Berlin
  31. Federal Aviation Administration FAA (1990) Emergency evacuation – cfr sec. 25.803. Regulation CFR Sec. 25.803
  32. Fischer H (1933) *Über die Leistungsfähigkeit von Türen, Gängen und Treppen bei ruhigem, dichtem Verkehr*. Dissertation, Technische Hochschule Dresden (in German)
  33. Frantzych H (1996) Study of movement on stairs during evacuation using video analysing techniques. Technical report, Department of Fire Safety Engineering, Lund Institute of Technology
  34. Frisch U, Hasslacher B, Pomeau Y (1986) Lattice-gas automata for the Navier-Stokes equation. *Phys Rev Lett* 56:1505
  35. Fruin JJ (1971) *Pedestrian Planning and Design*. Metropolitan Association of Urban Designers and Environmental Planners, New York
  36. Fruin JJ (1993) The causes and prevention of crowd disasters. In: Smith RA, Dickie JF (eds) *Engineering for Crowd Safety*. Amsterdam, Elsevier
  37. Fujiyama T (2006) Collision avoidance of pedestrians on stairs. Technical report, Centre for Transport Studies. University College London, London
  38. Fujiyama T, Tyler N (2004) An explicit study on walking speeds of pedestrians on stairs. In: 10th International Conference on Mobility and Transport for Elderly and Disabled People, Hamamatsu, Japan, May 2004
  39. Fujiyama T, Tyler N (2004) Pedestrian Speeds on Stairs: An Initial Step for a Simulation Model. In: Proceedings of 36th Universities Transport Studies Group Conference, Life Science Centre, Newcastle upon Tyne, Jan 2004
  40. Fukui M, Ishibashi Y (1999) Jamming transition in cellular automaton models for pedestrians on passageway. *J Phys Soc Jpn* 68:3738
  41. Fukui M, Ishibashi Y (1999) Self-organized phase transitions in cellular automaton models for pedestrians. *J Phys Soc Jpn* 68:2861
  42. Galbreath M (1969) Time of evacuation by stairs in high buildings. *Fire Res Note* 8, NRCC
  43. Galea ER (2002) Simulating evacuation and circulation in planes, trains, buildings and ships using the EXODUS software. In: Schreckenberg M, Sharma SD (eds) *Pedestrian and Evacuation Dynamics*. Springer, Berlin pp 203–226
  44. Galea ER (ed) (2003) *Pedestrian and Evacuation Dynamics 2003*. CMS Press, London
  45. Gipps PG, Marksjö B (1985) A micro-simulation model for pedestrian flows. *Math Comput Simul* 27:95–105
  46. Graat E, Midden C, Bockholts P (1999) Complex evacuation; effects of motivation level and slope of stairs on emergency egress time in a sports stadium. *Saf Sci* 31:127–141
  47. Grosshandler W, Sunder S, Snell J (2003) Building and fire safety investigation of the world trade center disaster. In: Galea ER (ed) *Pedestrian and Evacuation Dynamics 2003*. CMS Press, London, pp 279–281
  48. Hamacher HW, Tjandra SA (2002) Mathematical modelling of evacuation problems – a state of the art. In: Schreckenberg M, Sharma SD (eds) *Pedestrian and Evacuation Dynamics*. Springer, Berlin, pp 227–266
  49. Hankin BD, Wright RA (1958) Passenger flow in subways. *Oper Res Q* 9:81–88
  50. Helbing D (1992) A fluid-dynamic model for the movement of pedestrians. *Complex Syst* 6:391–415
  51. Helbing D (2001) Traffic and related self-driven many-particle systems. *Rev Mod Phys* 73:1067–1141
  52. Helbing D, Molnár P (1995) Social force model for pedestrian dynamics. *Phys Rev E* 51:4282–4286
  53. Helbing D, Farkas I, Vicsek T (2000) Freezing by heating in a driven mesoscopic system. *Phys Rev Lett* 84:1240–1243
  54. Helbing D, Farkas I, Vicsek T (2000) Simulating dynamical features of escape panic. *Nature* 407:487–490
  55. Helbing D, Farkas I, Molnár P, Vicsek T (2002) Simulation of pedestrian crowds in normal and evacuation situations. In: Schreckenberg M, Sharma SD (eds) *Pedestrian and Evacuation Dynamics*. Springer, Berlin, pp 21–58
  56. Helbing D, Buzna L, Werner T (2003) Self-organized pedestrian crowd dynamics and design solutions. *Traffic Forum*, pp 2003-12
  57. Helbing D, Buzna L, Johansson A, Werner T (2005) Self-organized pedestrian crowd dynamics: Experiments, simulations, and design solutions. *Transp Sci* 39:1–24
  58. Helbing D, Johansson A, Al-Abideen HZ (2007) The dynamics of crowd disasters: An empirical study. *Phys Rev E* 75:046109
  59. Helbing D, Johansson A, Al-Abideen HZ (2007) Crowd turbulence: the physics of crowd disasters. In: The Fifth International Conference on Nonlinear Mechanics, ICNM-V, Shanghai, pp 967–969
  60. Henderson LF (1971) The statistics of crowd fluids. *Nature* 229:381–383
  61. Henderson LF (1974) On the fluid mechanics of human crowd motion. *Transp Res* 8:509–515
  62. Hoogendoorn SP (2003) Walker behaviour modelling by differential games. In: Emmerich H, Nestler B, Schreckenberg M

- (eds) Interface and transport dynamics. Lecture notes in Computational Science and Engineering, vol 32. Springer, Berlin, pp 275–294
63. Hoogendoorn SP, Bovy PHL (2003) Simulation of pedestrian flows by optimal control and differential games. *Optim Control Appl Meth* 24:153
  64. Hoogendoorn SP, Daamen W (2005) Pedestrian behavior at bottlenecks. *Transp Sci* 39 2:0147–0159
  65. Hoogendoorn SP, Bovy PHL, Daamen W (2002) Microscopic pedestrian wayfinding and dynamics modelling. In: Schreckenberg M, Sharma SD (eds) *Pedestrian and Evacuation Dynamics*. Springer, Berlin, pp 123–154
  66. Hoogendoorn SP, Daamen W, Bovy PHL (2003) Microscopic pedestrian traffic data collection and analysis by walking experiments: Behaviour at bottlenecks. In: Galea ER (ed) *Pedestrian and Evacuation Dynamics 2003*. CMS Press, London, pp 89–100
  67. Hoskin K (2004) Fire protection and evacuation procedures of stadia venues in new zealand. Master's thesis, University of Canterbury
  68. Hughes RL (2000) The flow of large crowds of pedestrians. *Math Comput Simul* 53:367–370
  69. Hughes RL (2002) A continuum theory for the flow of pedestrians. *Transp Res Part B* 36:507–535
  70. International Maritime Organization (IMO) (2000) International Code of Safety for High-Speed Craft, 2000 (2000 HSC Code). Technical report, Resolution MSC 97(73)
  71. International Organization for Standardization (2000) ISO-TR-13387-8-1999 Fire safety engineering – part 8: Life safety – occupant behaviour, location and condition. Technical report
  72. Jian L, Lizhong Y, Daoling Z (2005) Simulation of bi-direction pedestrian movement in corridor. *Physica A* 354:619
  73. Johnson NR (1987) Panic at “The Who Concert Stampede”: An Empirical Assessment. *Soc Probl* 34(4):362–373
  74. Jungermann H, Göhlert C (2000) Emergency evacuation from double-deck aircraft. In: Cottam MP, Harvey DW, Pape RP, Tait J (eds) *Foresight and Precaution*. Proceedings of ESREL 2000, SARS and SRA. Europe Annual Conference, Rotterdam, pp 989–992
  75. Kashiwagi T (ed) (1994) *Fire Safety Science – 4th international Symposium Proceedings*. Interscience, West Yard House, Guildford. The International Association for Fire Safety Science. Grove, London
  76. Kaufman M (2007) Lane Formation in Counterflow Situations of Pedestrian Traffic. Master's thesis, Universität Duisburg-Essen
  77. Keating JP (1982) The myth of panic. *Fire J* May:57–62
  78. Kendik E (1983) Determination of the evacuation time pertinent to the projected area factor in the event of total evacuation of high-rise office buildings via staircases. *Fire Saf J* 5:223–232
  79. Kendik E (1984) Die Berechnung der Personenströme als Grundlage für die Bemessung von Gehwegen in Gebäuden und um Gebäude. Ph.D. thesis, TU Wien
  80. Kendik E (1986) Designing escape routes in buildings. *Fire Technol* 22:272–294
  81. Kerner BS (2004) *The Physics of Traffic*. Springer, Heidelberg
  82. Kirchner A (2003) Modellierung und statistische Physik biologischer und sozialer Systeme. Dissertation, Universität zu Köln
  83. Kirchner A, Schadschneider A (2002) Simulation of evacuation processes using a bionics-inspired cellular automaton model for pedestrian dynamics. *Physica A* 312:260–276
  84. Kirchner A, Namazi A, Nishinari K, Schadschneider A (2003) Role of conflicts in the floor field cellular automaton model for pedestrian dynamics. In: Galea ER (ed) *Pedestrian and Evacuation Dynamics 2003*. CMS Press, London, pp 51
  85. Kirchner A, Nishinari K, Schadschneider A (2003) Friction effects and clogging in a cellular automaton model for pedestrian dynamics. *Phys Rev E* 67:056122
  86. Kirchner A, Klüpfel H, Nishinari K, Schadschneider A, Schreckenberg M (2004) Discretization effects and the influence of walking speed in cellular automata models for pedestrian dynamics. *J Stat Mech* 10:P10011
  87. Klüpfel H (2003) A Cellular Automaton Model for Crowd Movement and Egress Simulation. Dissertation, University Duisburg-Essen
  88. Klüpfel H (2006) The simulation of crowds at very large events. In: Schadschneider A, Pöschel T, Kühne R, Schreckenberg M, Wolf DE (eds) *Traffic and Granular Flow 05*. Springer, Berlin, pp 341–346
  89. Klüpfel H, Meyer-König T, Wahle J, Schreckenberg M (2000) Microscopic simulation of evacuation processes on passenger ships. In: Bandini S, Worsch T (eds) *Theory and Practical Issues on Cellular Automata*. Springer, Berlin
  90. Klüpfel H, Meyer-König T, Schreckenberg M (2001) Empirical data on an evacuation exercise in a movie theater. Technical report, University Duisburg-Essen
  91. Ko SY (2003) Comparison of evacuation times using Simulex and EvacuationNZ based on trial evacuations. Fire Engineering Research Report 03/9, University of Canterbury
  92. Kretz T (2007) Pedestrian Traffic – Simulation and Experiments. Dissertation, Universität Duisburg-Essen
  93. Kretz T, Grünebohm A, Schreckenberg M (2006) Experimental study of pedestrian flow through a bottleneck. *J Stat Mech* P10014
  94. Kretz T, Grünebohm A, Kaufmann M, Mazur F, Schreckenberg M (2006) Experimental study of pedestrian counterflow in a corridor. *J Stat Mech* P10001
  95. Kretz T, Grünebohm A, Keßel A, Klüpfel H, Meyer-König T, Schreckenberg M (2008) Upstairs walking speed distributions on a long stair. *Saf Sci* 46:72–78
  96. Kuligowski ED, Milke JA (2005) A performance-based egress analysis of a hotel building using two models. *J Fire Prot Eng* 15:287–305
  97. Kuligowski ED, Peacock RD (2005) A review of building evacuation models. Technical report 1471. National Institute of Standards and Technology, Gaithersburg
  98. Lakoba TI, Kaup DJ, Finkelstein NM (2005) Modifications of the Helbing-Molnár-Farkas-Vicsek social force model for pedestrian evolution. *Simulation* 81(5):339–352
  99. Lam WHK, Lee JYS, Chan KS, Goh PK (2003) A generalised function for modeling bi-directional flow effects on indoor walkways in Hong Kong. *Transp Res A: Policy Pract* 37:789–810
  100. Laur U, Jaakula H, Metsaveer J, Lehtola K, Livonen H, Karpinen T, Eksborg AL, Rosengren H, Noord O (1997) Final Report on the Capsizing on 28 September 1994 in the Baltic Sea of the Ro-Ro Passenger Vessel MV Estonia. Technical report. The Joint Accident Investigation Commission of Estonia, Finland and Sweden

101. LeBon G (1895) *Lois Psychologiques De L'évolution Des Peuples*. Alcan, Paris
102. Leutzbach W (1988) *Introduction to the Theory of Traffic Flow*. Springer, Berlin
103. Lewin K (1951) *Field Theory in Social Science*. Harper, New York
104. Lord J, Meacham B, Moore A, Fahy R, Proulx G (2005) *Guide for evaluating the predictive capabilities of computer egress models*. Technical report NIST GCR 06-886, NIST, Gaithersburg
105. Lovas GG (1994) Modeling and simulation of pedestrian traffic flow. *Transp Res B* 28V:429
106. Maniccam S (2003) Traffic jamming on hexagonal lattice. *Physica A* 321:653
107. Maniccam S (2005) Effects of back step and update rule on congestion of mobile objects. *Physica A* 346:631
108. Marconi S, Chopard B (2002) A multiparticle lattice gas automata model for a crowd. In: *Cellular Automata. Lecture Notes Computer Science*, vol 2493. Springer, Berlin, pp 231
109. Mawson AR (2005) Understanding mass panic and other collective responses to threat and disaster. *Psychiatry* 68:95-113
110. Melinek SJ, Booth S (1975) An analysis of evacuation times and the movement of crowds in buildings. Technical report CP 96/75, BRE
111. Meyer-König T, Klüpfel H, Schreckenberg M (2001) A microscopic model for simulating mustering and evacuation processes onboard passenger ships. In: KH Dräger (ed) *Proceedings of the International Emergency Management Society Conference. The International Emergency Management Society, Oslo*
112. Meyer-König T, Klüpfel H, Schreckenberg M (2002) Assessment and analysis of evacuation processes on passenger ships by microscopic simulation. In: Schreckenberg M, Sharma SD (eds) *Pedestrian and Evacuation Dynamics*. Springer, Berlin, pp 297-302
113. Mintz A (1951) Non-adaptive group behaviour. *J Abnorm Soc Psychol* 46:150-159
114. Molnár P (1995) *Modellierung und Simulation der Dynamik von Fußgängerströmen*. Dissertation, Universität Stuttgart
115. Mori M, Tsukaguchi H (1987) A new method for evaluation of level of service in pedestrian facilities. *Transp Res* 21A(3): 223-234
116. Morrall JF, Ratnayake LL, Seneviratne PN (1991) Comparison of CBD pedestrian characteristics in Canada and Sri Lanka. In: *Transportation Research Record 1294*. TRB, National Research Council, Washington DC, pp 57-61
117. MSA (1997) *Report on Exercise Invicta*. Technical report. Marine Safety Agency, Southampton
118. MSC-Circ.1033. *Interim guidelines for evacuation analyses for new and existing passenger ships*. Technical report, International Maritime Organization, Marine Safety Committee, London, June, 6th 2002. MSC/Circ. 1033
119. MSC-Circ.1166. *Guidelines for a simplified evacuation analysis for high-speed passenger craft*. Technical report, International Maritime Organisation, 2005
120. Muir HC (1997) *Airplane of the 21st century: Challenges in safety and survivability*. International Conference on Aviation Safety and Security in the 21st Century, White House Commission on Aviation Safety and Security, Washington
121. Muir HC, Bottomley DM, Marrison C (1996) Effects of motivation and cabin configuration on emergency aircraft evacuation behavior and rates of egress. *Int J Aviat Psychol* 6(1):57-77
122. Müller K (1981) *Zur Gestaltung und Bemessung von Fluchtwegen für die Evakuierung von Personen aus Bauwerken auf der Grundlage von Modellversuchen*. Dissertation, Technische Hochschule Magdeburg
123. Müller W (1966) Die Beurteilung von Treppen als Rückzugsweg in mehrgeschossigen Gebäuden. *Unser Brandschutz – Wissenschaftlich-Technische Beil* 3:65-70; to be continued in 4/1966
124. Müller W (1966) Die Beurteilung von Treppen als Rückzugsweg in mehrgeschossigen Gebäuden. *Unser Brandschutz – Wissenschaftlich-Technische Beil* 4:93-96; continuation from 3/1966
125. Müller W (1968) Die Überschneidung der Verkehrsströme bei dem Berechnen der Räumungszeit von Gebäuden. *Unser Brandschutz – Wissenschaftlich-Technische Beil* 4:87-92
126. Müller W (1970) *Untersuchung über zulässige Räumungszeiten und die Bemessung von Rückzugswegen in Gebäuden*. Habilitation, TU Dresden, Dresden
127. Muramatsu M, Nagatani T (2000) Jamming transition in two-dimensional pedestrian traffic. *Physica A* 275:281-291
128. Muramatsu M, Nagatani T (2000) Jamming transition of pedestrian traffic at crossing with open boundary conditions. *Physica A* 286:377-390
129. Muramatsu M, Irie T, Nagatani T (1999) Jamming transition in pedestrian counter flow. *Physica A* 267:487-498
130. Argebau (2005) *MVStättV – Erläuterungen: Musterverordnung über den Bau und Betrieb von Versammlungsstätten*. Erläuterungen, Juni 2005
131. Nagai R, Nagatani T (2006) Jamming transition in counter flow of slender particles on square lattice. *Physica A* 366:503
132. Nagai R, Fukamachi M, Nagatani T (2006) Evacuation of crawlers and walkers from corridor through an exit. *Physica A* 367:449-460
133. Nagel K, Schreckenberg M (1992) A cellular automaton model for freeway traffic. *J Phys I* 2:2221
134. Nakayama A, Hasebe K, Sugiyama Y (2005) Instability of pedestrian flow and phase structure in a two-dimensional optimal velocity model. *Phys Rev E* 71:036121
135. Navin PD, Wheeler RJ (1969) Pedestrian flow characteristics. *Traffic Eng* 39:31-36
136. Nelson HE, Mowrer FW (2002) Emergency movement. In: Di-Nenno PJ (ed) *SFPE Handbook of Fire Protection Engineering*, 3rd edn. National Fire Protection Association, Bethesda, p 367
137. National Fire Protection Association (2007) *NFPA 130: Standard for Fixed Guideway Transit and Passenger Rail Systems*.
138. Norwegian Ministry of Justice and Police (2000) *The High-Speed Craft MS Sleipner Disaster, 26 November 1999*. Official Norwegian Reports 2000:31, Oslo
139. Oeding D (1963) *Verkehrsbelastung und Dimensionierung von Gehwegen und anderen Anlagen des Fußgängerverkehrs*. Forschungsbericht, vol 22. Technische Hochschule Braunschweig
140. O'Flaherty CA, Parkinson MH (1972) Movement in a city centre footway. *Traffic Eng Control*, p 434
141. Okazaki S, Matsushita S (1993) A study of simulation model for pedestrian movement with evacuation and queuing. In: Smith RA, Dickie JF (eds) *Proceedings International Con-*

- ference Engineering Crowd Safety. Elsevier, Amsterdam, pp 271
142. Older SJ (1968) Movement of pedestrians on footways in shopping streets. *Traffic Eng Control* 10:160–163
  143. Owen M, Galea ER, Lawrence PJ, Filippidis L (1998) AASK – aircraft accident statistics and knowledge: a database of human experience in evacuation, derived from aviation accident reports. *Aero J* 102:353–363
  144. Pauls JL (1971) Evacuation drill held in the b. c. hydro building, 26 June 1969. Building Research Note 80, National Republican Congressional Committee
  145. Pauls JL (1995) Movement of people. In: DiNenno PJ, Beyler CL, Custer RLP, Walton WD, Watts JM, Drysdale D, Hall JR (eds) *SFPE Handbook of Fire Protection Engineering*, 2nd edn. National Fire Protection Association, Quincy, p 263
  146. Pauls JL, Fruin JJ, Zupan JM (2007) Minimum stair width for evacuation, overtaking movement and counterflow – technical bases and suggestions for the past, present and future. In: Waldau N, Gattermann P, Knoflacher H, Schreckenberg M (eds) *Pedestrian and Evacuation Dynamics 2005*. Springer, Berlin, pp 57–69
  147. Paulsen T, Soma H, Schneider V, Wiklund J, Lovas G (1995) Evaluation of simulation models of evacuating from complex spaces. SINTEF Report STF75 A95020. SINTEF, Trondheim
  148. Polus A, Joseph JL, Ushpiz A (1983) Pedestrian flow and level of service. *J Transp Eng* 109(1):46–56
  149. Popkov V, Schütz G (1999) Steady-state selection in driven diffusive systems with open boundaries. *Europhys Lett* 48:257
  150. Predtechenskii VM, Milinskii AI (1969) *Planing for foot traffic flow in buildings*. Amerind Publishing, New Dehli, 1978. Translation of: Proektirovanie Zhdaniis Uchetom Organizatsii Dvizheniya Lyudskikh Potokov, Stroiizdat Publishers, Moscow
  151. Predtetschenski WM, Milinski AI (1971) *Personenströme in Gebäuden – Berechnungsmethoden für die Modellierung*. Müller, Köln-Braunsfeld
  152. Predtetschenski WM, Cholstschewnikow WW, Völkel H (1972) Vereinfachte Berechnung der Umformung von Personenströmen auf Wegabschnitten mit begrenzter Länge. *Unser Brandschutz Wissenschaftlich-Technische Beil* 6:90–94
  153. Purser DA, Bensilium M (2001) Quantification of behaviour for engineering design standards and escape time calculations. *Saf Sci* 38(2):158–182
  154. Pushkarev B, Zupan JM (1975) Capacity of walkways. *Transp Res Rec* 538:1–15
  155. Quarantelli EL (1960) Images of withdrawal behavior in disasters: Some basic misconceptions. *Soc Probl* 8:63–79
  156. Quarantelli EL (2001) The sociology of panic. In: Smelser NJ, Baltes PB (eds) *International Encyclopedia of the Social and Behavioral Sciences*. Pergamon, New York, pp 11020–11030
  157. Revi A, Singh AK (2007) Cyclone and storm surge, pedestrian evacuation and emergency response in India. In: Waldau N, Gattermann P, Knoflacher H, Schreckenberg M (eds) *Pedestrian and Evacuation Dynamics 2005*. Springer, Berlin, pp 119–130
  158. Rex M, Löwen H (2007) Lane formation in oppositely charged colloids driven by an electric field: Chaining and two-dimensional crystallization. *Phys Rev E* 75:051402
  159. Rickert M, Nagel K, Schreckenberg M, Latour A (1996) Two lane traffic simulations using cellular automata. *Physica A* 231:534
  160. Rogsch C (2005) Vergleichende Untersuchungen zur dynamischen Simulation von Personenströmen. Technical report JUEL-4185. Forschungszentrum Jülich
  161. Rogsch C, Klingsch W, Seyfried A, Weigel H (2007) How reliable are commercial software-tools for evacuation calculation? In: *Interflam 2007 – Conference Proceedings*. Interscience Communication Ltd, Greenwich, London, pp 235–245
  162. Rogsch C, Klingsch W, Seyfried A, Weigel H (2007) Prediction accuracy of evacuation times for high-rise buildings and simple geometries by using different software-tools. In *Traffic and Granular Flow 2007*. Preprint
  163. Roitman MJ (1966) *Die Evakuierung von Menschen aus Bauwerken*. Staatsverlag der Deutschen Demokratischen Republik
  164. Rothman DH, Zaleski S (1994) Lattice-gas models of phase separation: Interfaces, phase transitions, and multiphase flow. *Rev Mod Phys* 66:1417
  165. Rothman DH, Zaleski S (1997) *Lattice-Gas Cellular Automata*. Cambridge University Press, Cambridge
  166. Saloma C, Perez GJ (2007) Herding in real escape panic. In: Waldau N, Gattermann P, Knoflacher H, Schreckenberg M (eds) *Pedestrian and Evacuation Dynamics 2005*. Springer, Berlin, pp 471–479
  167. Schadschneider A (2002) Cellular automaton approach to pedestrian dynamics – theory. In: Schreckenberg M, Sharma SD (eds) *Pedestrian and Evacuation Dynamics*. Springer, Berlin, pp 75–86
  168. Schelajew J, Schelajewa E, Semjonow N (2000) Nikolaus II. Der letzte russische Zar. Bechtermünz, Augsburg
  169. Schneider U, Kath K, Oswald M, Kirchberger H (2006) *Evakuierung und Verhalten von Personen im Brandfall unter spezieller Berücksichtigung von schienengebundenen Fahrzeugen*. Technical report 12, TU Wien
  170. Schreckenberg M, Sharma SD (eds) (2007) *Pedestrian and Evacuation Dynamics*. Springer, Berlin
  171. Schultz M, Lehmann S, Fricke H (2007) A discrete microscopic model for pedestrian dynamics to manage emergency situations in airport terminals. In: Waldau N, Gattermann P, Knoflacher H, Schreckenberg M (eds) *Pedestrian and Evacuation Dynamics 2005*. Springer, Berlin, pp 389–395
  172. Schütz GM (2001) Exactly solvable models for many-body systems. In: Domb C, Lebowitz JL (eds) *Phase Transitions and Critical Phenomena*, vol 19. Academic Press, Amsterdam
  173. Seeger PG, John R (1978) Untersuchung der Räumungsabläufe in Gebäuden als Grundlage für die Ausbildung von Rettungswegen, Teil III: Reale Räumungsversuche. Technical report T395. Forschungsstelle für Brandschutztechnik an der Universität Karlsruhe (TH)
  174. Seyfried A, Steffen B, Klingsch W, Boltes M (2005) The fundamental diagram of pedestrian movement revisited. *J Stat Mech* P10002
  175. Seyfried A, Steffen B, Lippert T (2006) Basics of modelling the pedestrian flow. *Physica A* 368:232–238
  176. Seyfried A, Rupperecht T, Passon O, Steffen B, Klingsch W, Boltes M (2007) Capacity estimation for emergency exits and bottlenecks. In: *Interflam 2007 – Conference Proceedings*. Interscience Communication Ltd, Greenwich, London
  177. Shestopal VO, Grubits SJ (1994) Evacuation model for merg-

- ing traffic flows in multi-room and multi-storey buildings. In: Kashiwagi T (ed) *Fire Safety Science – 4th international Symposium Proceedings*. Interscience, West Yard House, Guildford. The International Association for Fire Safety Science. Grove, London, pp 625–632
178. Sime JD (1990) *The Concept of Panic*. In: Canter D (ed) *Fires and Human Behaviour*, vol 1. Wiley, London, pp 63–81
  179. Smelser NJ (1962) *Theory of Collective Behavior*. Free Press, New York
  180. Still KG (2001) *Crowd Dynamics*, Ph.D. thesis, University of Warwick
  181. Tajima Y, Nagatani T (2002) Clogging transition of pedestrian flow in t-shaped channel. *Physica A* 303:239–250
  182. Taylor PM (1990) *The Hillsborough Stadium Disaster: Inquiry Final Report*. Technical report, Great Britain Home Office
  183. Templer J (1992) *The Staircase*. MIT Press, Cambridge
  184. Thompson PA, Marchant EW (1994) Simulex; developing new computer modelling techniques for evaluation. In: Kashiwagi T (ed) *Fire Safety Science – 4th international Symposium Proceedings*. Interscience, West Yard House, Guildford. The International Association for Fire Safety Science. Grove, London, pp 613–624
  185. Togawa K (1955) Study on fire escapes basing on the observation of multitude currents. Report of the building research institute. Ministry of Construction, Japan (in Japanese)
  186. Tsuji Y (2003) Numerical simulation of pedestrian flow at high densities. In: Galea ER (ed) *Pedestrian and Evacuation Dynamics 2003*. CMS Press, London, p 27
  187. Tubbs JS, Meacham B (2007) *Egress Design Solutions – A Guide to Evacuation and Crowd Management Planning*. Wiley, New Jersey
  188. Virkler MR, Elayadath S (1994) Pedestrian density characteristics and shockwaves. In: Akcelik R (ed) *Proceedings of the Second International Symposium on Highway Capacity*, vol 2. Australian Road Research Board, Sydney, pp 671–684
  189. Waldau N (2002) *Massenpanik in Gebäuden*. Diploma thesis, Technische Universität Wien
  190. Waldau N, Gattermann P, Knoflacher H, Schreckenberg M (eds) (2006) *Pedestrian and Evacuation Dynamics 2005*. Springer, Berlin
  191. Weckman LS, Mannikkö S (1999) Evacuation of a theatre: Exercise vs calculations. *Fire Mater* 23:357–361
  192. Weidmann U (1993) *Transporttechnik der Fußgänger – Transporttechnische Eigenschaften des Fußgängerverkehrs (Literaturauswertung)*. Schriftenreihe des IVT 90, ETH Zürich, 3 1993. Zweite, ergänzte Auflage (in German)
  193. Weifeng F, Lizhong Y, Weicheng F (2003) Simulation of bi-directional pedestrian movement using a cellular automata model. *Physica A* 321:633–640
  194. Wolf DE, Grassberger P (eds) (1996) *Friction, Arching, Contact Dynamics*. World Scientific, Singapore
  195. Yamamoto K, Kokubo S, Nishinari K (2006) New approach for pedestrian dynamics by real-coded cellular automata (rca). In: El Yacoubi S, Chopard B, Bandini S (eds) *Cellular Automata – 7th International Conference on Cellular Automata for Research and Industry, ACRI 2006, Perpignan*. Springer, Berlin, pp 728–731
  196. Yamamoto K, Kokubo S, Nishinari K (2007) Simulation for pedestrian dynamics by real-coded cellular automata (rca). *Physica A* 379:654
  197. Yamori K (1998) Going with the flow: Micro-macro dynamics in the macrobehavioral patterns of pedestrian crowds. *Psychol Rev* 105(3):530–557

### Books and Reviews

- Chopard B, Droz M (1998) *Cellular automaton modeling of physical systems*. Cambridge University Press, Cambridge
- Chowdhury D, Nishinari K, Santen L, Schadschneider A (2008) *Stochastic transport in complex systems: From molecules to vehicles*. Elsevier, Amsterdam
- DiNenno PJ (ed) (2002) *SFPE Handbook of Fire Protection Engineering*. National Fire Protection Association, Quincy
- Galea ER (ed) (2003) *Pedestrian and Evacuation Dynamics '03*. CMS Press, London
- Ped-Net collaboration. Webpage [www.ped-net.org](http://www.ped-net.org) (including discussion forum)
- Predtechenskii VM, Milinskii AI (1978) *Planing for foot traffic flow in buildings*. Amerint Publishing, New Delhi
- Schadschneider A, Pöschel T, Kühne R, Schreckenberg M, Wolf DE (eds) (2007) *Traffic and Granular Flow '05*. Springer, Berlin (see also previous issues of this conference series)
- Tubbs JS, Meacham BJ (2007) *Egress Design Solution – A Guide to Evacuation and Crowd Management Planning*. Wiley, New Jersey
- Waldau N, Gattermann P, Knoflacher H, Schreckenberg M (eds) (2007) *Pedestrian and Evacuation Dynamics '05*. Springer, Berlin

## Extreme Events in Socio-economic and Political Complex Systems, Predictability of

VLADIMIR KEILIS-BOROK<sup>1,2</sup>, ALEXANDRE SOLOVIEV<sup>2,3</sup>, ALLAN LICHTMAN<sup>4</sup>

<sup>1</sup> Institute of Geophysics and Planetary Physics and Department of Earth and Space Sciences, University of California, Los Angeles, USA

<sup>2</sup> International Institute of Earthquake Prediction Theory and Mathematical Geophysics, Russian Academy of Science, Moscow, Russia

<sup>3</sup> Abdus Salam International Centre for Theoretical Physics, Trieste, Italy

<sup>4</sup> American University, Washington D.C., USA

### Article Outline

Glossary

Definition of the Subject

Introduction

Common Elements of Data Analyzes

Elections

US Economic Recessions

Unemployment

Homicide Surges

Summary: Findings and Emerging Possibilities

Bibliography

### Glossary

**Complexity** A definitive feature of nonlinear systems of interacting elements. It comprises high instability with respect to initial and boundary conditions, and complex but non-random behavior patterns (“order in chaos”).

**Extreme events** Rare events having a large impact. Such events are also known as critical phenomena, disasters, catastrophes, and crises. They persistently reoccur in hierarchical complex systems created, separately or jointly, by nature and society.

**Fast acceleration of unemployment (FAU)** The start of a strong and lasting increase of the unemployment rate.

**Pattern recognition of rare events** The methodology of artificial intelligence’ kind aimed at studying distinctive features of complex phenomena, in particular – at formulating and testing hypotheses on these features.

**Premonitory patterns** Patterns of a complex system’s behavior that emerge most frequently as an extreme event approaches.

**Recession** The American National Bureau of Economic Research defines recession as “a significant decline in economic activity spread across the economy, lasting more than a few months”. A recession may involve simultaneous decline in coincident measures of overall economic activity such as industrial production, employment, investment, and corporate profits.

**Start of the homicide surge (SHS)** The start of a strong and lasting increase in the smoothed homicide rate.

### Definition of the Subject

At stake in the development of accurate and reliable methods of prediction for social systems is the capacity of scientific reason to improve the human condition. Today’s civilization is highly vulnerable to crises arising from extreme events generated by complex and poorly understood systems. Examples include external and civil wars, terrorist attacks, crime waves, economic downturns, and famines, to name just a few. Yet more subtle effects threaten modern society, such as the inability of democratic systems to produce policies responsive to challenges like climate change, global poverty, and resource depletion.

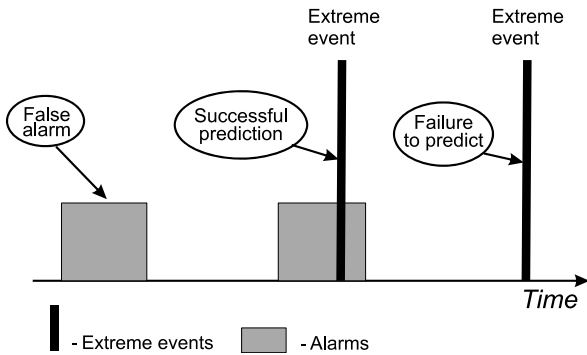
Our capacity to predict the course of events in complex social systems is inherently limited. However, there is a new and promising approach to predicting and understanding complex systems that has emerged through the integration of studies in the social sciences and the mathematics of prediction. This entry describes and analyzes that approach and its real-world applications. These include algorithmic prediction of electoral fortunes of incumbent parties, economic recessions, surges of unemployment, and outbursts of crimes. This leads to important inferences for averting and responding to impending crises and for improving the functioning of modern democratic societies.

That approach was successfully applied also to natural disasters such as earthquakes. Ultimately, improved prediction methods enhance our capacity for understanding the world and for protecting and sustaining our civilization.

**Extreme events.** Hierarchical complex systems persistently generate extreme events – the rare fast changes that have a strong impact on the system. Depending on connotation they are also known as critical phenomena, disasters, catastrophes, and crises. This article examines the development and application of the algorithmic prediction of extreme socio-economic and political events.

**The prediction problem** is formulated as follows:

*given* are time series that describe dynamics of the sys-



Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 1  
Possible outcomes of prediction

tem up to the current moment of time  $t$  and contain potential precursors of an extreme event;

to predict whether an extreme event will or will not occur during the subsequent time period  $(t, t + \tau)$ ; if the answer is “yes”, this will be the “*period of alarm*”.

As the time goes by, predictions form a discrete sequence of alarms. The possible outcomes of such a prediction are shown in Fig. 1. The actual outcome is determined unambiguously, since the extreme events are identified independently of the prediction either by the actual happening (e. g. by an election result) or by a separate algorithm (e. g. homicide surge) after they occur.

Such “yes or no” prediction is aimed not at analyzing the whole dynamics of the system, but only at identifying the occurrence of rare extreme events. In a broad field of prediction studies this prediction is different from and complementary to the classical Kolmogoroff–Wiener prediction of continuous functions, and to traditional cause-and-effect analysis.

The problem includes estimating the predictions’ accuracy: the rates of false alarms and failures to predict, and the total duration of alarms in relation to the total time considered. These characteristics represent the inevitable *probabilistic component* of prediction; they provide for statistical validation of a prediction algorithm and for optimizing preparedness to predicted events (e. g. recessions or crime surges).

**Twofold importance.** The prediction problem is pivotal in two areas:

- *Fundamental understanding of complex systems.* Prediction algorithms quantitatively define phenomena that anticipate extreme events. Such quantitative definition is pivotal for fundamental understanding of a complex system where these events occur, including

the intertwined mechanisms of system’s development and its basic features, e. g. multiple scaling, correlation range, clustering, fragmentation etc. (see Sects. “**Common Elements of Data Analyzes**”, “**Elections**”, “**US Economic Recessions**”, “**Unemployment**”). The understanding of complex systems remains a major unsolved problem of modern science, tantamount to transforming our understanding of the natural and human world.

- *Disaster preparedness.* On the practical side prediction is pivotal for coping with a variety of disasters, commonly recognized as major threats to the survival and sustainability of our civilization (e. g. [22]; see also materials of G8-UNESCO World Forum on “Education, Innovation and Research: New Partnership for Sustainable Development”, <http://g8forum.ictp.it>). The reliable advance prediction of extreme events can save lives, contribute to social and economic stability, and to improving the governing of modern societies.

## Introduction

### Predictability vs. Complexity: The Need for Holistic Approach [7,12,13,15,17,27,32]

Natural science had for many centuries regarded the Universe as a completely predictable machine. As Pierre Simon de Laplace wrote in 1776, “... if we knew exactly the laws of nature and the situation of the universe at the initial moment, we could predict exactly the situation of the same universe at a succeeding moment.” However, at the turn of the 20th century (1905) Jules Henry Poincare discovered, that “... this is not always so. It may happen that small differences in the initial conditions will produce very great ones in the final phenomena. Prediction becomes impossible”.

This instability to initial conditions is indeed a definitive attribute of complex systems. Nonetheless, through the robust integral description of such systems, it is possible to discover regular behavior patterns that transcend the inherent complexity. For that reason studying complexity requires the holistic approach that proceeds from the whole to details, as opposed to the reductionism approach that proceeds from details to the whole. It is in principle not possible “to understand a complex system by breaking it apart” [13].

Among the regular behavior patterns of complex systems are “premonitory” ones that emerge more frequently as an extreme event approaches. These premonitory patterns make complex systems predictable. The accuracy of predictions, however, is inevitably limited due to the systems’ complexity and observational errors.

Premonitory patterns and extreme events are consecutive manifestations of a system's dynamics. These patterns may not trigger extreme events but merely signal the growth of instability, making the system ripe for the emergence of extreme events.

### Methodology

The prediction algorithms described here are based on discovering premonitory patterns. The development of the algorithms requires the integration of complementary methods:

- Theoretical and numerical modeling of complex systems; this includes “universal” models considered in statistical physics and non-linear dynamics (e. g. [1,3,5,8,12,15,20,42]), and system-specific models, if available.
- Exploratory data analysis.
- Statistical analysis of limited samples, which is relevant since the prediction targets are by definition rare.
- Practical expertise, even if it is intuitive.
- Risk analysis and theory of optimal control for optimizing prediction strategy along with disaster preparedness.

**Pattern Recognition of Rare Events** This methodology provides an efficient framework for integrating diverse information into prediction algorithms [4,11,19]. This methodology has been developed by the artificial intelligence school of I. Gelfand for the study of rare phenomena of a highly complex origin. In terminology of pattern recognition, the “object of recognition” is the time moment  $t$ . The problem is to recognize whether it belongs to the period of alarm, i. e. to a time interval  $\Delta$  preceding an extreme event. An alarm starts when certain combinations of premonitory patterns emerges.

Several features of that methodology are important for predicting extreme events in the absence of a complete closed theory that would unambiguously define a prediction algorithm. First, this kind of pattern recognition relies on simple, robust parameters that overcome the bane of complexity analysis – incomplete knowledge of the system's causal mechanisms and chronic imperfections in the available data. In its efficient robustness, pattern recognition of rare events is akin to exploratory data analysis as developed by J. Tukey [50]. Second, unlike other statistical methods, e. g. regression analysis, that methodology can be used for small samples such as presidential elections or economic recessions. Also, it integrates quantitative and judgmental parameters and thereby more fully captures

the full dimensions of the prediction problem than procedures that rely strictly on quantitative variables.

Summing up, the methodology described here can help in prediction when there are (1) many causal variables, (2) qualitative knowledge about which variables are important, and (3) limited amounts of data [2].

Besides societal predictions, pattern recognition of rare events has been successfully applied in seismology and earthquake prediction (e. g. [11,19,20,44,46]), geological prospecting (e. g. [45]) and in many other fields. Review can be found in [21,47]. Tutorial materials are available at the web site of the Abdus Salam International Centre for Theoretical Physics ([http://cdsagenda5.ictp.it/full\\_display.php?da=a06219](http://cdsagenda5.ictp.it/full_display.php?da=a06219)).

**Validation of Prediction Algorithms** The algorithms include many adjustable elements, from selecting the data and defining the prediction targets, to specifying numerical parameters involved. In lieu of theory that would unambiguously determine these elements they have to be developed retrospectively, by “predicting” past extreme events. The application of the methodology to known events creates the danger of self-deceptive data-fitting: As J. von Neumann put it “*with four exponents I can fit an elephant*”. The proper validation of the prediction algorithms requires three consecutive tests.

- *Sensitivity analysis*: testing whether predictions are sensitive to variations of adjustable elements.
- *Out of sample analysis*: application of an algorithm to past data that has not been used in the algorithm's development. The test is considered successful if algorithm retains its accuracy.
- *Predicting future events* – the only decisive test of a prediction algorithm (see for example Sect. “Elections” below).

A highly efficient tool for such tests is the error Diagram, showing major characteristics of prediction accuracy [33,34,35,36,37,38,39]. Its example is given in Fig. 10. Exhaustive sets of these tests are described in [10,11,24,52].

### Common Elements of Data Analyses

The methodology discussed above was used for predicting various kinds of extreme events, as illustrated in the next four Sections. Naturally, from case to case this methodology was used in different ways, according to specifics of phenomena considered. However in all cases data analysis has essential common elements described below.

**Sequence of analysis** comprises four stages: (i) Defining prediction targets. (ii) Choosing the data (time series),



where premonitory patterns will be looked for and summing up a priori constrains on these patterns. (iii) Formulating hypothetical definition of these patterns and developing prediction algorithm; determining the error diagram. (iv) Validating and optimizing that algorithm.

**Preliminary transformation of raw data.** In predicting recessions (Sect. “US Economic Recessions”), fast acceleration of unemployment (Sect. “Unemployment”) and crime surges (Sect. “Homicide Surges”) raw data were time series of relevant monthly indicators, hypothetically containing premonitory patterns. Let  $f(m)$  be such an indicator, with integer  $m$  showing time in months. Premonitory behavior of some indicators is better captured by their linear trends.

Let  $W^f(l/q, p)$  be the local linear least-squares regression of a function  $f(m)$  within the sliding time window  $(q, p)$ :

$$W^f(l/q, p) = K^f(q, p)l + B^f(q, p), \quad q \leq l \leq p, \quad (1)$$

where integers  $l$ ,  $q$ , and  $p$  stand for time in months.

Premonitory behavior of most indicators was captured by the following two functions:

- The trend of  $f(m)$  in the  $s$  months long window,  $(m - s, m)$ . For brevity we denote

$$K^f(m/s) = K^f(m - s, m) \quad (2)$$

- The deviation of  $f(m)$  from extrapolation of its long-term regression (i. e. regression on a long time window  $(q, m - 1)$ ):

$$R^f(m/q) = f(m) - W^f(m/q, m - 1). \quad (3)$$

Both functions can be used for prediction since their values do not depend on the information about the future (after the month  $m$ ) which would be anathema in prediction.

**Discretization.** The prediction algorithms use one or several premonitory patterns. Each pattern is defined at the lowest – binary – level of resolution, as 0 or 1, distinguishing only the presence of absence of a pattern at each moment of time. Then the objects of recognition are described by binary vectors of the same length. This ensures the robustness of the prediction algorithms.

**Simple algorithm called Hamming distance** is used for classification of binary vectors in applications considered here, [14,20,28]. Each vector is either premonitory or not. Analyzing the samples of vectors of each class (“the learning material”), the algorithm determines a reference binary vector (“kernel”) with components typical for the premonitory vector. Let  $D$  be the Hamming distance of

a vector from the kernel (the number of non-coinciding binary components). The given vector is recognized as premonitory class, if  $D$  is below a certain threshold  $D^*$ . This criterion takes advantage of the clustering of precursors in time.

**Summing up**, these elements of the pattern recognition approach are common for its numerous applications, their diversity notwithstanding. Experience in the specific applications is described in Sects. “Elections”, “US Economic Recessions”, “Unemployment”, “Homicide Surges”. The conceptual summary of the accumulated experience is given in the final Sect. “Summary: Findings and Emerging Possibilities”.

## Elections

This Section describes algorithms for predicting the outcome of the US Presidential and mid-term Senatorial elections [28,29,30,31]. Elections’ time is set by the law as follows.

- National elections are held every even-numbered year, on the first Tuesday after the first Monday in November (i. e., between November 2 and November 8, inclusively).
- Presidential elections are held once every 4 years, i. e. on every other election day. People in each of the 50 states and District of Columbia are voting separately for “electors” pledged to one or another of the Presidential candidates. These electors make up the “Electoral College” which directly elects the President. Since 1860, when the present two-party system was basically established, the Electoral College reversed the decision of the popular vote only three times, in 1888, 1912, and 2000. Algorithmic prediction of such reversals is not developed so far.
- A third of Senators are elected for a 6-year term every election day; “mid-term” elections held in the middle of a Presidential term are considered here.

## Methodology

*The prediction target* is an electoral defeat of an “incumbent” party, i. e. the party holding the contested seat. Accordingly, the prediction problem is formulated as whether the incumbent party will retain this seat or lose it to the challenging party (*and not whether Republican or Democrat will win*). As is shown below, that formulation is crucial for predicting the outcomes of elections considered.

*Data.* The pre-election situation is described by robust common sense parameters defined at the lowest (binary)

level of resolution, as the *yes* or *no* answers to the questionnaires given below (Tables 1, 2). The questions are formulated in such a way that the answer *no* favors the victory of the challenging party. According to the Hamming distance analysis (Sect. “Common Elements of Data Analyzes”) the victory of the challenging party is predicted when the number of answers *no* exceeds a threshold  $D^*$ .

**Mid-term Senatorial Elections**

The prediction algorithm was developed by a retrospective analysis of the data on three elections, 1974, 1978, and 1982. The questionnaire is shown in Table 1. Victory of the challenger is predicted if the number of answers *no* is 5 or more [28,29,30].

The meaning of these questions may be broader than their literal interpretation. For example, financial contri-

butions (key 5 in Table 2) not only provide the resources required for an effective campaign, but may also constitute a poll in which the preferences are weighed by the money attached.

*Predicting future elections.* This algorithm (without any changes from year to year and from state to state) was applied in advance to the five subsequent elections, 1986–2002. Predictions are shown in Fig. 2. Altogether, 150 seats were put up for election. For each seat a separate prediction was made, 128 predictions were correct, and 22 – wrong.

*Statistical significance* of this score is 99.9%. In other words the probability to get such a score by chance is below 0.1% [28,29,30]. For some elections these predictions might be considered as trivial, since they coincide with prevailing expectation of experts. Such elections are identified by *Congressional Review*. Eliminating them from the score still results in 99% significance.

Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Table 1  
Questionnaire for mid-term Senatorial Elections [28]

1.	(Incumbency): The incumbent -party candidate is the sitting senator.
2.	(Stature): The incumbent -party candidate is a major national figure.
3.	(Contest): There was no serious contest for the incumbent -party nomination.
4.	(Party mandate): The incumbent party won the seat with 60% or more of the vote in the previous election.
5.	(Support): The incumbent -party candidate outspends the challenger by 10% or more.
6.	(Obscurity): The challenging -party candidate is not a major national figure or a past or present governor or member of Congress.
7.	(Opposition): The incumbent party is not the party of the President.
8.	(Contest): There is no serious contest for the challenging -party nomination (the nominee gains a majority of the votes cast in the first primary and beats the second-place finisher at least two to one).

Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Table 2  
Questionnaire for Presidential elections [29,30]

KEY 1	(Party Mandate): After the midterm elections, the incumbent party holds more seats in the US House of Representatives than it did after the previous midterm elections.
KEY 2	(Contest): There is no serious contest for the incumbent -party nomination.
KEY 3	(Incumbency): The incumbent -party candidate is the sitting president.
KEY 4	(Third party): There is significant third-party or independent campaign.
KEY 5	(Short-term economy): The economy is not in recession during the election campaign.
KEY 6	(Long-term economy): Real per -capita economic growth during the term equals or exceeds mean growth during the previous two terms.
KEY 7	(Policy change): The incumbent administration effects major changes in national policy.
KEY 8	(Social unrest): There is no sustained social unrest during the term.
KEY 9	(Scandal): The incumbent administration is unattained by a major scandal.
KEY 10	(Foreign/military failure): The incumbent administration suffers no major failure in foreign or military affairs.
KEY 11	(Foreign/military success): The incumbent administration achieves a major success in foreign or military affairs.
KEY 12	(Incumbent charisma): The incumbent -party candidate is charismatic or a national hero.
KEY 13	(Challenger charisma): The challenging -party candidate is not charismatic or a national hero.

## Presidential Elections

The prediction algorithm was developed by a retrospective analysis of the data on the past 31 elections, 1860–1980; that covers the period between victories of A. Lincoln and R. Reagan inclusively. The questionnaire is shown in Table 2. Victory for the challenger is predicted if the number of answers *no* is 6 or more [29,30].

*Predicting of future elections.* This algorithm (without any changes from year to year state) was applied in advance to the six subsequent elections, 1984–2004. Predictions are shown in Fig. 3. All of them happened to be correct. In 2000 the decision of popular majority was reversed by the Electoral College; such reversals are not targeted by this algorithm [29,30].

## Understanding Elections

*Collective behavior.* The finding that aggregate-level parameters can reliably anticipate the outcome of both presidential and senatorial elections points to an electoral behavior highly integrated not only for the nation as a whole but also within the diverse American states.

- A presidential election is determined by the collective, integrated estimation of performance of incumbent administration during the previous four years.
- In case of senatorial elections the electorate has more diffused expectations of performance but puts more importance on political experience and status than in the case of presidential elections. Senate incumbents, unlike presidential ones, do not suffer from a bad economy or benefit from a good one. (This suggests that rather than punishing the party holding a Senate seat for hard times, the voters may instead regard the incumbent party as a safe port in a storm).

*Similarity.* For each election year in all states the outcomes of elections follow the same pattern that transcends the diversities of the situations in each of the individual elections.

The same pattern of the choice of the US President prevails since 1860, i. e. since election of A Lincoln, despite all the overwhelming changes in the electorate, the economy, the social order and the technology of politics during these 130 years. (For example, the electorate of 1860 did not include the groups, which constitute 3/4 of present electorate, such as women, African Americans, or most of the citizens of the Latin American, South European, Eastern European, and Jewish descent [30].

*An alternative (and more traditional) concept* of American elections focuses on the division of voters into interest and attitudinal groups. By this concept the goal of

the contestants is to attract the maximum number of voting blocks with minimal antagonism from other blocks. Electoral choice depends strongly on the factors irrelevant to the essence of the electoral dilemma (e. g. on the campaign tactics). The drawbacks of this concept are discussed in [18,30]. In sum, the work on presidential and senatorial elections described above suggests the following new ways of understanding American politics and perhaps the politics of other societies as well.

1. Fundamental shifts in the composition of the electorate, the technology of campaigning, the prevailing economic and social conditions, and the key issues of campaigns do not necessarily change the pragmatic basis on which voters choose their leaders.
2. It is governing not campaigning that counts in the outcomes of presidential elections.
3. Different factors may decide the outcome of executive as compared to legislative elections.
4. Conventional campaigning will not improve the prospects for candidates faced with an unfavorable combination of fundamental historical factors. Disadvantaged candidates have an incentive to adopt innovative campaigns that break the pattern of conventional politics.
5. All candidates would benefit from using campaigns to build a foundation for governing in the future.

## US Economic Recessions

US National Bureau of Economic Research (NBER) has identified the seven recessions that occurred in the US since 1960 (Table 3). The starting points of a recession and of the recovery from it follow the months marked by a peak and a trough of economic activity, respectively.

A peak indicates the last month before a recession, and a trough – the last month of a recession.

*Prediction targets* considered are the first month after the peak and after the trough (“the turns to the worst and to the best”, respectively). The start of the first recession, in 1960, is not among the targets, since the data do not cover a sufficient period of time preceding the recession.

*The data* used for prediction comprise the following six monthly leading economic indicators obtained from the CITIBASE data base, Jan. 1960–June 2000 (abbreviations are the same, as in [49]).

**G10FF = FYGT10 – FEDFUN** Difference between the annual interest rate on 10 year US Treasury bonds, and federal fund annual interest rate.

0	1	2	3	4	5	6	7
			OK98				
			CO98				
			FL98				
			GA98				
			HA98	TN02			
			ID98	SC02			
			MA98	NC02			
			ND98	NE02			
			PN98	KY02			
			SD98	IA02			
			UT98	CO02			
			FL94	AL02			
			HA94	AK98			
			IN94	CA98			
			MT94	CT98			
			NB94	NE98			
			NJ94	OR98			
			TX94	SC98			
			WA94	VT98			
		AS98	WV94	WA98			
		KA98	WI94	CT94			
		LA98	AK90	MD94			
		MI98	IN90	NV94			
		NH98	KN90	WY94			
		MS94	ME90	CO90			
	AL98	NM94	MA90	HA90			
	AZ98	ND94	MT90	KY90			
	IO98	RI94	NB90	MI90			
	DL94	VT94	NC90	AZ86			
	MA94	AS90	TX90	CO86			
	NY94	IO90	WY90	ID86			
	AL90	MS90	AR86	LA86			
	DE90	NM90	CA86	NY86			
	IL90	OR90	IL86	OK86	WI98	MN94	
	LA90	RI90	IN86	WI86	CA94	MO94	
	OK90	SD90	IA86	NC86	ID90	VA94	
	SC90	VA90	NH86	WA86	PA86	NH90	
	TN90	WV90	OR86	MN90	IL98	IN98	
	HI86	AK86	VT86	OK94	ME94	OH98	
	OH86	CT86	TN94	PA94	AL86	MI94	
UT94	SC86	KS86	TX02	TN294	FL86	MD86	KY98
GA90	UT86	KY86	OK02	NC98	GA86	NV86	AZ94
NJ90	NH02	ND86	NJ02	NY98	MO86	SD86	OH94
0	1	2	3	4	5	6	7

OK98 – incumbent won, KY98 – challenger won, errors are highlighted.

Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 2

Made-in-advance predictions of the mid-term senatorial elections (1986–2002). Each election is represented by the two-letter state abbreviation with the election year shown by two last digits. Each column shows elections with certain number *D* of answers “no” to the questionnaire given in Table 1 (such answers are favorable to challenging party). Value of *D*, indicated at the top, is the Hamming distance from the kernel

D (number of answers NO)	0	1	2	3	4	5	6	7	8	9
Predictions published months in advance			1984	1988	2004	2000* 1996	1992			
Learning				1964 1928 1916 1908					1980 1976 1968 1952 1932	
		1956	1944 1940	1900 1872	1972 1924	1948	1912*	188 4	1920	1960
	1904	1936	1868	1864	1880	1888*	1892	186 0	1896	1876*

1904 years when incumbent won popular vote  
 1892 years when challenger won popular vote  
 \* years when popular vote was reversed by electoral vote

Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 3  
 Division of presidential elections (1860–2004) by the number *D* of answers “no” to the questionnaire given in Table 2 (such answers are favorable to challenging party). *D* is the Hamming distance from the kernel

Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Table 3  
 American Economic Recessions since 1960

#	Peaks	Troughs
1	1960:04	1961:02
2	1969:12	1970:11
3	1973:11	1975:03
4	1980:01	1980:07
5	1981:07	1982:11
6	1990:07	1991:03
7	2001:03	2001:11

- IP** Industrial Production, total: index of real (constant dollars, dimensionless) output in the entire economy. This represents mainly the manufacturing industry, because of the difficulties in measuring the quantity of the output in services (such as travel agents, banking, etc.).
- LHELL** Index of “help wanted” advertising. This is put together by a private publishing company that measures the amount of job advertising (column-inches) in a number of major newspapers.
- LUINC** Average weekly number of people claiming unemployment insurance.
- INVMTQ** Total inventories in manufacturing and trade, in real dollars. Includes intermediate inventories (for example held by manufacturers, ready to be sent to retailers) and final goods inventories (goods on the shelves in stores).

**FYGM3** Interest rate on 90 day US treasury bills at an annual rate (in percent).

These indicators were already known [48,49], as those that correlate with a recession’s approach.

**Prediction of a Recession Start**

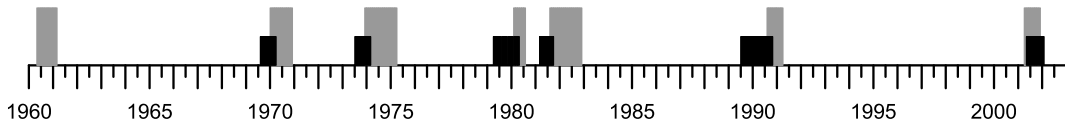
*Single indicators* exhibit the following premonitory patterns:

- G10FF:** small value
- IP and INVMTQ:** small deviation from the long-term trend  $R^f$  (3)
- FYGM3:** large deviation from the long-term trend  $R^f$
- LHELL:** small trend  $K^f$  (2)
- LUINC:** large trend  $K^f$

The prediction algorithm triggers an alarm after a month when most of the patterns emerge simultaneously. It lasts  $\Delta$  months and can be extended by the same rule, if premonitory patterns keep emerging. Formal quantitative definition of the algorithm can be found in [23] along with its validation by sensitivity and out-of-sample analyzes.

Alarms and recessions are juxtaposed in Fig. 4. We see that five recessions occurring between 1961 and 2000 were predicted by an alarm. The sixth recession started in April 2001, one month before the corresponding alarm. (Recession of 1960 was not considered for prediction, since data analyzed start just before it.)

Only the first six recessions listed in Table 1 were considered in the developing of the algorithm [23]. Duration



Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 4  
Alarms (black bars) and recessions (gray bars)

of each alarm was between 1 and 14 months. Total duration of all alarms was 38 months, or 13.6% of the time interval considered. There were no false alarms. No alarms were yielded so far by subsequent prediction in advance and no recession was identified during that time.

**Prediction of a Recession End**

Prediction targets are the starting points of recovery from recessions; these points are indicated in the last column of Table 3.

The data comprise the same six indicators that indicate the approach of a recession (see Subject. "Prediction of a Recession Start"); they are analyzed only within the recessions' periods.

Data analysis shows intriguing regularity illustrated in Fig. 5:

- Financial indicators change in opposite directions before the recession and before the recovery.
- Economic indicators change in the same direction before the recession and the recovery; but the change is stronger before the recovery, i. e., the economic situation worsens.

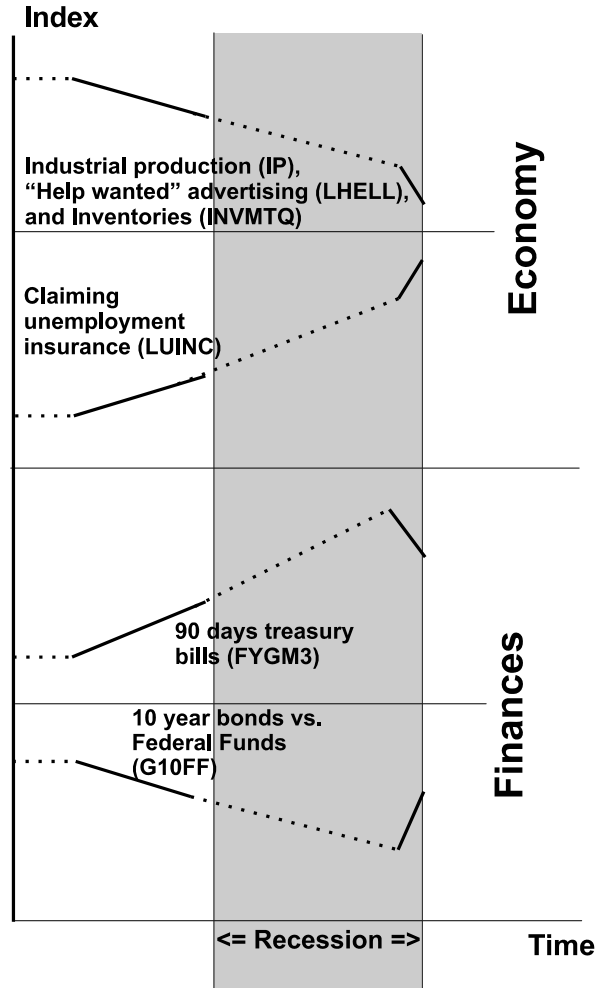
Prediction algorithm is formulated in the same terms as in the previous case but an alarm is triggered after three consecutive months when most of the patterns emerge simultaneously. The alarms predict when the recovery will start. Alarms and prediction targets are juxtaposed in Fig. 6. Duration of a single alarm is one to five months. Total duration of alarms is 16 months, which is 22% of time covered by all recessions. There are neither false alarms nor failures to predict.

**Unemployment**

Here we describe uniform prediction of the sharp and lasting unemployment surge in France, Germany, Italy, and the USA [25].

**Prediction Target**

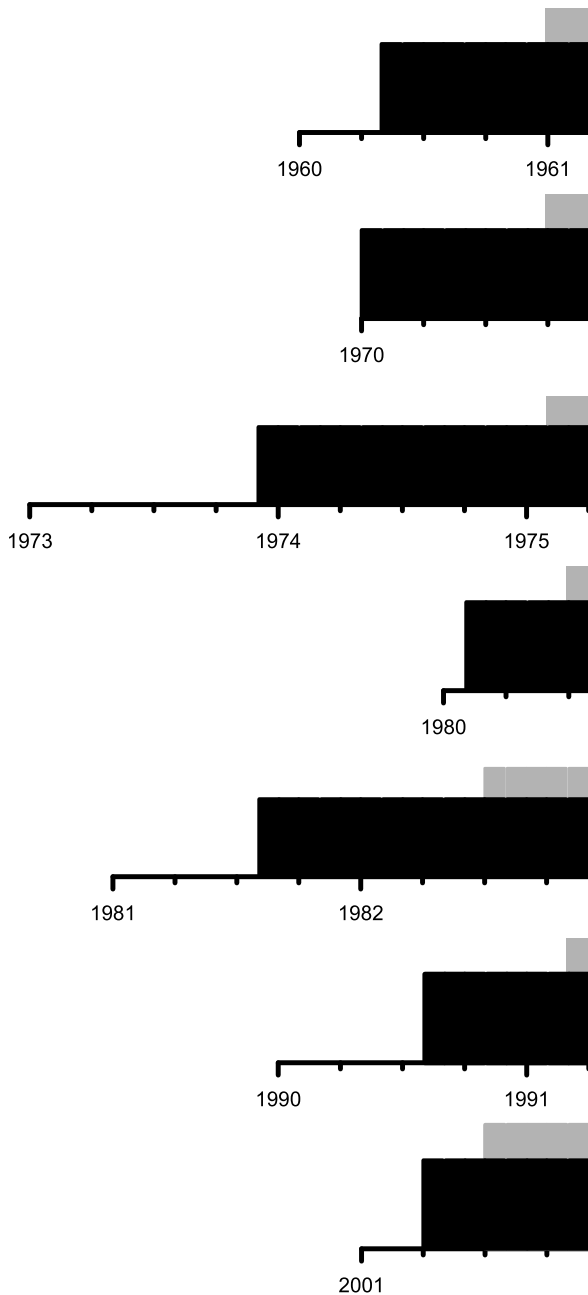
A prediction target is schematically illustrated in Fig. 7. Thin curve shows monthly unemployment with seasonal



Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 5  
Premonitory changes of indicators before the start of a recession and before its end. See explanations in the text

variations. On the thick curve seasonal variations are smoothed away. The arrow indicates a sharp upward bend of the smoothed curve. The moment of that bend is the prediction target. It is called by the acronym FAU, for "Fast Acceleration of Unemployment".

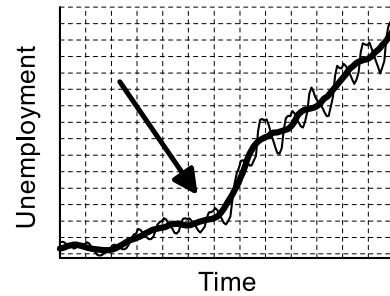
Smoothing was done as follows: Let  $u(m)$  be number of unemployed in a month  $m = 1, 2, \dots$ . After smooth-



Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 6

Prediction of recovery from a recession. *Black bars* – periods of recessions. *Gray bars* – alarms preceding the end of a recession

ing out the seasonal variation we obtain time series  $U(m) = W^u(m/m - 6, m + 6)$ ; this is the linear regression over the year-long time interval  $(m - 6, m + 6)$ . A natural robust measure of unemployment acceleration at the time  $m$



Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 7

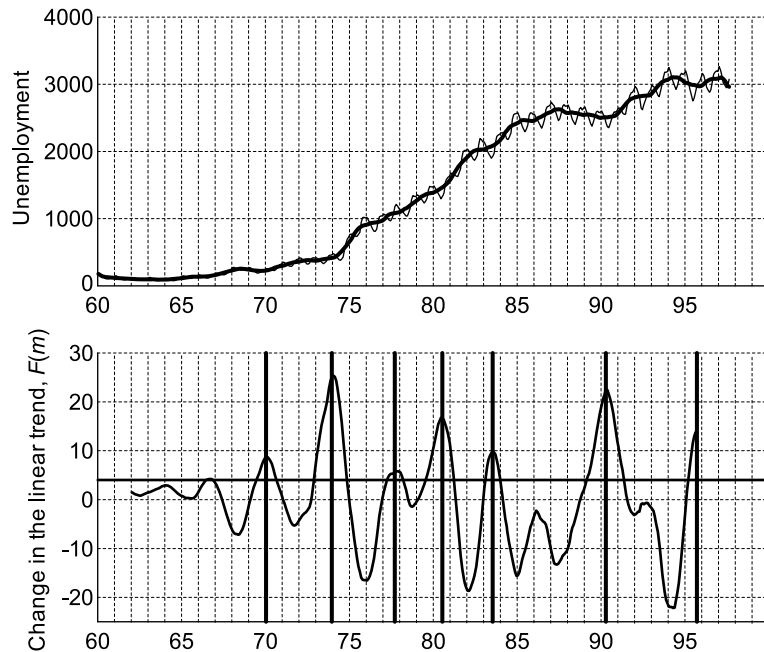
Fast acceleration of unemployment (FAU): schematic definition. *Thin line* – monthly unemployment; with seasonal variations. *Thick line* – monthly unemployment, with seasonal variations smoothed away. The *arrow* indicates a FAU – the sharp bend of the smoothed curve. The moment of a FAU is the target of prediction

is the bend of the linear trend of  $U$ ; in notations used in (1) this is the function  $F(m/s) = K^U(m + s, m) - K^U(m, m - s)$ . The FAUs are identified by the local maxima of  $F(m)$  exceeding a certain threshold  $F$ . The time  $m^*$  and the height  $F^*$  of such a maximum are, respectively, the time and the magnitude of a FAU. Subsequent local minimum of  $F(m)$  identifies the month  $m_e$  when acceleration ends. Figure 8 shows thus defined FAUs for France.

### The Data

The analysis has been initially made for France and three groups of data have been analyzed.

- *Composite macroeconomic indicators of national economy*
  1. **IP**: Industrial production indicator, composed of weighted production levels in numerous sectors of the economy, in % relative to the index for 1990.
  2. **L**: Long-term interest rate on 10-year government bonds, in %.
  3. **S**: Short-term interest rate on 3-month bills, in %.
- *Characteristics of more narrow areas of French economy*
  4. **NC**: The number of new passenger car registrations, in thousands of units.
  5. **EI**: Expected prospects for the national industrial sector.
  6. **EP**: Expected prospects for manufacturers.
  7. **EO**: Estimated volume of current orders.
- *Indicators related to US economy.*



Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 8

Unemployment in France. *Top*: Monthly unemployment, thousands of people. *Thin line*:  $u(m)$ , data from the OECD database; note the seasonal variations. *Thick line*:  $U(m)$ , data smoothed over one year. *Bottom*: Determination of FAUs.  $F(m)$  shows the change in the linear trend of unemployment  $U(m)$ . FAUs are attributed to the local maxima of  $F(m)$  exceeding threshold  $F = 4.0$  shown by horizontal line. The thick vertical lines show moments of the FAUs

8. **FF/\$**: Value of US dollar in French francs.
9. **AR**: The state of the American economy: is it close to a recession or not? This indicator shows the presence or absence of a current pre-recession alarm (see Subsect. “**Prediction of a Recession Start**”).

The data bases with above indicators for Europe are issued by the Organization for Economic Cooperation and Development [43] and the International Monetary Fund [16].

American analogues of indicators **IP**, **L**, and **S** are provided by CITIBASE; they are described in Sect. “**US Economic Recessions**” under abbreviations **IP**, **FYGM3** and **FIGT10** respectively.

### Prediction

Single indicators exhibit the following premonitory behavior.

- Steep upward trends of composite indicators (#1–#3). This behavior reflects “overheating” of the economy and may sound counterintuitive for industrial production (#1), since the rise of production is supposed to create more jobs. However, a particularly steep rise may create oversupply.

- Steep downward trends of economic expectations by general public (#4) and business community (#5–#8).
- Proximity of an American recession (#9). Before analysis was made such and opposite precursors might be expected for equally plausible reasons, so that this finding, if further confirmed, does provide a constraint on understanding unemployment’s dynamics.

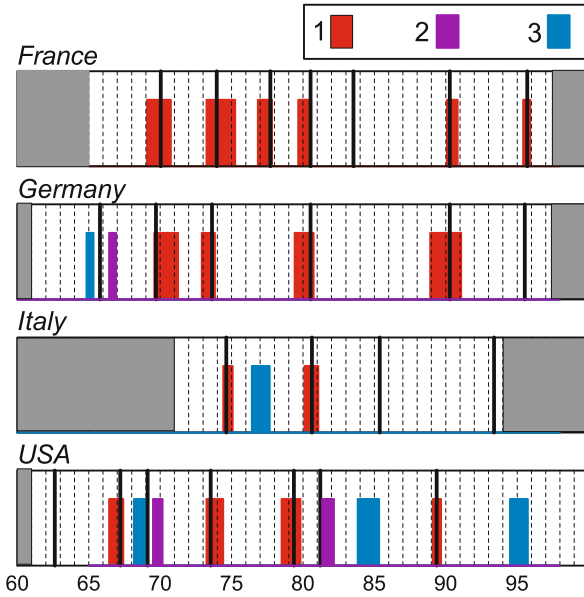
Among different combinations of indicators the macroeconomic ones (#1–#3) jointly give relatively better predictions, with smallest rates of errors and highest stability in sensitivity tests.

*Retrospective prediction.* Macroeconomic indicators were used jointly in the Hamming distance prediction algorithm (Sect. “**Common Elements of Data Analyzes**”). Being robust and self-adjusting to regional conditions, this algorithm was applied without any changes to the four countries considered here.

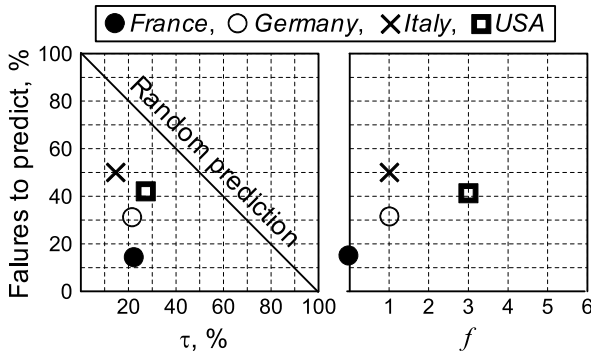
Alarms and FAUs are juxtaposed in Fig. 9. Error diagram in Fig. 10 shows quality of prediction for different countries. For US the quality is lower than for European countries, though still higher than in random predictions.

*Prediction of the future FAUs* was launched for USA. The results are shown in Fig. 11. It shows that by Jan-





Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 9  
 Retrospective predictions for four countries: FAUs and alarms obtained by the prediction algorithm. The thick vertical lines show the moments of FAUs in a country. Bars – the alarms starting shortly after FAUs within the periods of unemployment surge, 3 – false alarms. Shaded areas on both sides indicate the times, for which data on economic indicators were unavailable



Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 10  
 Error diagram for prediction of FAUs in different countries;  $\tau$  is total duration of alarms in % to the time interval considered,  $f$  – total number of false alarms

uary 2008 two correct predictions have been made, without other false alarms or failures to predict. In November 2006 the second prediction was filed on the web site of the Anderson School of Management, University of California, Los Angeles (<http://www.uclaforecast.com/>). This

started the documented experiment in testing the algorithm by predicting future FAUs on that website.

### Homicide Surges

This section analyzes the prediction of homicide rates in an American megacity – Los Angeles, CA [24].

### Prediction Target

A prediction target is the start of a sharp and lasting acceleration of the homicide rate; it is called by the acronym SHS, for “Start of the Homicide Surge.” It is formally determined by the analysis of monthly homicides rates, with seasonal variations smoothed out, as described in Subject. “Prediction Target”. Prediction targets thus identified are shown by vertical lines in Figs. 12 and 14 below.

### The Data

The analyzed data include monthly rates of the homicides and 11 types of lesser crimes, listed in Table 2. Definitions of these crimes are given in [6].

The data are taken from two sources:

- The National Archive of Criminal Justice Data, placed on the web site (NACJD), 1975–1993.
- Data bank of the Los Angeles Police Department (LAPD) Information Technology Division), 1990–2003.

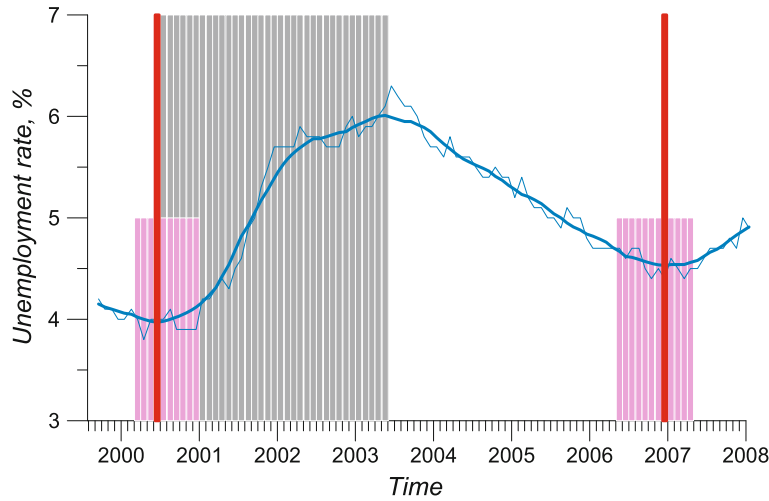
The algorithm does not use socio-economic determinants of crime, or other data that might be also useful. The objective was to develop a simple, efficient prediction model; development of comprehensive causal model would be a complementary objective.

### Prediction

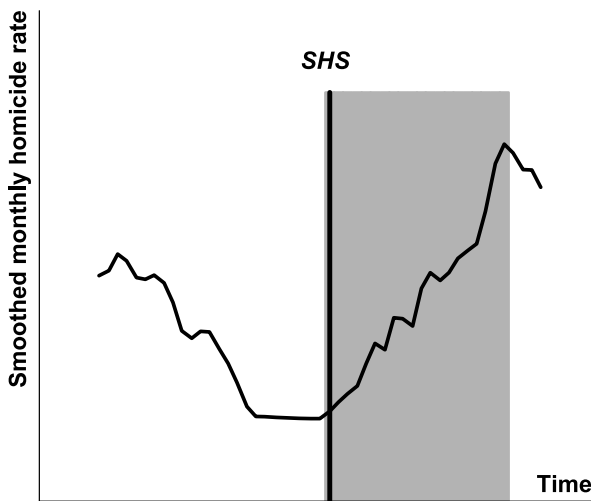
Premonitory behavior of indicators is illustrated in Fig. 13. The first phase is characterized by an escalation of burglaries and assaults, but not of robberies. Later on, closer to a homicide surge, robberies also increase.

The Prediction algorithm based on Hamming distance (see Sect. “Common Elements of Data Analyzes”) uses seven indicators listed in Table 4. Other five indicators marked by \* are used in sensitivity tests; and the homicide rate is used for identification of targets SHS.

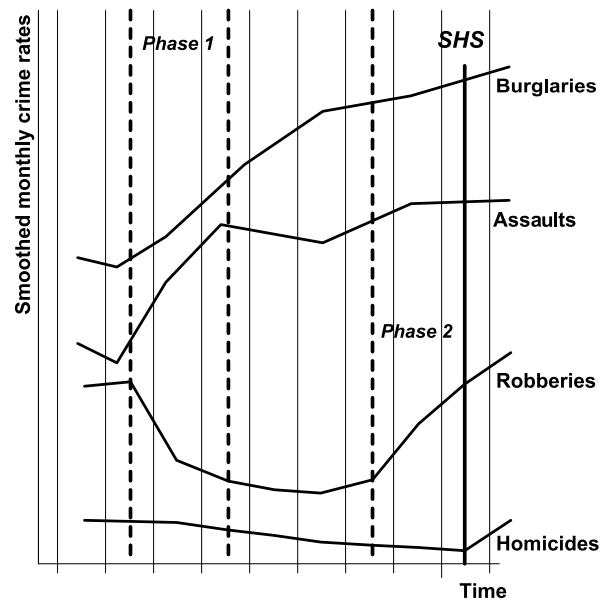
Alarms and homicide surges are juxtaposed in Fig. 14. The SHS episode in November 1994 has occurred simultaneously with the corresponding alarm. It is captured by an alarm, which starts in the month of SHS without a lead time. Prediction missed the October 1999 episode: it occurred two months before the start of the corresponding



Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 11  
 Experiment in predicting future FAUs, September (1999)–January (2008). *Thin blue curve* shows monthly unemployment rate in USA, according to data of Bureau of Labor Statistics, US Department of Labor (<http://www.data.bls.gov>). *Thick curve* shows this rate with seasonal variation smoothed away. *Vertical red lines* show prediction targets – the moments of FAU, *gray bar* – the period of unemployment’s growth; *pink bars* – periods of alarms



Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 12  
 Target of prediction – the Start of the Homicide Surge (“SHS”); schematic definition. *Gray bar* marks the period of homicide surge



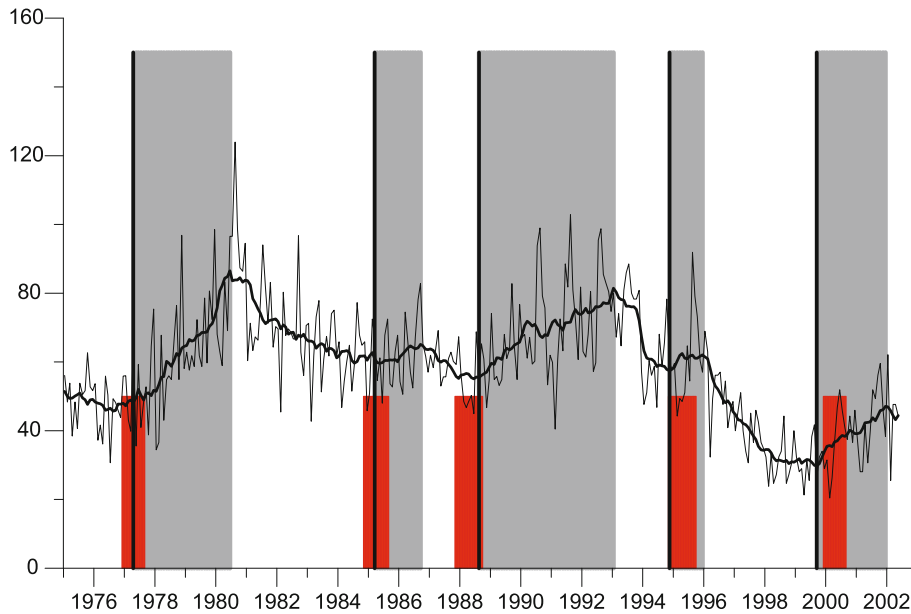
Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 13  
 Scheme of premonitory changes in crime statistics

alarm. Such delays should be taken into account for validating the algorithm. Note, however, that the last prediction did remain informative.

Altogether alarms occupy 15% of the time considered. During phase 2 (as defined in Fig. 13) this rate might be reduced [24].

**Summary: Findings and Emerging Possibilities**

The findings described above enhance predictive understanding of extreme events and indicate yet untapped possibilities for further R&D in that field.



Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 14  
 Performance of prediction algorithm through 1975–2002. *Thin curve* – original time series, total monthly number of homicides in Los Angeles city, per 3,000,000 inhabitants. Data from NACJD [6] have been used for 1975–1993 and from the Data Bank of the Los Angeles Police Department (LAPD Information Technology Division) for subsequent 9 years. *Thick curve* – smoothed series, with seasonal variations eliminated. *Vertical lines* show the targets of prediction – episodes of SHS (Subsect. “Prediction Target”). *Gray bars* show the periods of homicide surge. *Red bars* show the alarms declared by the prediction algorithm [24]

Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Table 4  
 Types of crimes considered (after [6])

Homicide	Robberies	Assaults	Burglaries
● All	● All	● All*	● Unlawful not forcible entry
	● With firearms	● With firearms	● Attempted forcible entry*
	● With knife or cutting instrument	● With knife or cutting instrument	
	● With other dangerous weapon	● With other dangerous weapon*	
	● Strong-arm robberies*	● Aggravated injury assaults*	

\*Analyzed in sensitivity tests only

### Pattern Recognition Approach

Information extracted from the already available data is indeed increased by this approach. To each problem considered here one may apply the following conclusion of J. Stock, a leading expert in the field: “Prediction/of recessions/requires fitting non-linear, high-dimensional models to a handful of observations generated by a possibly non-stationary economic environment... The evidence presented here suggests that these simple binary transformations of economic indicators have significant predictive content for recessions. It is striking that these models, in which the information in the data is reduced to binary indicators, have predictive contents comparable to or, in

many cases, better than that of more conventional models.” Importantly, this is achieved by using not more detailed data and models, but more robust aggregation (Subsect. “Predictability vs. Complexity: The Need for Holistic Approach”).

Partial “universality” of premonitory patterns is established by broad research in modeling and data analysis. This includes the common definition of the patterns, their self-adjustment, scaling, and similarity [9,10,20,26,42]; see also references in Sects. “Elections”, “US Economic Recessions”, “Unemployment”, “Homicide Surges”).

Relation to “cause and effect” analysis (perpetrators or witnesses?). Premonitory patterns might be either “perpetrators” contributing to causing extreme events, or the

“witnesses” – parallel manifestations of the system’s development. The cause that triggered a specific extreme event is usually identified, at least in retrospect. It may be, for example, a certain governmental decision, a change in the international situation, a natural disaster, the depletion of natural resources etc. However an actual extreme event might materialize only if the system is destabilized and “ripe” for it. Patterns of each kind signal such a ripe situation.

*What premonitory patterns to use for prediction?* Existing theories and experience reduce the number of such patterns, but too many of them remain hypothetically promising and have to be chosen by a trial and error procedure. Inevitably a prediction algorithm begins with a limited number of promising patterns. They should be sufficient for prediction, but other patterns may be equally or more useful and should be considered in further development of the algorithm. Most relevant “perpetrators” might not be included in the most useful patterns (e.g. due to their sensitivity to too many factors).

*Relation to policy-making: prediction and disaster preparedness.* Reliable predictions of future extreme events in complex societal systems would allow policy-makers to take remedial action before rather than after the onset of such afflictions as economic disasters, crime surges, etc. As in case of military intelligence predictions would be useful if their accuracy is known, albeit not necessarily high. Analysis of error diagrams allows to regulate the tradeoff between the rates of failures to predict and false alarms according to the needs of a decision-maker.

*Relation to governing and campaigning.* The findings presented here for the USA elections show that top elected officials would have better chances for reelection, if they focus on effective governing, and not on rhetoric, packaging and image-making. Candidates will benefit themselves and their parties if they run substantive campaigns that build a foundation for governing during the next term.

### Further Possibilities

**A wealth of yet untapped data and models** is readily available for the continuation of the kinds of studies described and analyzed in this article. Following are some immediate possibilities; specific examples can be found in the given references.

- *Continuing experiments in advance prediction*, for which the above findings set up a base (Sect. “**Elections**”). Successes and errors are equally important [37,38].
- *Incorporating other available data into the analysis* (Sects. “**US Economic Recessions**”, “**Unemployment**”)
  - *Predicting the same kind of extreme events in different contexts* (Sect. “**Unemployment**”)
  - *Predicting the end of a crisis* (Sect. “**US Economic Recessions**”).
  - *Multistage prediction with several lead times* (Sect. “**Homicide Surges**”)
    - Less imminent, but within reach are:
      - “*Universal*” scenarios of extreme development and low-parametric definition of an ensemble of premonitory patterns [9,51,52].
      - *Validation of an algorithm and joint optimization of prediction and preparedness strategy* [38].
      - *Developing prediction algorithms for other types of extreme events.*

The authors would be glad to provide specific information upon request.

### Generalizations

**The problems considered here** have the following common features:

- *The absence of a closed theory* that would unambiguously determine prediction methodology. This leads to the need for intense intertwining of mathematics, statistical physics and non-linear dynamics, a range of societal sciences, and practical experience (Subsect. “**Methodology**”). In reality this requires long-term collaboration of respective experts. As can be seen from the references to Sects. “**Elections**”, “**US Economic Recessions**”, “**Unemployment**”, “**Homicide Surges**” previous applications inevitably involved the teams of such experts.
- *Predictions in advance* is the only final validation of the results obtained.
- *The need for holistic analysis* driven to extreme robustness.
- *Considerable, albeit limited, universality* of the premonitory phenomena.

Two classical quotations shed the light on these features:

A. N. Kolmogoroff. “It became clear for me that it is unrealistic to have a hope for the creation of a pure theory [of the turbulent flows of fluids and gases] closed in itself. Due to the absence of such a theory we have to rely upon the hypotheses obtained by processing of the experimental data.”

M. Gell-Mann: “... if the parts of a complex system or the various aspects of a complex situation, all defined in advance, are studied carefully by experts on those parts or aspects, and the results of their work are pooled, an adequate description of the whole system or situation does

not usually emerge. ... The reason, of course, is that these parts or aspects are typically entangled with one another. ... We have to supplement the partial studies with a transdisciplinary crude look at the whole.”

*In the general scheme of things* the problem considered belongs to a much wider field – the quest for a universal theory of complex systems extended to predicting extreme events – the Holy Grail of complexity studies. This quest encompasses the natural and human-made complex systems that comprise what some analysts have called “the global village”. It requires entirely new applications of modern science, such as algebraic geometry, combinatorics, and thermodynamics. As a means for anticipating, preventing and responding to natural and manmade disasters and for improving the outcomes of economic and political systems, the methods described here may hold one key for the survival and sustainability of our civilization.

## Bibliography

### Primary Literature

- Allègre CJ, Le Mouél J-L, Ha Duyen C, Narteau C (1995) Scaling organization of fracture tectonics (SOFT) and earthquake mechanism. *Phys Earth Planet Inter* 92:215–233
- Armstrong JS, Cuzan AG (2005) Index methods for forecasting: An application to american presidential elections. *Foresight Int J Appl Forecast* 3:10–13
- Blanter EM, Shnirman MG, Le Mouél JL, Allègre CJ (1997) Scaling laws in blocks dynamics and dynamic self-organized criticality. *Phys Earth Planet Inter* 99:295–307
- Bongard MM, Vaintsveig MI, Guberman SA, Izvekova ML, Smirnov MS (1966) The use of self-learning prog in the detection of oil containing layers. *Geol Geofiz* 6:96–105
- Burridge R, Knopoff L (1967) Model and theoretical seismicity. *Bull Seismol Soc Am* 57:341–360
- Carlson SM (1998) Uniform crime reports: Monthly weapon-specific crime and arrest time series 1975–1993 (National, State, 12-City Data), ICPSR 6792 Inter-university Consortium for Political and Social Research. Ann Arbor
- Farmer JD, Sidorowich J (1987) Predicting chaotic time series. *Phys Rev Lett* 59:845
- Gabrielov A, Keilis-Borok V, Zaliapin I, Newman WI (2000) Critical transitions in colliding cascades. *Phys Rev E* 62:237–249
- Gabrielov A, Keilis-Borok V, Zaliapin I (2007) Predictability of extreme events in a branching diffusion model. [arXiv:0708.1542](https://arxiv.org/abs/0708.1542)
- Gabrielov AM, Zaliapin IV, Newman WI, Keilis-Borok VI (2000) Colliding cascade model for earthquake prediction. *Geophys J Int* 143(2):427–437
- Gelfand IM, Guberman SA, Keilis-Borok VI, Knopoff L, Press F, Ranzman IY, Rotwain IM, Sadovsky AM (1976) Pattern recognition applied to earthquake epicenters in California. *Phys Earth Planet Inter* 11:227–283
- Gell-Mann M (1994) *The quark and the jaguar: Adventures in the simple and the complex*. Freeman, New York
- Crutchfield JP, Farmer JD, Packard NH, Shaw RS (1986) *Chaos* *Sci Am* 255:46–57
- Gvishiani AD, Kosobokov VG (1981) On found of the pattern recognition results applied to earthquake-prone areas. *Izvestiya Acad Sci USSR. Phys Earth* 2:21–36
- Holland JH (1995) *Hidden order: How adaptation builds complexity*. Addison, Reading
- IMF (1997) *International monetary fund, international financial statistics*. CD-ROM
- Kadanoff LP (1976) Scaling, universality and operator algebras. In: Domb C, Green MS (eds) *Phase transitions and critical phenomena*, vol 5a. Academic, London, pp 1–34
- Keilis-Borok VI, Lichtman AJ (1993) The self-organization of American society in presidential and senatorial elections. In: Kravtsov YA (ed) *Limits of predictability*. Springer, Berlin, pp 223–237
- Keilis-Borok VI, Press F (1980) On seismological applications of pattern recognition. In: Allègre CJ (ed) *Source mechanism and earthquake prediction applications*. Editions du centre national du la recherché scientifique, Paris, pp 51–60
- Keilis-Borok VI, Soloviev AA (eds) (2003) *Nonlinear dynamics of the lithosphere and earthquake prediction*. Springer, Berlin
- Keilis-Borok V, Soloviev A (2007) Pattern recognition methods and algorithms. Ninth workshop on non-linear dynamics and earthquake prediction, Trieste ICTP 1864-11
- Keilis-Borok VI, Sorondo MS (2000) (eds) *Science for survival and sustainable development*. The proceedings of the study-week of the Pontifical Academy of Sciences, 12–16 March 1999. Pontificiae Academiae Scientiarum Scripta Varia, Vatican City
- Keilis-Borok V, Stock JH, Soloviev A, Mikhalev P (2000) Pre-recession pattern of six economic indicators in the USA. *J Forecast* 19:65–80
- Keilis-Borok VI, Gascon DJ, Soloviev AA, Intriligator MD, Pichardo R, Winberg FE (2003) On predictability of homicide surges in megacities. In: Beer T, Ismail-Zadeh A (eds) *Risk science and sustainability*. Kluwer, Dordrecht (NATO Sci Ser II Math, Phys Chem 112), pp 91–110
- Keilis-Borok VI, Soloviev AA, Allègre CB, Sobolevskii AN, Intriligator MD (2005) Patterns of macroeconomic indicators preceding the unemployment rise in Western Europe and the USA. *Pattern Recogn* 38(3):423–435
- Keilis-Borok V, Soloviev A, Gabrielov A, Zaliapin I (2007) Change of scaling before extreme events in complex systems. In: Proceedings of the plenary session on “predictability in science: Accuracy and limitations”, Pontificiae Academiae Scientiarum Scripta Varia, Vatican City
- Kravtsov YA (ed) (1993) *Limits of predictability*. Springer, Berlin
- Lichtman AJ, Keilis-Borok VI (1989) Aggregate-level analysis and prediction of midterm senatorial elections in the United States, 1974–1986. *Proc Natl Acad Sci USA* 86(24):10176–10180
- Lichtman AJ (1996) *The keys to the White House*. Madison Books, Lanham
- Lichtman AJ (2005) *The keys to the White House: Forecast for 2008*. *Foresight Int J Appl Forecast* 3:5–9
- Lichtman AJ (2008) *The keys to the White House*, 2008 edn. Rowman/Littlefield, Lanham
- Ma Z, Fu Z, Zhang Y, Wang C, Zhang G, Liu D (1990) *Earthquake prediction: Nine major earthquakes in china*. Springer, New York
- Mason IB (2003) Binary events. In: Jolliffe IT, Stephenson DB (eds) *Forecast verification. A practitioner’s guide in atmospheric science*. Wiley, Chichester, pp 37–76

34. Molchan GM (1990) Strategies in strong earthquake prediction. *Phys Earth Planet Inter* 61:84–98
35. Molchan GM (1991) Structure of optimal strategies of earthquake prediction. *Tectonophysics* 193:267–276
36. Molchan GM (1994) Models for optimization of earthquake prediction. In: Chowdhury DK (ed) *Computational seismology and geodynamics*, vol 1. Am Geophys Un, Washington, pp 1–10
37. Molchan GM (1997) Earthquake prediction as a decision-making problem. *Pure Appl Geophys* 149:233–237
38. Molchan GM (2003) Earthquake prediction strategies: A theoretical analysis. In: Keilis-Borok VI, Soloviev AA (eds) *Nonlinear dynamics of the lithosphere and earthquake prediction*. Springer, Berlin, pp 209–237
39. Molchan G, Keilis-Borok V (2008) Earthquake prediction: Probabilistic aspect. *Geophys J Int* 173(3):1012–1017
40. NACJD: <http://www.icpsr.umich.edu/NACJD/index.html>
41. NBER: <http://www.nber.org/cycles/cyclesmain.html>
42. Newman W, Gabrielov A, Turcotte DL (eds) (1994) *Nonlinear dynamics and predictability of geophysical phenomena*. Am Geophys Un, Int Un Geodesy Geophys, Washington
43. OECD (1997) *Main economic indicators: Historical statistics 1960–1996*. Paris, CD-ROM
44. Press F, Briggs P (1975) Chandler wobble, earthquakes, rotation and geomagnetic changes. *Nature* 256:270–273, London
45. Press F, Briggs P (1977) Pattern recognition applied to uranium prospecting. *Nature* 268:125–127
46. Press F, Allen C (1995) Patterns of seismic release in the southern California region. *J Geophys Res* 100(B4):6421–6430
47. Soloviev A (2007) Application of the pattern recognition techniques to earthquake-prone areas determination. Ninth workshop on non-linear dynamics and earthquake prediction, Trieste ICTP 1864-9
48. Stock JH, Watson MW (1989) New indexes of leading and coincident economic indicators. *NBER Macroecon Ann* 4:351–394
49. Stock JH, Watson MW (1993) A procedure for predicting recessions with leading indicators. In: Stock JH, Watson MW (eds) *Business cycles, indicators, and forecasting* (NBER Studies in Business Cycles, vol 28), pp 95–156
50. Tukey JW (1977) *Exploratory data analysis*. Addison-wesley series in behavioral science: Quantitative methods. Addison, Reading
51. Turcotte DL, Newman WI, Gabrielov A (2000) A statistical physics approach to earthquakes. In: *Geocomplexity and the physics of earthquakes*. Am Geophys Un, Washington
52. Zaliapin I, Keilis-Borok V, Ghil M (2003) A Boolean delay model of colliding cascades, II: Prediction of critical transitions. *J Stat Phys* 111(3–4):839–861

### Books and Reviews

- Bongard MM (1967) *The problem of recognition*. Nauka, Moscow
- Brito DL, Intriligator MD, Worth ER (1998) In: Eliasson G, Green C (eds) *Microfoundations of economic growth: A Schumpeterian perspective*. University of Michigan Press, Ann Arbor
- Bui Trong L (2003) Risk of collective youth violence in french suburbs. A clinical scale of evaluation, an alert system. In: Beer T, Ismail-Zadeh A (eds) *Risk science and sustainability*. Kluwer, Dordrecht (NATO Sci Ser II Math Phys Chem 112)
- Engle RF, McFadden DL (1994) (eds) *Handbook of econometrics*, vol 4. North-Holland, Amsterdam
- Klein PA, Niemira MP (1994) *Forecasting financial and economic cycles*. Wiley, New York
- Messner SF (1983) Regional differences in the economic correlates of the urban homicide rate. *Criminology* 21:477–488
- Mitchell WC (1951) *What happens during business cycles: A progress report*. NBER, New York
- Mitchell WC, Burns AF (1946) *Measuring business cycles*. NBER, New York
- Moore GH (ed) (1961) *Business cycle indicators*. NBER, New York
- Mostaghimi M, Rezayat F (1996) Probability forecast of a downturn in US economy using classical statistical theory. *Empir Econ* 21:255–279
- Watson MW (1994) In: Engle RF, McFadden DL (eds) *Handbook of econometrics*, vol IV. North-Holland, Amsterdam

## Fractals in Geology and Geophysics

DONALD L. TURCOTTE  
 Department of Geology,  
 University of California,  
 Davis, USA

### Article Outline

Glossary  
 Definition of the Subject  
 Introduction  
 Drainage Networks  
 Fragmentation  
 Earthquakes  
 Volcanic Eruptions  
 Landslides  
 Floods  
 Self-Affine Fractals  
 Topography  
 Earth's Magnetic Field  
 Future Directions  
 Bibliography

### Glossary

**Fractal** A collection of objects that have a power-law dependence of number on size.

**Fractal dimension** The power-law exponent in a fractal distribution.

### Definition of the Subject

The scale invariance of geological phenomena is one of the first concepts taught to a student of geology. When a photograph of a geological feature is taken, it is essential to include an object that defines the scale, for example, a coin or a person. It was in this context that Mandelbrot [7] introduced the concept of fractals. The length of a rocky coastline is obtained using a measuring rod with a specified length. Because of scale invariance, the length of the coastline increases as the length of the measuring rod decreases according to a power law. It is not possible to obtain a specific value for the length of a coastline due to small indentations down to a scale of millimeters or less.

A fractal distribution requires that the number of objects  $N$  with a linear size greater than  $r$  has an inverse power-law dependence on  $r$  so that

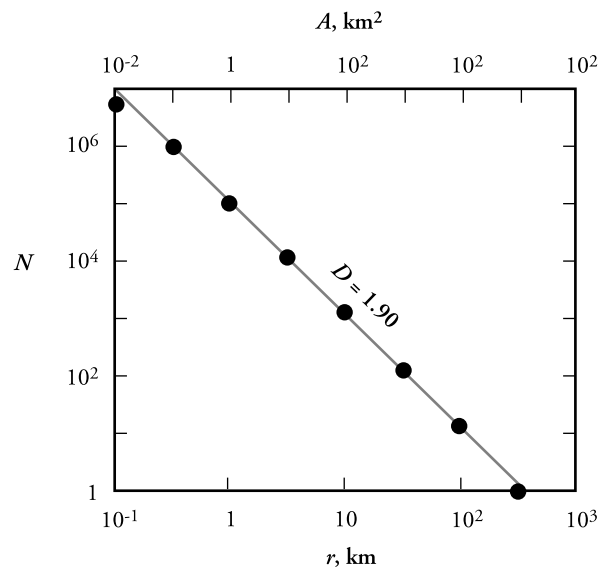
$$N = \frac{C}{r^D} \quad (1)$$

where  $C$  is a constant and the power  $D$  is the fractal dimension. This power-law scaling is the only distribution that is

scale invariant. However, the power-law dependence cannot be used to define a statistical distribution because the integral of the distribution diverges to infinity either for large values or small values of  $r$ . Thus fractal distributions never appear in compilations of statistical distributions. A variety of statistical distributions have power-law behavior either at large scales or small scales, but not both. An example is the Pareto distribution.

Many geological phenomena are scale invariant. Examples include the frequency-size distributions of fragments, faults, earthquakes, volcanic eruptions, and landslides. Stream networks and landforms exhibit scale invariance. In terms of these applications there must always be upper and lower cutoffs to the applicability of a fractal distribution. As a specific application consider earthquakes on the Earth. The number of earthquakes has a power-law dependence on the size of the rupture over a wide range of sizes. But the largest earthquake cannot exceed the size of the Earth, say  $10^4$  km. Also, the smallest earthquake cannot be smaller than the grain size of rocks, say 1 mm. But this range of scales is  $10^{10}$ . Actual earthquakes appear to satisfy fractal scaling over the range 1 m to  $10^3$  km.

An example of fractal scaling is the number-area distribution of lakes [10], this example is illustrated in Fig. 1. Excellent agreement with the fractal relation given in Eq. (1)



Fractals in Geology and Geophysics, Figure 1  
 Dependence of the cumulative number of lakes  $N$  with areas greater than  $A$  as a function of  $A$ . Also shown is the linear dimension  $r$  which is taken to be the square root of  $A$ . The straight-line correlation is with Eq. (1) taking the fractal dimension  $D = 1.90$

is obtained taking  $D = 1.90$ . The linear dimension  $r$  is taken to be the square root of the area  $A$  and the power-law (fractal) scaling extends from  $r = 100$  m to  $r = 300$  km.

## Introduction

Fractal scaling evolved primarily as an empirical means of correlating data. A number of examples are given below. More recently a theoretical basis has evolved for the applicability of fractal distributions. The foundation of this basis is the concept of self-organized criticality. A number of simple computational models have been shown to yield fractal distributions. Examples include the sand-pile model, the forest-fire model, and the slider-block model.

## Drainage Networks

Drainage networks are a universal feature of landscapes on the Earth. Small streams merge to form larger streams, large streams merge to form rivers, and so forth. Strahler [16] quantified stream networks by introducing an ordering system. When two like-order streams of order  $i$  merge they form a stream of order  $i + 1$ . Thus two  $i = 1$  streams merge to form a  $i = 2$  stream, two  $i = 2$  streams merge to form a  $i = 3$  stream and so forth. A bifurcation ratio  $R_b$  is defined by

$$R_b = \frac{N_i}{N_{i+1}} \quad (2)$$

where  $N_i$  is the number of streams of order  $i$ . A length order ratio  $R_r$  is defined by

$$R_r = \frac{r_{i+1}}{r_i} \quad (3)$$

where  $r_i$  is the mean length of streams of order  $i$ . Empirically both  $R_b$  and  $R_r$  are found to be nearly constant for a range of stream orders in a drainage basin. From Eq. (1) the fractal dimension of a drainage basin

$$D = \frac{\ln(N_i/N_{i+1})}{\ln(r_{i+1}/r_i)} = \frac{\ln R_b}{\ln R_r} \quad (4)$$

Typically  $R_b = 4.6$ ,  $R_r = 2.2$ , and the corresponding fractal dimension is  $D = 1.9$ . This scale invariant scaling of drainage networks was recognized some 20 years before the concept of fractals was introduced in 1967.

A major advance in the quantification of stream networks was made by Tokunaga [17]. This author was the first to recognize the importance of side branching, that is some  $i = 1$  streams intersect  $i = 2$ ,  $i = 3$ , and all higher-order streams. Similarly,  $i = 2$  streams intersect  $i = 3$  and higher-order streams and so forth. A fully

self-similar, side-branching topology was developed. Applications to drainage networks have been summarized by Peckham [11] and Pelletier [13].

## Fragmentation

An important application of power-law (fractal) scaling is to fragmentation. In many examples the frequency-mass distributions of fragments are fractal. Explosive fragmentation of rocks (for example in mining) give fractal distributions. At the largest scale the frequency size distribution of the tectonic plates of plate tectonics are reasonably well approximated by a power-law distribution. Fault gouge is generated by the grinding process due to earthquakes on a fault. The frequency-mass distribution of the gouge fragments is fractal. Grinding (comminution) processes are common in tectonics. Thus it is not surprising that fractal distributions are ubiquitous in geology.

As a specific example consider the frequency-mass distribution of asteroids. Direct measurements give a fractal distribution. Since asteroids are responsible for the impact craters on the moon, it is not surprising that the frequency-area distribution of lunar craters is also fractal.

Using evidence from the moon and a fractal extrapolation it is estimated that on average, a 1m diameter meteorite impacts the earth every year, that a 100 m diameter meteorite impacts every 10,000 years, and that a 10 km diameter meteorite impacts the earth every 100,000,000 years. The classic impact crater is Meteor Crater in Arizona, it is over 1 km wide and 200 m deep. Meteor Crater formed about 50,000 years ago and it is estimated that the impacting meteorite had a diameter of 30 m. The largest impact to occur in the 20th century was the June 30, 1908 Tunguska event in central Siberia. The impact was observed globally and destroyed over 1000 km<sup>2</sup> of forest. It is believed that this event was the result of a 30 m diameter meteorite that exploded in the atmosphere.

One of the major global extinctions occurred at the Cretaceous/Tertiary boundary 65 million years ago. Some 65% of the existing species were destroyed including dinosaurs. This extinction is attributed to a massive impact at the Chicxulub site on the Yucatan Peninsula, Mexico. It is estimated that the impacting meteorite had a 10 km diameter. In addition to the damage done by impacts there is evidence that impacts on the oceans have created massive tsunamis. The fractal power-law scaling can be used to quantify the risk of future impacts.

## Earthquakes

Earthquakes universally satisfy several scaling laws. The most famous of these is Gutenberg-Richter frequency-



magnitude scaling. The magnitude  $M$  of an earthquake is an empirical measure of the size of an earthquake. If the magnitude is increased by one unit it is observed that the cumulative number of earthquakes greater than the specified magnitude is reduced by a factor of 10.

For the entire earth, on average, there is 1 magnitude 8 earthquake per year, 10 magnitude 7 earthquakes per year, and 100 magnitude 6 earthquakes per year. When magnitude is converted to rupture area a fractal relation is obtained. The numbers of earthquakes that occur in a specified region and time interval have a power-law dependence on the rupture area.

The validity of this fractal scaling has important implications for probabilistic seismic risk assessment. The number of small earthquakes that occur in a region can be extrapolated to estimate the risk of larger earthquakes [1]. As an example consider southern California. On average there are 30 magnitude 4 or larger earthquakes per year. Using the fractal scaling it is estimated that the expected intervals between magnitude 6 earthquakes will be 3 years, between magnitude 7 earthquakes will be 30 years, and between magnitude 8 earthquakes will be 300 years.

The fractal scaling of earthquakes illustrate a useful aspect of fractal distributions. The fractal distribution requires two parameters. The first parameter, the fractal dimension  $D$  (known as the  $b$ -value in seismology), gives the dependence of number on size (magnitude). For earthquakes the fractal dimension is almost constant independent of the tectonic setting. The second parameter gives the level of activity. For example, this can be the number of earthquakes greater than a specified magnitude in a region. This level of activity varies widely and is an accepted measure of seismic risk. The level is essentially zero in states like Minnesota and is a maximum in California.

### Volcanic Eruptions

There is good evidence that the frequency-volume statistics of volcanic eruptions are also fractal [9]. Although it is difficult to quantify the volumes of magma and ash associated with older eruptions, the observations suggest that an eruption with a volume of  $1 \text{ km}^3$  would be expected each 10 years,  $10 \text{ km}^3$  each 100 years, and  $100 \text{ km}^3$  each 1000 years. For example, the 1991 Mount Pinatubo, Philippines eruption had an estimated volume of about  $5 \text{ km}^3$ . The most violent eruption in the last 200 years was the 1815 Tambora, Indonesia eruption with an estimated volume of  $150 \text{ km}^3$ . This eruption influenced the global climate in 1816 which was known as the year without a summer. It is estimated that the Long Valley, California eruption with an age of about 760,000 years had a volume of about

$600 \text{ km}^3$  and the Yellowstone eruptions of about 600,000 years ago had a volume of about  $2000 \text{ km}^3$ .

Although the validity of the power-law (fractal) extrapolation of volcanic eruption volumes to long periods in the past can be questioned, the extrapolation does give some indication of the risk of future eruptions to global climate. There is no doubt that the large eruptions that are known to have occurred on time scales of  $10^5$  to  $10^6$  years would have a catastrophic impact on global agricultural production.

### Landslides

Landslides are a complex natural phenomenon that constitutes a serious natural hazard in many countries. Landslides also play a major role in the evolution of landforms. Landslides are generally associated with a trigger, such as an earthquake, a rapid snowmelt, or a large storm. The landslide event can include a single landslide or many thousands. The frequency-area distribution of a landslide event quantifies the number of landslides that occur at different sizes. It is generally accepted that the number of large landslides with area  $A$  has a power-law dependence on  $A$  with an exponent in the range 1.3 to 1.5 [5].

Unlike earthquakes, a complete statistical distribution can be defined for landslides. A universal fit to an inverse-gamma distribution has been found for a number of event inventories. This distribution has a power-law (fractal) behavior for large landslides and an exponential cut-off for small landslides. The most probable landslides have areas of about  $40 \text{ m}^2$ . Very few small landslides are generated.

As a specific example we consider the 11,111 landslides generated by the magnitude 6.7 Northridge (California) earthquake on January 17, 1994. The total area of the landslides was  $23.8 \text{ km}^2$  and the area of the largest landslide was  $0.26 \text{ km}^2$ . The inventory of landslide areas had a good power-law dependence on area for areas greater than  $10^3 \text{ m}^2$  ( $10^{-3} \text{ km}^2$ ). The number of landslides generated by earthquakes have a strong dependence on earthquake magnitude. Typically earthquakes with magnitudes  $M$  less than 4 do not generate any landslides [6].

### Floods

Floods are a major hazard to many cities and estimates of flood hazards have serious economic implications. The standard measure of the flood hazard is the 100-year flood. This is quantified as the river discharge  $Q_{100}$  expected during a 100 year period. Since there is seldom a long enough history to establish  $Q_{100}$  directly, it is necessary to extrapolate smaller floods that occur more often.

One extrapolation approach is to assume that flood discharges are fractal (power-law) [3,19]. This scale invariant distribution can be expressed in terms of the ratio  $F$  of the peak discharge over a 10 year interval to the peak discharge over a 1 year interval,  $F = Q_{10}/Q_1$ . With self-similarity the parameter  $F$  is also the ratio of the 100 year peak discharge to the 10 year peak discharge,  $F = Q_{100}/Q_{10}$ . Values of  $F$  have a strong dependence on climate. In temperate climates such as the northeastern and northwestern US values are typically in the range  $F = 2-3$ . In arid and tropical climates such as the southwestern and southeastern US values are typically in the range  $F = 4-6$ .

The applicability of fractal concepts to flood forecasting is certainly controversial. In 1982, the US government adopted the log-Pearson type 3 (LP3) distribution for the legal definition of the flood hazard [20]. The LP3 is a thin-tailed distribution relative to the thicker tailed power-law (fractal) distribution. Thus the forecast 100 year flood using LP3 is considerably smaller than the forecast using the fractal approach. This difference is illustrated by considering the great 1993 Mississippi River flood. Considering data at the Keukuk, Iowa gauging station [4] this flood was found to be a typical 100 year flood using the power-law (fractal) analysis and a 1000 to 10,000 year flood using the federal LP3 formulation. Concepts of self-similarity argue for the applicability of fractal concepts for flood-frequency forecasting. This applicability also has important implications for erosion. Erosion will be dominated by the very largest floods.

### Self-Affine Fractals

Mandelbrot and Van Ness [8] extended the concept of fractals to time series. Examples of time series in geology and geophysics include global temperature, the strength of the Earth's magnetic field, and the discharge rate in a river. After periodicities and trends have been removed, the remaining values are the stochastic (noise) component of the time series. The standard approach to quantifying the noise component is to carry out a Fourier transform on the time series [2]. The power-spectral density coefficients  $S_i$  are proportional to the squares of the Fourier coefficients. The time series is a self-affine fractal if the power-spectral density coefficients have an inverse power-law dependence on frequency  $f_i$ , that is

$$S_i = \frac{C}{f_i^\beta} \quad (5)$$

where  $C$  is a constant and  $\beta$  is the power-law exponent.

For a Gaussian white noise the values in the time series are selected randomly from a Gaussian distribution.

Adjacent values are not correlated with each other. In this case the spectrum is flat and the power spectral density coefficients are not a function of frequency,  $\beta = 0$ . The classic example of a self-affine fractal is a Brownian walk. A Brownian walk is obtained by taking the running sum of a Gaussian white noise. In this case we have  $\beta = 2$ . Another important self-affine time series is a red (or pink) noise with power spectral density coefficients proportional to  $1/f$ , that is  $\beta = 1$ . We will see that the variability in the Earth's magnetic field is well approximated by a  $1/f$  noise.

Self-affine fractal time series in the range  $\beta = 0$  to 1 are known as fractional Gaussian noises. These noises are stationary and the standard deviation is a constant independent of the length of the time series. Self-affine time series with  $\beta$  larger than 1 are known as fractional Brownian walk. These motions are not stationary and the standard deviation increases as a power of the length of the time series, there is a drift. For a Brownian walk the standard deviation increases with the square root of the length of the time series.

### Topography

The height of topography along linear tracks can be considered to be a continuous time series. In this case we consider the wave number  $k_i$  (1/wave length) instead of frequency. Topography is a self-affine fractal if

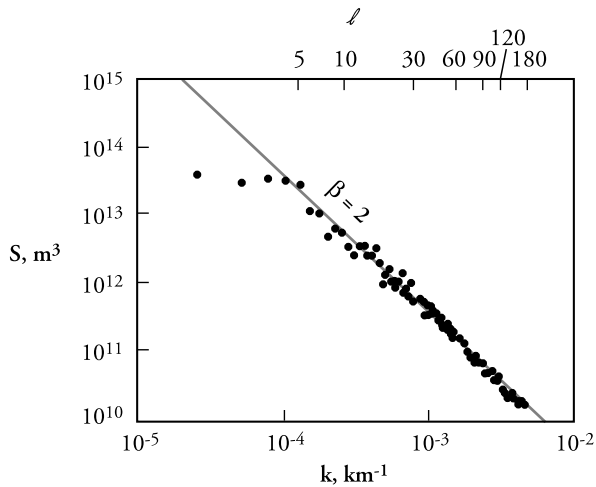
$$S_i = \frac{C}{k_i^\beta} \quad (6)$$

Spectral expansions of global topography have been carried out, an example [15] is given in Fig. 2. Excellent agreement with the fractal relation given in Eq. (6) is obtained taking  $\beta = 2$ , topography is well approximated by a Brownian walk. It has also shown that this fractal behavior of topography is found for the moon, Venus, and Mars [18].

### Earth's Magnetic Field

Paleomagnetic studies have given the strength and polarity of the Earth's magnetic field as a function of time over millions of years. These studies have also shown that the field has experienced a sequence of reversals.

Spectral studies of the absolute amplitude of the field have been shown that it is a self-affine fractal [12,14]. The power-spectral density is proportional to one over the frequency, it is a  $1/f$  noise. When the fluctuations of the  $1/f$  noise take the magnitude to zero the polarity of the field reverses. The predicted distribution of polarity intervals is fractal and is in good agreement with the observed polarity intervals.



Fractals in Geology and Geophysics, Figure 2

Power spectral density  $S$  as a function of wave number  $k$  for a spherical harmonic expansion of the Earth's topography (degree  $l$ ). The straight-line correlation is with Eq. (6) taking  $\beta = 2$ , a Brownian walk

### Future Directions

There is no question that fractals are a useful empirical tool. They provide a rational means for the extrapolation and interpolation of observations. The wide applicability of power-law (fractal) distributions is generally accepted, but does this applicability have a more fundamental basis? Fractality appears to be fundamentally related to chaotic behavior and to numerical simulations exhibiting self-organized criticality. The entire area of fractals, chaos, self-organized criticality, and complexity remains extremely active, and it is impossible to predict with certainty what the future holds.

### Bibliography

#### Primary Literature

1. Kossobokov VG, Keilis-Borok VI, Turcotte DL, Malamud BD (2000) Implications of a statistical physics approach for earthquake hazard assessment and forecasting. *Pure Appl Geophys* 157:2323
2. Malamud BD, Turcotte DL (1999) Self-affine time series: I. Generation and analyses. *Adv Geophys* 40:1
3. Malamud BD, Turcotte DL (2006) The applicability of power-law frequency statistics to floods. *J Hydrol* 332:168

4. Malamud BD, Turcotte DL, Barton CC (1996) The 1993 Mississippi river flood: A one hundred or a one thousand year event? *Env Eng Geosci* 2:479
5. Malamud BD, Turcotte DL, Guzzetti F, Reichenbach P (2004) Landslide inventories and their statistical properties. *Earth Surf Process Landf* 29:687
6. Malamud BD, Turcotte DL, Guzzetti F, Reichenbach P (2004) Landslides, earthquakes, and erosion. *Earth Planet Sci Lett* 229:45
7. Mandelbrot BB (1967) How long is the coast of Britain? Statistical self-similarity and fractional dimension. *Science* 156:636
8. Mandelbrot BB, Van Ness JW (1968) Fractional Brownian motions, fractional noises and applications. *SIAM Rev* 10:422
9. McClelland L et al (1989) *Global Volcanism 1975-1985*. Prentice-Hall, Englewood Cliffs
10. Meybeck M (1995) Global distribution of lakes. In: Lerman A, Imboden DM, Gat JR (eds) *Physics and Chemistry of Lakes*, 2nd edn. Springer, Berlin, pp 1-35
11. Peckham SD (1989) New results for self-similar trees with applications to river networks. *Water Resour Res* 31:1023
12. Pelletier JD (1999) Paleointensity variations of Earth's magnetic field and their relationship with polarity reversals. *Phys Earth Planet Int* 110:115
13. Pelletier JD (1999) Self-organization and scaling relationships of evolving river networks. *J Geophys Res* 104:7259
14. Pelletier JD, Turcotte DL (1999) Self-affine time series: II. Applications and models. *Adv Geophys* 40:91
15. Rapp RH (1989) The decay of the spectrum of the gravitational potential and the topography of the Earth. *Geophys J Int* 99:449
16. Strahler AN (1957) Quantitative analysis of watershed geomorphology. *Trans Am Geophys Un* 38:913
17. Tokunaga E (1978) Consideration on the composition of drainage networks and their evolution. *Geogr Rep Tokyo Metro Univ* 13:1
18. Turcotte DL (1987) A fractal interpretation of topography and geoid spectra on the earth, moon, Venus, and Mars. *J Geophys Res* 92:E597
19. Turcotte DL (1994) Fractal theory and the estimation of extreme floods. *J Res Natl Inst Stand Technol* 99:377
20. US Water Resources Council (1982) *Guidelines for Determining Flood Flow Frequency*. Bulletin 17B. US Geological Survey, Reston

#### Books and Reviews

- Feder J (1988) *Fractals*. Plenum Press, New York
- Korvin G (1992) *Fractal Models in the Earth Sciences*. Elsevier, Amsterdam
- Mandelbrot BB (1982) *The Fractal Geometry of Nature*. Freeman, San Francisco
- Turcotte DL (1997) *Fractals and Chaos in Geology and Geophysics*, 2nd edn. Cambridge University Press, Cambridge

## Geo-complexity and Earthquake Prediction

VLADIMIR KEILIS-BOROK<sup>1,2</sup>, ANDREI GABRIELOV<sup>3</sup>,  
ALEXANDRE SOLOVIEV<sup>2,4</sup>

<sup>1</sup> Institute of Geophysics and Planetary Physics and  
Department of Earth and Space Sciences,  
University of California, Los Angeles, USA

<sup>2</sup> International Institute of Earthquake Prediction Theory  
and Mathematical Geophysics,  
Russian Academy of Sciences, Moscow, Russia

<sup>3</sup> Departments of Mathematics and Earth  
and Atmospheric Sciences, Purdue University,  
West Lafayette, USA

<sup>4</sup> The Abdus Salam International Center for Theoretical  
Physics, Trieste, Italy

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Lithosphere as a Hierarchical Complex System](#)

[General Scheme of Prediction](#)

[Four Paradigms](#)

[Earthquake Prediction and Earthquake Preparedness](#)

[Further Goals](#)

[Acknowledgment](#)

[Bibliography](#)

### Glossary

**Chaos** Apparently random or unpredictable behavior in systems governed by deterministic laws. The common element in these systems is a very high sensitivity to initial conditions and to the way in which a system is set in motion (Encyclopedia Britannica).

**Complexity** An attribute of nonlinear (chaotic) systems. It comprises instability and complex but not random behavior patterns – “order in chaos”.

**Earthquake** An episode of rupture and discontinuous displacement within the solid Earth. Part of the energy accumulated around the rupture is released by inelastic deformation and seismic waves. Both may cause destructive shaking of the ground, if the energy release is sufficiently large.

**Earthquake forecasting** Probabilistic extrapolation of seismic activity comprising many earthquakes.

**Earthquake prediction** Prediction of time interval, geographic area, and magnitude range where an individual future strong earthquake will occur. The prediction

is meaningful if it includes an estimated rate of false alarms.

**Earthquake preparedness** A set of actions reducing the damage from the future earthquakes. There are different levels of preparedness.

**Extreme events** Rare events of low probability but high impact on a system where they occur. In different connotations they are also known as critical transitions, disasters, catastrophes, and crises. Over time they persistently recur in both natural and constructed complex systems. In this article the extreme events are the strong earthquakes. An earthquake might be an extreme event in a certain volume of the lithosphere and part of the background seismicity in a larger volume.

**Lithosphere** The earthquake-prone outer shell of the solid Earth. In prediction research it is regarded as a hierarchical complex system.

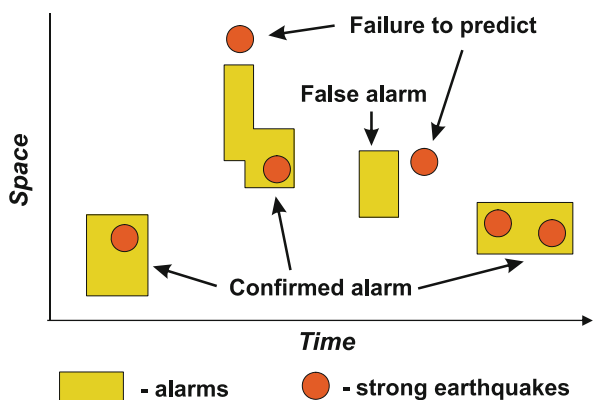
**Premonitory seismicity patterns** Space-time-magnitude patterns of earthquake occurrences that signal the approach of a strong earthquake.

### Definition of the Subject

#### Definition

The *problem* of earthquake prediction is to find when and where a strong earthquake will occur. A prediction is formulated as a discrete sequence of alarms (Fig. 1). The accuracy of a prediction method is captured by probabilities of errors (false alarms and failures to predict) and by the total space-time occupied by alarms. (Sect. “Error Diagram”).

In terms of prediction studies this is algorithmic prediction of individual extreme events having low probability but large impact. This problem is necessarily intertwined with problems of disaster preparedness, dynamics



Geo-complexity and Earthquake Prediction, Figure 1  
Possible outcomes of prediction

of solid Earth, and modeling of extreme events in hierarchical complex systems.

*Predictability* (“*order in chaos*”). Complex systems, lithosphere included, are not predictable with unlimited precision. However, after a coarse-graining (i. e., in a not-too-detailed scale) certain regular behavior patterns emerge and a system becomes predictable, up to certain limits ([13,20,24,26,36,46,52,83]). Accordingly, earthquake prediction requires a holistic analysis, “from the whole to details”. Such analysis makes it possible to overcome the geo-complexity itself and the chronic imperfection of observations as well.

*Premonitory patterns*. Certain behavior patterns emerge more frequently as a strong earthquake draws near. Called premonitory patterns, they signal destabilization of the earthquake-prone lithosphere and thus an increase in the probability of a strong earthquake. Premonitory patterns do not necessarily contribute to causing a subsequent strong earthquake; both might be parallel manifestations of the same underlying process – the tectonic development of the Earth in multiple time-, space-, and energy- scales. For that reason premonitory patterns might emerge in a broad variety of observable fields reflecting lithosphere dynamics, and in different scales.

The algorithms considered here, based on premonitory seismicity patterns, provide alarms lasting years to months. There is ample evidence that major findings made in developing these algorithms are applicable to premonitory patterns in other fields, to predicting other geological and geotechnical disasters, and probably to determining shorter and longer alarms (Sect. “**Further Goals**”).

## Importance

Algorithmic earthquake prediction provides pivotal constraints for fundamental understanding of the dynamics of the lithosphere and other complex systems. It is also critically important for protecting the global population, economy, and environment. Vulnerability of our world to the earthquakes is rapidly growing, due to proliferation of high-risk construction (nuclear power plants, high dams, radioactive waste disposals, lifelines, etc.), deterioration of ground and infrastructure in megacities, destabilization of environment, population growth, and escalating socio-economic volatility of the global village. Today a single earthquake with its ripple effects may take up to a million lives; destroy a megacity; trigger a global economic depression (e. g. if it occurs in Tokyo); trigger an ecological catastrophe, rendering a large territory uninhabitable; or destabilize military balance in a region. Regions of low

seismicity have become highly vulnerable, e. g. European and Indian platforms, and Central and Eastern parts of the U.S. As a result the earthquakes joined the ranks of the major disasters that, in the words of J. Wisner, have become “a threat to civilization survival, as great as was ever posed by Hitler, Stalin or the atom bomb”. Earthquake prediction is necessary to reduce the damage by escalating disaster preparedness. Predictions useful for preparedness should have known, but not necessarily high, accuracy. Such is the standard practice in preparedness for all disasters, wars included.

## Introduction

Earthquakes occur in some parts of the outer shell of the solid Earth, called the lithosphere; its thickness ranges from a few kilometers near the mid-ocean ridges to a few hundred kilometers in certain continental regions. At many continental margins the lithosphere bends downward penetrating underlying mantle as seismically active subduction zones. In seismically active regions a significant part of tectonic development is realized through the earthquakes.

About a million earthquakes with magnitude 2 (energy about  $10^{15}$  erg) or more are detected each year worldwide by seismological networks. About a hundred of these cause considerable damage and few times in a decade a catastrophic earthquake occurs.

Catalogs of earthquakes provide the data for detecting premonitory seismicity patterns. Typically for complexity studies we do not have a complete set of fundamental equations that govern dynamics of seismicity and unambiguously define earthquake prediction algorithms. This is due to the multitude of mechanisms controlling seismicity – see Sect. “**Generalization: Complexity and Extreme Events**”. In lieu of such equations “... *we have to rely upon the hypotheses obtained by processing of the experimental data*” (A. Kolmogorov on transition to turbulence). Formulating and testing such hypotheses involves exploratory data analysis, numerical and laboratory modeling, and theoretical studies (Sect. “**General Scheme of Prediction**”).

Diversity of methods and urgency of the problem makes learning by doing a major if not *the* major form of knowledge transfer in prediction of extreme events ([http://cdsagenda5.ictp.it/full\\_display.php?da=a06219](http://cdsagenda5.ictp.it/full_display.php?da=a06219)).

*Reliability of the existing algorithms* has been tested by continuous prediction of future strong earthquakes in numerous regions worldwide. Each algorithm is self-adapting, i. e. applicable without any changes in the regions with different seismic regimes. Predic-

tions are filed in advance at the websites (<http://www.mitp.ru/predictions.html>; <http://www.phys.ualberta.ca/mirrors/mitp/predictions.html>; and <http://www.igpp.ucla.edu/prediction/rtp/>).

Following is the scoring for four different algorithms.

- *Algorithms M8 [32] and MSc [44]* (MSc stands for the Mendocino Scenario). Algorithm M8 gives alarms with characteristic duration years. MSc gives a second approximation to M8, reducing the area of alarm. An example of their application is shown in Fig. 2.

Continually applied since 1992, algorithm M8 has predicted 10 out of 14 large earthquakes (magnitude 8 or more) which have occurred in the major seismic belts. Alarms occupied altogether about 30% of the time-space considered. Both algorithms applied together reduced the time-space alarms to 15%, but three more target earthquakes were missed by prediction.

- *Algorithm SSE or Second Strong Earthquake [43,91]*. Its aim is to predict whether or not a second strong earthquake will follow the one that had just occurred. An alarm lasts 18 months after the first strong earthquake. An example of prediction is shown in Fig. 3. Testing by prediction in advance is set up for California, Pamir and Tien Shan, Caucasus, Iberia and Maghreb, the Dead Sea rift, and Italy. Since 1989 this algorithm

made 29 predictions; 24 of which were correct and 5 were wrong.

These scores demonstrate predictability of individual earthquakes. A predictions' accuracy is indeed limited, but sufficient to prevent a considerable part of the damage.

- *Algorithm RTP or Reverse Tracing of Precursors [37,81]*. This algorithm gives alarms with a characteristic duration of months. An example of this prediction is shown in Fig. 4. Testing by prediction in advance started only few years ago for California, Japan, the Northern Pacific, Eastern Mediterranean, and Italy with adjacent areas.

*Perspective.* It is encouraging that only a small part of readily available relevant data, models and theories have been used for prediction so far. This suggests a potential for a substantial increase of prediction accuracy.

### Lithosphere as a Hierarchical Complex System

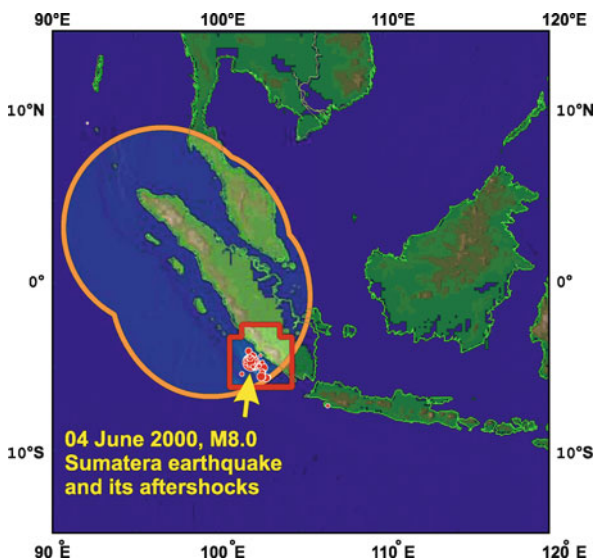
Two major factors turn the lithosphere into a hierarchical dissipative complex system [29,36,87]. The first one is a *hierarchical structure* extending from tectonic plates to grains of rocks. The second factor is *instability* caused by a multitude of nonlinear mechanisms destabilizing the strength and stress fields.

Among extreme events in that system are the strong earthquakes. *An earthquake may be an extreme event in a certain volume of the lithosphere and a part of the background seismicity in a larger volume.*

### Structure

**Blocks** The structure of the lithosphere presents a hierarchy of volumes, or blocks, which move relative to each other. The largest blocks are the major tectonic plates, of continental size. They are divided into smaller blocks, such as shields or mountain belts. After 15–20 consecutive divisions we come to about  $10^{25}$  grains of rocks of millimeter size.

**Boundary zones** Blocks are separated by relatively thin and less rigid boundary zones. They are called fault zones high in the hierarchy, then faults, sliding surfaces, and, finally, interfaces between grains of rock. Except at the bottom of the hierarchy, a boundary zone presents a similar hierarchical structure with more dense division. Some segments of the boundary zones, particularly in tectonically young regions, might be less explicitly expressed, presenting a bundle of small ruptures not yet merged into a fault, of a flexure not yet ruptured, etc.



Geo-complexity and Earthquake Prediction, Figure 2 Prediction of the Sumatra earthquake, June 4th, 2000,  $M = 8.0$  by algorithms M8 and MSc. The orange oval curve bounds the area of alarm determined by algorithm M8, the red rectangle is its reducing made by algorithm MSc. Circles show epicenters of the Sumatra earthquake and its aftershocks. After [43]



Geo-complexity and Earthquake Prediction, Figure 3

Prediction of the Northridge, California earthquake, January 28th, 1994,  $M = 6.8$  by algorithm SSE. The prediction was made by analysis of aftershocks of the Landers earthquake, June 28th, 1992,  $M = 7.6$ . An earthquake with  $M = 6.6$  or larger was expected during the 18 months after the Landers earthquake within the 169-km distance from its epicenter (shown by a circle). The Northridge earthquake occurred on January 28th, 1994, 20 days after the alarm expired. After [43]

**Nodes** These are even more densely fractured mosaic structures formed around the intersections and junctions of boundary zones. Their origin is due, roughly saying, to collision of the corners of blocks [16,39,40,55]. The nodes play a singular role in the dynamics of the lithosphere. A special type of instability is concentrated within the nodes and strong earthquakes nucleate in nodes. The epicenters of strong earthquakes worldwide are located only within some specific nodes that can be identified by pattern recognition [19,22].

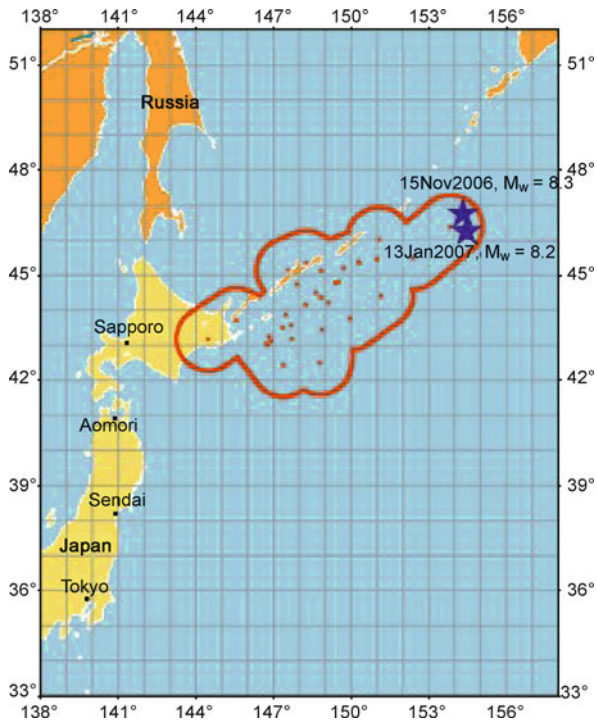
Nodes are well known in the structural geology and geomorphology and play a prominent textbook role in geological prospecting. However their connection with earthquakes is less widely recognized.

The formalized procedure for dividing a territory into blocks  $\Rightarrow$  faults  $\Rightarrow$  nodes is given in [2].

### Fault Network – A Stockpile of Instability

For brevity, the systems of boundary zones and nodes are called here fault networks. They range from the Circum Pacific seismic belt, with the giant triple junctions for the nodes, to interfaces between the grains of rocks, with the corners of grains for the nodes. Their great diversity notwithstanding, fault networks play a similar role in the lithosphere dynamics. Specifically, while tectonic energy is stored in the whole volume of the lithosphere and well beneath, the energy release is to a large extent controlled by the processes in relatively thin fault networks. This contrast is due to the following.

First, the strength of a fault network is smaller than the strength of blocks it separates: fault networks are weakened by denser fragmentation and higher permeability to



Geo-complexity and Earthquake Prediction, Figure 4  
 Prediction of Simushir, Kuril Islands earthquakes, November 15th, 2006,  $M_w = 8.3$  and January 13th, 2007,  $M_w = 8.2$  by Algorithm RTP. An earthquake with magnitude  $M_w \geq 7.2$  is predicted to occur within the time interval from September 30th, 2006, to June 30th, 2007 in the area bordered by the red curve. The red dots show epicenters of an earthquake-forming premonitory chain. The blue stars show epicenters of the predicted earthquakes

fluids. For that reason, tectonic deformations are concentrated in fault networks, whereas blocks move essentially as a whole, with a relatively smaller rate of internal deformations. In other words, in the time scale directly relevant to earthquake prediction (hundreds of years or less) the major part of the lithosphere dynamics is realized through deformation of fault networks and relative movement of blocks.

Second, the strength of a fault network is not only smaller, but also highly unstable, sensitive to many processes there. There are two different kinds of such instability. The “physical” one is originated at the micro level by a multitude of physical and chemical mechanisms reviewed in the next section. “Geometric” instability is originated at a macro level controlled by the geometry of the fault network (Sect. “**Geometric Instability**”). These instabilities largely control dynamics of seismicity, the occurrence of strong earthquakes included.

### “Physical” Instability [23,29]

As in any solid body, deformations and fracturing in the lithosphere are controlled by the relation of the strength field and stress field. The strength is in turn controlled by a great multitude of interdependent mechanisms concentrated in the fault network. We describe, for illustration, several such mechanisms starting with the impact of fluids.

### Rehbinder Effect, or Stress Corrosion [14,85]

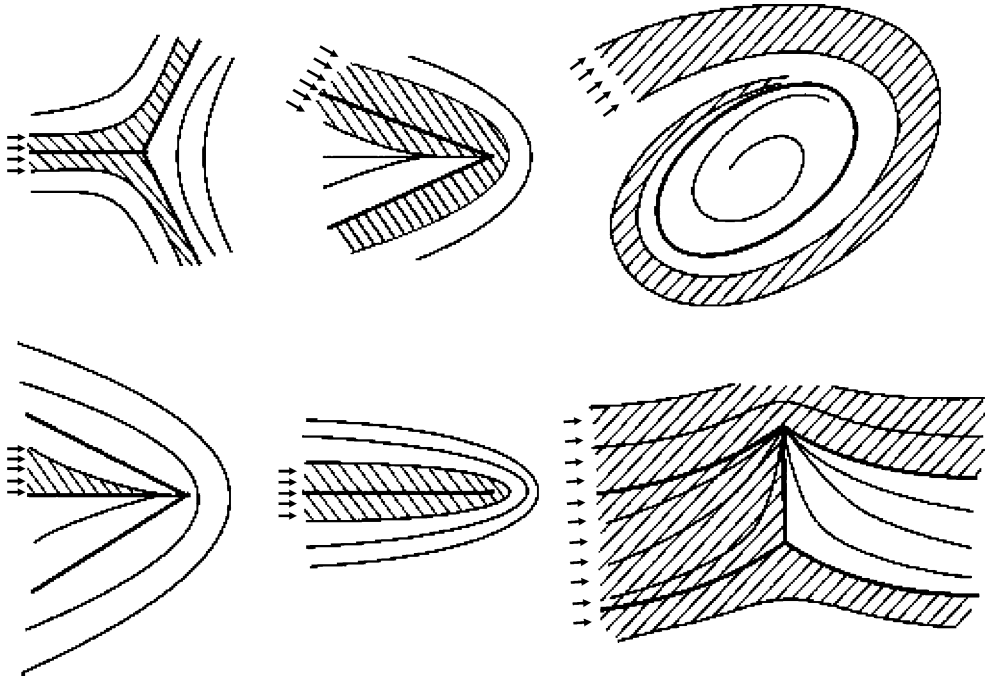
*Mechanism* Many solid substances lose their strength when they come in contact with certain surface-active liquids. The liquid diminishes the surface tension  $\mu$  and consequently the strength, which is proportional to  $\sqrt{\mu}$  by the Griffiths criterion. When the strength drops, cracks may emerge under small stress. Then liquid penetrates the cracks and they grow, with drops of liquid propelling forward, until they dissipate. This greatly reduces the stress required to generate the fracturing. Stress corrosion was first discovered for metals and ceramics. Then such combinations of solid substances and surface-active liquids were recognized among the common ingredients of the lithosphere, e.g. basalt and sulphur solutions. When they meet, the basalt is permeated by a grid of cracks and the efficient strength may instantly drop by a factor of 10 or more due to this mechanism alone.

*Geometry of Weakened Areas* Orientation of such cracks at each point is normal to the main tensile stress. The stress field in the lithosphere may be very diverse. However, the shape of weakened areas where the cracks concentrate may be of only a few types, determined by the theory of singularities. Some examples are shown in Fig. 5, where thin lines show the trajectories of cracks; each heavy line is a separatrix, dividing the areas with different patterns of trajectories.

If a liquid infiltrates from a place shown in Fig. 5 by arrows, the cracks concentrate in the shaded area, and its strength plummets. A slight displacement of the source across the separatrix may strongly change the geometry of such fatigue; it may be diverted to quite a different place and take quite a different shape, although not an arbitrary one. Furthermore evolution of the stress field may change the type of a singularity, make it disappear or create a new one, and the geometry of fatigue will follow suit.

*Stress Corrosion is Highly Sensitive to Geochemistry of Fluids* For example, gabbro and dolerite are affected only in the presence of iron oxides; Kamchatka ultrabasic rocks





Geo-complexity and Earthquake Prediction, Figure 5  
Instability caused by stress corrosion. The geometry of weakened areas depends on the type of singularity and the place where the chemically active fluid comes in. After [14]

are affected by the andesite lava liquids only in the presence of copper oxide, etc. Migration of fluids would cause observable variations of electromagnetic and geochemical fields.

*Summing Up* Stress corrosion brings into lithosphere a strong and specific instability, which may explain many observed premonitory seismicity patterns. However the basic configurations of fatigue, as shown in Fig. 5 might be realizable only in not-too-large areas. This limitation stems from the dissipation of fluids and/or from the inhomogeneity of stress field.

**Other Mechanisms** Boundary zones feature several other mechanisms, potentially as important and certainly as complicated. A few more examples follow.

*Mechanical Lubrication* by fluids migrating through a boundary zone [7]. The ensuing instability will be enhanced by *fingers of fluids* springing out at the front of migration [6].

*Dissolution of Rocks* Its impact is magnified by the *Rikke effect* – an increase of solubility of rocks with pressure. This effect leads to a mass transfer. Solid material is dis-

solved under high stress and carried out in solution along the stress gradient to areas of lower stress, where it precipitates. The Rikke effect might be easily triggered in a crystalline massif at the corners of rock grains, where stress is likely to concentrate.

*Petrochemical Transitions* Some of them tie up or release fluids, as in the formation or decomposition of serpentines. Other transitions cause a rapid drop of density, such as in the transformation of calcite into aragonite. (This would create a vacuum and unlock the fault; the vacuum will be closed at once by hydrostatic pressure, but a rupture may be triggered.)

Instability is created also by sensitivity of dynamic friction to local physical environment [50], mechanical processes, such as multiple fracturing, buckling, viscous flow, and numerous other mechanisms [49,70].

Most of the above mechanisms are sensitive to variations of pressure and temperature.

### Geometric Instability [16]

The geometry of fault networks might be, and often is, incompatible with kinematics of tectonic movements, including earthquakes. This leads to stress accumulation, de-

formation, fracturing, and the change of fault geometry, jointly destabilizing the fault network. Two integral measures of this instability, both concentrated in the nodes, are *geometric* and *kinematic incompatibility* [16].

Each measure estimates the integrated effect of tectonic movements in a wide range of time scales, from seismicity to geodetic movements to neotectonics.

**Geometric Incompatibility** The intersection of two strike-slip faults separating moving blocks. Figure 6 is a simple example of geometric incompatibility. If the movements indicated by arrows in Fig. 6a could occur, the corners A and C would penetrate each other and an intersection point would split into a parallelogram (Fig. 6c). In the general case of a finite number of intersecting faults their intersection point would split into a polygon. Such splitting is not possible in reality; the collision at the corners leads to the accumulation of stress and deformations near the intersection followed by fracturing and changes of fault geometry. The divergence of the corners will be realized by normal faulting.

The expansion of that unrealizable polygon with time,  $S(t) = Gt^2/2$ , measures the intensity of this process. Here,  $S$  is the area of the polygon, determined by the slip rates on intersecting faults;  $t$  is the elapsed time from the collision, and  $G$  is the measure of geometric incompatibility.

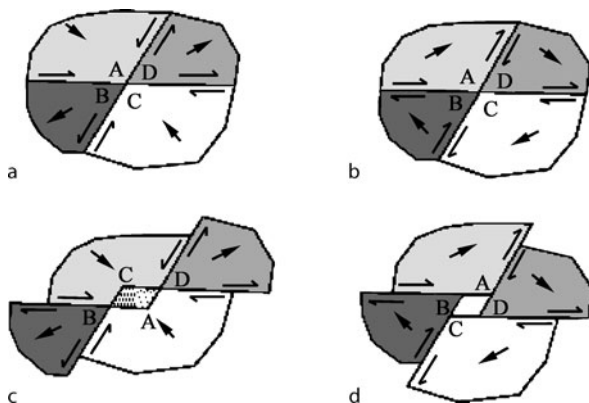
Such incompatibility of structure and kinematics was first described in [55] for a triple junction. The study established a condition under which a single junction can retain its geometry as the plates move, so that the stress will not accumulate. It was suggested in [39,40] that the

general case, when that condition is not satisfied, the ensuing fracturing would not dissolve the stress accumulation, but only redistribute it among newly formed corners. This triggers further similar fracturing with the result that a hierarchy of progressively smaller and smaller faults is formed about an initial intersection. This is a node, recognizable by the dense mosaic structure, with probably self-similar fractal geometry [39].

A real fault network contains many interacting nodes. Incompatibility  $G$  is additive, and can be estimated for a network as a whole. An analogue of the Stokes theorem connects the total value of  $G$  within a territory with observations on its boundary. This removes the nearly impossible task – to take into account complex internal structure of the nodes. One can instead surround the system of nodes by a contour crossing the less complicated areas. Then the geometric incompatibility can be realistically evaluated from the movements of the fewer faults that cross the contour.

Geometric incompatibility in different nodes is interdependent, because they are connected through the movements of blocks-and-faults system. A strong earthquake in a node would redistribute values  $G$  in other nodes thus affecting the occurrence of earthquakes there. Observations indicating the interaction of nodes have been described by [73,74]. These studies demonstrate phenomenon of long-range aftershocks: a rise of seismic activity in the area, where the next strong earthquake is going to occur within about 10 years.

So far, the theory of geometric incompatibility has been developed for the two-dimensional case, with rigid blocks and horizontal movements.



Geo-complexity and Earthquake Prediction, Figure 6  
Geometric incompatibility near a single intersection of faults. a, b initial position of the blocks; c, d extrapolation of the blocks' movement; a, c the locked node: movement is physically unrealizable without fracturing or a change in the fault geometry; b, d the unlocked node. After [16]

**Kinematic Incompatibility** Relative movements on the faults would be in equilibrium with the absolute movements of blocks separated by these faults (one could be realized through the other) under the well known Saint-Venant condition of kinematic compatibility [8,56,57]. In the simplest case, shown in Fig. 6, this condition is  $K = \sum v_i = 0$ , where  $v_i$  are slip rates on the faults meeting at the intersection (thin arrows in Fig. 6). The value of  $K$  is the measure of the kinematic incompatibility, causing accumulation of stress and deformation in the blocks. A simple illustration of that phenomenon is the movement of a rectangular block between two pairs of parallel faults. The movement of the block as a whole has to be compensated for by relative movements on all the faults surrounding it: if, for example, the movement takes place on only one fault, the stress will accumulate at other faults and within the block itself thus creating kinematic incompatibility.

Like geometric incompatibility the values of  $K$  are also additive: one may sum up values at different parts of the network. And an analogue of the Stokes theorem links the value of  $K$  for a region with observations on its boundary.

### Generalization: Complexity and Extreme Events

Summing up, dynamics of the lithosphere is controlled by a wide variety of mutually dependent mechanisms concentrated predominantly within fault networks and interacting across and along the hierarchy. Each mechanism creates strong instability of the strength-stress field, particularly of the strength. Except for very special circumstances, none of these mechanisms alone prevails in the sense that the others can be neglected.

Even the primary element of the lithosphere, a grain of rock, may act simultaneously as a material point, a viscoelastic body, an aggregate of crystals, a source or absorber of energy, fluids, volume, with its body and surface involved in different processes.

Assembling the set of governing equations is unrealistic and may be misleading as well: A well-known maxim in nonlinear dynamics tells that *one cannot understand chaotic system by breaking it apart* [12]. One may rather hope for a generalized theory (or at least a model), which directly represents the gross integrated behavior of the lithosphere. That brings us to the concept that *the mechanisms destabilizing the strength of fault networks altogether turn the lithosphere into a nonlinear hierarchical dissipative system, with strong earthquakes among the extreme events*. At the emergence of that concept the lithosphere was called a chaotic system [29,66,87]; the more general term is *complex system* [20,24,31,53,78,83].

### General Scheme of Prediction

Typically for a complex system, the solid Earth exhibits a permanent background activity, a mixture of interacting processes providing the raw data for earthquake prediction. Predictions considered here are based on detecting premonitory patterns of that activity (Sect. “Definition”).

### Pattern Recognition Approach

Algorithms described here consider prediction as the pattern recognition problem: *Given* the dynamics of relevant fields in a certain area prior to some time  $t$ , *to predict* whether a strong earthquake will or will not occur within that area during the subsequent time interval  $(t, t + \Delta)$ . Some algorithms also reduce the area where it will occur.

In terms of pattern recognition, the object of recognition is the time  $t$ . The problem is to recognize whether it

belongs or not to the time interval  $\Delta$  preceding a strong earthquake. That interval is often called the *TIP* (an acronym for the *time of increased probability* of a strong earthquake). Such prediction is aimed not at the whole dynamics of seismicity but only at the rare extraordinary phenomena, strong earthquakes.

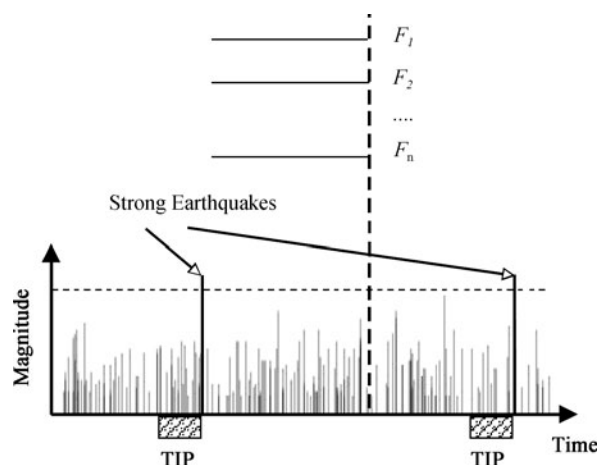
*Pattern recognition of rare events* proves to be very efficient in that approach to prediction. This methodology has been developed by the school of I. Gelfand for the study of rare phenomena of complex origin [9,19,34,71].

### Data Analysis

Prediction algorithms are designed by analysis of the learning material – a sample of past critical events and the time series hypothetically containing premonitory patterns. Analysis comprises four following steps:

1. *Detecting premonitory patterns*. Each time series considered is robustly described by the functionals  $F_k(t)$ ,  $k = 1, 2, \dots$ , capturing hypothetical patterns (Fig. 7). Hypotheses on what these patterns may be are provided by universal modeling of complex systems (Sect. “Fourth Paradigm: Dual Nature of Premonitory Phenomena”), modeling of Earth-specific processes, exploratory data analysis, and practical experience, even if it is intuitive. Pattern recognition of rare events is an efficient common framework for formulating and testing such hypotheses, their diversity notwithstanding.

With a few exceptions the functionals are defined in sliding time windows; the value of a functional is attributed to the end of the window. In the algorithms



Geo-complexity and Earthquake Prediction, Figure 7  
General scheme of prediction. After [29]

described here the time series were earthquake sequences.

2. *Discretization.* Emergence of a premonitory pattern is defined by the condition  $F_k(t) \geq C_k$ . The threshold  $C_k$  is chosen in such a way that a premonitory pattern emerges on one side of the threshold more frequently than on another side. That threshold is usually defined as a certain percentile of the functional  $F_k$ . In such robust representation of the data pattern recognition is akin to exploratory data analysis developed in [86].
3. *Formulating an algorithm.* A prediction algorithm will trigger an alarm when a certain combination of premonitory patterns emerges. This combination is determined by further application of pattern recognition procedures [36,71].
4. *Estimating reliability of an algorithm.* This is necessary, since an algorithm inevitably includes many adjustable elements, from selecting the data used for prediction and definition of prediction targets, to the values of numerical parameters. In lieu of the closed theory a priori determining all these elements they have to be adjusted retrospectively, by predicting the past extreme events. That creates the danger of self-deceptive data-fitting: *If you torture the data long enough, it will confess to anything.* Validation of the algorithms requires three consecutive tests.
  - *Sensitivity analysis:* varying adjustable elements of an algorithm.
  - *Out of sample analysis:* applying an algorithm to past data that has not been used in the algorithm's development.

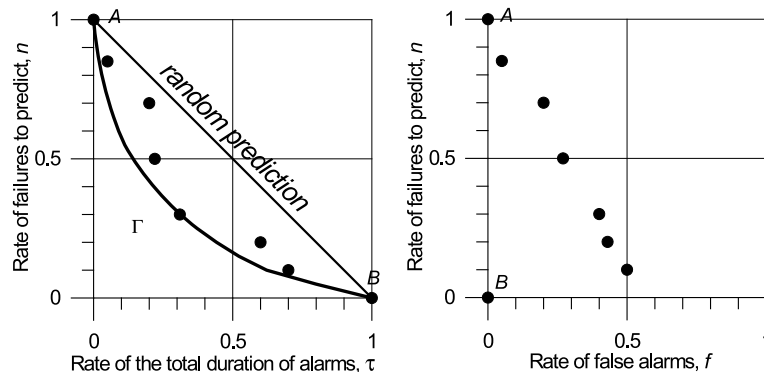
- *Predicting in advance* – the only decisive test of a prediction algorithm.

Such tests take a lion's share of data analysis [17,19,36,93]. A prediction algorithm makes sense only if its performance is (i) sufficiently better than a random guess, and (ii) not too sensitive to variation of adjustable elements. Error diagrams described in the next section show whether these conditions are satisfied.

### Error Diagram

**Definition** An error diagram shows three major characteristics of a prediction's accuracy. Consider an algorithm applied to a certain territory during the time period  $T$ . During the test  $N$  strong earthquakes have occurred there and  $N_m$  of them have been missed by alarms. Altogether,  $A$  alarms have been declared and  $A_f$  of them happened to be false. The total duration of alarms is  $D$ .

Performance of an algorithm is characterized by three dimensionless parameters: the relative duration of alarms,  $\tau = D/T$ ; the rate of failures to predict,  $n = N_m/N$ ; and the rate of false alarms,  $f = A_f/A$ . These three parameters are necessary in any test of a prediction algorithm regardless of a particular methodology. They are juxtaposed on the error diagrams schematically illustrated in Fig. 8. Also called Molchan diagrams, they are used for validation and optimization of prediction algorithms and for joint optimization of prediction and preparedness [59,60,61,62,63]. In many applications parameter  $f$  is not yet considered. In



Geo-complexity and Earthquake Prediction, Figure 8

Scheme of an error diagram. Each point shows the performance of a prediction method: the rate of failures to predict,  $n$ , the relative duration of alarms,  $\tau$ , and the rate of false alarms,  $f$ . Different points correspond to different algorithms. The *diagonal* in the *left plot* corresponds to the random guess. Point *A* corresponds to the trivial optimistic strategy, when an alarm is never declared; point *B* marks the trivial pessimistic strategy, when an alarm takes place all the time; other points correspond to non-trivial predictions. Best combinations  $(n, \tau)$  lie on the envelope of these points  $\Gamma$ . After [63]

early applications they are called ROC diagrams for relative operating characteristics (e. g., [54]).

#### Four Paradigms

Central for determining premonitory patterns is what we know about them a priori. In other words – what are a priori constraints on the functionals  $F_k(t)$  that would capture these patterns (Sect. “Data Analysis”). These constraints are given by the four paradigms described in this section. They have been first found in the quest for premonitory seismicity patterns in the observed and modeled seismicity. There are compelling reasons to apply them also in a wide variety of prediction problems.

*Prehistory.* New fundamental understanding of the earthquake prediction problem was formed during the last 50 or so years, triggering entirely new lines of research. In hindsight this understanding stems from the following unrelated developments in the early sixties.

- F. Press initiated the installation of the state-of-the-art World-Wide Standardized Seismographic Network (WWSSN) later on succeeded by the Global Seismographic Network (GSN). Thus a uniform data base began to accumulate, augmented by expanding satellite observations.
- E. Lorenz discovered deterministic chaos in an ordinary natural process, thermal convection in the atmosphere [51]. This triggered recognition of deterministic chaos in a multitude of natural and socio-economic processes; however, the turn of seismicity and geodynamics in general came about 30 years later [4,29,66,87]. The phenomenon of deterministic chaos was eventually generalized by less rigorously defined and more widely applicable concept of complexity [20, 24,25].
- I. Gelfand and J. Tukey, working independently, created a new culture of exploratory data analysis that allows coping with the complexity of a process (e. g., [19,86]).
- R. Burridge and L. Knopoff [11] demonstrated that a simple system of interacting elements may reproduce a realistically complex seismicity, fitting many basic heuristic constraints. The models of interacting elements developed in statistical physics extended to seismology.
- L. Malinovskaya found a premonitory seismicity pattern reflecting the rise of seismic activity [33]. This is the first reported earthquake precursor formally defined and featuring long-range correlations and world-wide similarity.

With broader authorship:

- Plate tectonics established the connection between seismicity and large-scale dynamics of the lithosphere [41].
- Research in experimental mineralogy and rocks mechanics revealed a multitude of mechanisms that may destabilize the strength in the fault zones [70].

#### First Paradigm: Basic Types of Premonitory Patterns

*The approach of a strong earthquake is indicated by the following premonitory changes in the basic characteristics of seismicity:*

- *Rising:* Seismic activity, earthquakes clustering in space-time, earthquake correlation range, and irregularity of earthquake sequences. Rise of activity sometimes alternates with seismic quiescence.
- *Transforming:* Magnitude distribution (the Gutenberg–Richter relation). Its right end (at larger magnitudes) bends upward, and left end bends downward.
- *Reversing:* territorial distribution of seismicity.
- *Patterns of two more kinds* yet less explored: Rising response to excitation and decreasing dimensionality of the process considered (i. e. rising correlation between its components).

These patterns resemble asymptotic behavior of a thermodynamical system near the critical point in phase transition. Some patterns have been found first in observations and then in models; other patterns have been found in the opposite order. More specifics are given in [15,17,30,31,35,36,67,79,80,83,84,93].

Patterns capturing rise of intensity and clustering, have been validated by statistically significant predictions of real earthquakes [43,65]; other patterns undergo different stages of testing.

#### Second Paradigm: Long-Range Correlations

*The generation of an earthquake is not localized about its future source. A flow of earthquakes is generated by a fault network, rather than each earthquake – by a segment of a single fault. Accordingly, the signals of an approaching earthquake come not from a narrow vicinity of the source but from a much wider area.*

*What is the size of such areas?* Let  $M$  and  $L(M)$  be the earthquake magnitude and the characteristic length of its source, respectively. In the intermediate-term prediction (on a time scale of years) that size may reach  $10L(M)$ ; it might be reduced down to  $3L$  or even to  $L$  in a second approximation [43]. On a time scale of about 10 years that size reaches about  $100L$ . For example, according to [71],

the Parkfield (California) earthquake with  $M$  about 6 and  $L \approx 10$  km “... is not likely to occur until activity picks up in the Great Basin or the Gulf of California”, about 800 km away.

*Historical perspective.* An early estimate of the area where premonitory patterns are formed was obtained in [33] for a premonitory rise of seismic activity. C. Richter, who was sceptical about the feasibility of earthquake prediction, made an exception to that pattern, specifically because it was defined in large areas. He wrote [75]: “... It is important that (the authors) confirm the necessity of considering a very extensive region including the center of the approaching event. It is very rarely true that the major event is preceded by increasing activity in its immediate vicinity.”

However, such spreading of premonitory patterns has been often regarded as counterintuitive in earthquake prediction research on the grounds that earthquakes can't trigger each other at such distances. The answer is that earthquakes forming such patterns do not trigger each other but reflect an underlying large-scale dynamics of the lithosphere. Among the indisputable manifestations of that correlation are the following phenomena: migration of earthquakes along fault zones [47,52,58,90]; alternate rise of seismicity in distant areas [71] and even in distant tectonic plates [76]. Global correlations have been found also between major earthquakes and other geophysical phenomena, such as Chandler wobble, variations of magnetic field, and the velocity of Earth's rotation [34,72]. These correlations may be explained by several mechanisms not mutually exclusive. Such mechanisms range from micro-fluctuations of large scale tectonic movements to impact of migrating fluids (e. g., [1,5,7,10,69,71,82,84,89]).

### Third Paradigm: Similarity

*Premonitory phenomena are similar (identical after normalization) in the extremely diverse environments and in a broad energy range (e. g., [1,33,36]). The similarity is not unlimited however and regional variations of premonitory phenomena do emerge.*

Normalized prediction algorithms retain their prediction power in active regions and platforms, with the magnitude of target earthquakes ranging from 8.5 to 4.5. Furthermore, similarity extends to induced seismicity, and to multiple fracturing in engineering constructions and laboratory samples (e. g., [3,35,43]). Ultimately, a single but explicit demonstration of similarity was obtained for starquakes – ruptures of the crust of neutron star [45], where the conditions are extremely different than in the Earth.

Altogether the corresponding elastic energy release ranges from ergs to  $10^{25}$  ergs (even to  $10^{46}$  ergs if the starquake is counted in).

However, the performance of prediction algorithms does vary from region to region (see [21,35,63]). It is not yet clear whether this is due to imperfect normalization, or to limitations on similarity itself.

### Fourth Paradigm: Dual Nature of Premonitory Phenomena

*Some premonitory patterns are “universal”, common for hierarchical complex systems of different origin; other are specific to geometry of fault networks or to a certain physical mechanism controlling the strength and stress fields in the lithosphere.*

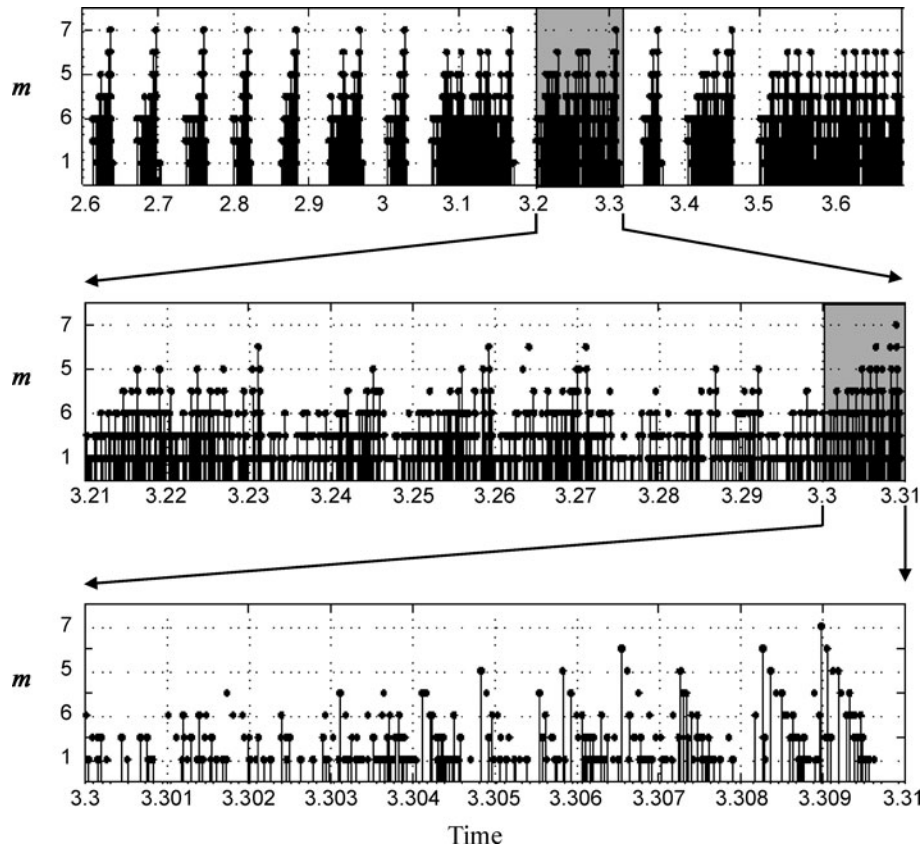
*Universal patterns.* These are most of the patterns so far known. They can be reproduced on models not specific to the Earth only, e. g. models of a statistical physics type (direct or inverse cascade, colliding cascades, percolation, dynamical clustering), models of critical phenomena in fluid dynamics, as well as Earth-specific models themselves.

Complete analytical definition of premonitory patterns was obtained recently on the branching diffusion model [18]. Definition includes only three control parameters, thus strongly reducing uncertainty in data analysis (Sect. “Data Analysis”).

Reviews of such models can be found in [15,17,36,66,83,89,93]. Discussion of particular patterns is given also in [25,42,67,68,88,92].

An example of an earthquake sequence generated by a universal model is shown in Fig. 9 [17]. The modeled seismicity exhibits major features of real seismicity: seismic cycle, switching of seismic regime, the Gutenberg–Richter relation, foreshocks and aftershocks, long-range correlation, and, finally, the premonitory seismicity patterns.

*Earth-specific patterns* are not yet incorporated in prediction algorithms. We discuss here the patterns reflecting the state of the nodes – structures where the strong earthquakes are nucleated (see Sect. “Structure”). Quantitative characteristics of that state are geometric incompatibility  $G$  (Sect. “Geometric Instability”). It shows whether the nodes are locked up or unlocked and quantifies their tendency to fracture and change of the faults geometry. Change of  $G$  might create or dissolve such feature as asperities, relaxation barriers, weak links, and replacement of seismicity by creep or “silent” earthquakes [16]. These features would migrate from node to node with velocity typical of seismicity migration: tens to hundreds km/year [90].



Geo-complexity and Earthquake Prediction, Figure 9

Synthetic earthquake sequence consecutively zoomed. *Shaded areas* mark zoomed intervals. The model shows the rich variety of behavior on different timescales. Note that the ratio of timescales for the *top* and *bottom* panels is  $10^2$ . After [17]

All this makes monitoring of  $G$  highly relevant to detecting premonitory patterns. A simple pattern of that kind is seismic quiescence around the soon-to-break nodes (e. g., [44,58,77]). A simple highly promising possibility is considering separately premonitory phenomena inside and outside of nodes (e. g., [77]).

### Earthquake Prediction and Earthquake Preparedness

Given the limited accuracy of predictions, how do we use them for damage reduction? The key to this is to escalate or de-escalate preparedness depending on the following: content of the current alarm (what and where is predicted), probability of a false alarm, and cost/benefit ratio of disaster preparedness measures. Prediction might be useful if its accuracy is *known*, even if it is not high. Such is the standard practice in preparedness for all disasters, war included.

### Diversity of Damage

Earthquakes hurt population, economy, and environment in very different ways: destruction of buildings, lifelines, etc; triggering fires; releasing of toxic, radioactive and genetically active materials; triggering other natural disasters, such as floods, avalanches, landslides, tsunamis, etc.

Equally dangerous are the socio-economic and political consequences of earthquakes: disruption of vital services (supply, medical, financial, law enforcement, etc.), epidemics, drop of production, slowdown of economy, unemployment, disruptive anxiety of population, profiteering and crime. The socio-economic consequences may be inflicted also by the undue release of predictions.

Different kinds of damage are developing at different time and space scales, ranging from immediate damage to chain reaction, lasting tens of years and spreading regionally if not worldwide.

**Diversity of Disaster Preparedness Measures** Such diversity of damage requires a hierarchy of disaster preparedness measures, from building code and insurance to mobilization of post disaster services to red alert. It takes different times, from decades to seconds to undertake different measures; having different cost they can be maintained for different time periods; and they have to be spread over different territories, from selected sites to large regions. No single stage can replace another one for damage reduction and no single measure is sufficient alone.

On the other hand many important measures are inexpensive and do not require high accuracy of prediction. An example is the Northridge, California, earthquake, 1994, which caused economic damage exceeding \$30 billion. Its prediction, published well in advance [48], was not precise – the alarm covered a time period of 18 months and an area 340 km in diameter with dramatically uneven vulnerability. However, low-cost actions, undertaken in response to this prediction (e. g. an out of turn safety inspection) would be well justified if even just a few percent of the damage were prevented.

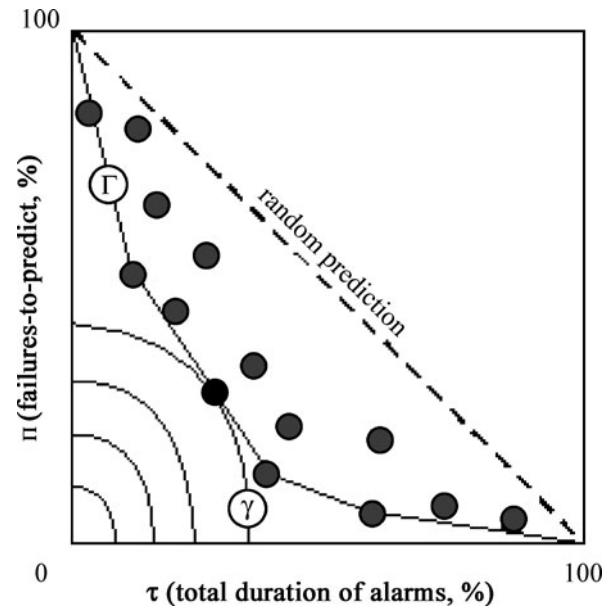
### Joint Optimization of Prediction and Preparedness

The choice of preparedness measures is by no means unique. Different measures may supersede or mutually exclude one another, leaving the decision-maker a certain freedom of choice [38]. The definition of the prediction algorithm is not unique either. The designer of the algorithm has certain freedom to choose the tradeoff between different characteristics of its accuracy (rate of failures to predict, duration of alarms, and rate of failures to predict) by varying adjustable elements of the algorithm (Sect. “General Scheme of Prediction”). That leads to the problem, typical for decision-making with incomplete information: to optimize jointly prediction and preparedness. Figure 10 shows the scheme of such optimization. This figure shows also advantages of a new formulation of prediction: parallel applications of several versions of an algorithm.

Further discussion can be found in [27,28,63,64].

### Further Goals

Particularly encouraging for further earthquake prediction research is the wealth of relevant data, models, and theories that are available and yet untapped (the *want amidst plenty* pattern, Conference and School on Predictability of Natural Disasters for our Planet in Danger. A System View: Theory, Models, Data Analysis, 25 June – 6 July 2007, Trieste, ICTP, [http://cdsagenda5.ictp.it/full\\_display.php?ida=a06204](http://cdsagenda5.ictp.it/full_display.php?ida=a06204)). Likely within reach is a new generation



Geo-complexity and Earthquake Prediction, Figure 10

Joint optimization of prediction and preparedness based on the theory of optimal control. Dots show points on the error diagram.  $\Gamma$  is their envelope. Thin contours ( $\gamma$ ) show loss curves with constant value of a prevented loss. Optimal strategy is the tangent point of contours  $\Gamma$  and  $\gamma$ . After [63]

of prediction algorithms, about five- to ten-fold more accurate than existing ones.

In the general scheme of things, this is a part of wider developments: Emergence of the newly integrated dynamics of the solid Earth, extending from a fundamental concept succeeding plate tectonics to predictive understanding and (with luck) control of geological and geotechnical disasters. And predictive understanding of extreme events (critical phenomena) in the complex systems formed, separately and jointly, by nature and society.

### Acknowledgment

The authors are sincerely grateful for insightful comments of Edo Nyland, William Lee, Michele Caputo, and Antoni Correig.

### Bibliography

#### Primary Literature

1. Aki K (1996) Scale dependence in earthquake phenomena and its relevance to earthquake prediction. Proc Natl Acad Sci USA 93:3740–3747
2. Alekseevskaya MA, Gabrielov AM, Gvishiani AD, Gelfand IM, Ranzman EY (1977) Formal morphostructural zoning of mountain territories. J Geophys 43:227–233



3. Allegre CJ, Le Mouél J-L, Provost V (1982) Scaling rules in rock fracture and possible implications for earthquake prediction. *Nature* 297:47–49
4. Bak P, Chen K, Tang C (1992) A forest-fire model and some thoughts on turbulence. *Phys Lett A* 147:297–300
5. Barenblatt GI (1993) Micromechanics of fracture. In: Bodner ER, Singer J, Solan A, Hashin Z (eds) *Theoretical and Applied Mechanics*. Elsevier, Amsterdam, pp 25–52
6. Barenblatt G (1996) *Scaling, Self-similarity, and Intermediate Asymptotics*. Cambridge University Press, Cambridge
7. Barenblatt GI, Keilis-Borok VI, Monin AS (1983) Filtration model of earthquake sequence. *Trans (Doklady) Acad Sci SSSR* 269:831–834
8. Bird P (1998) Testing hypotheses on plate-driving mechanisms with global lithosphere models including topography, thermal structure, and faults. *J Geophys Res* 103(B5):10115–10129
9. Bongard MM, Vaintsveig MI, Guberman SA, Izvekova ML, Smirnov MS (1966) The use of self-learning programs in the detection of oil containing layers. *Geol Geofiz* 6:96–105 (in Russian)
10. Bowman DD, Ouillon G, Sammis GG, Sornette A, Sornette D (1998) An observational test of the critical earthquake concept. *J Geophys Res* 103:24359–24372
11. Burridge R, Knopoff L (1967) Model and theoretical seismicity. *Bull Seismol Soc Am* 57:341–360
12. Crutchfield JP, Farmer JD, Packard NH, Shaw RS (1986) *Chaos*. *Sci Am* 255:46–57
13. Farmer JD, Sidorowich J (1987) Predicting chaotic time series. *Phys Rev Lett* 59:845
14. Gabriellov AM, Keilis-Borok VI (1983) Patterns of stress corrosion: Geometry of the principal stresses. *Pure Appl Geophys* 121:477–494
15. Gabriellov A, Dmitrieva OE, Keilis-Borok VI, Kossobokov VG, Kuznetsov IV, Levshina TA, Mirzoev KM, Molchan GM, Negmatullaeov SK, Pisarenko VF, Prozoroff AG, Rinehart W, Rotwain IM, Shebalin PN, Shnirman MG, Shreider SY (1986) *Algorithm of Long-term Earthquakes' Prediction*. Centro Regional de Sismologia para America del Sur, Lima
16. Gabriellov AM, Keilis-Borok VI, Jackson DD (1996) Geometric incompatibility in a fault system. *Proc Natl Acad Sci USA* 93:3838–3842
17. Gabriellov AM, Zaliapin IV, Newman WI, Keilis-Borok VI (2000) Colliding cascade model for earthquake prediction. *Geophys J Int* 143(2):427–437
18. Gabriellov A, Keilis-Borok V, Zaliapin I (2007) Predictability of extreme events in a branching diffusion model. *arXiv: 0708.1542 [nlin.AO]*
19. Gelfand IM, Guberman SA, Keilis-Borok VI, Knopoff L, Press F, Ranzman IY, Rotwain IM, Sadovsky AM (1976) Pattern recognition applied to earthquake epicenters in California. *Phys Earth Planet Inter* 11:227–283
20. Gell-Mann M (1994) *The Quark and the Jaguar: Adventures in the Simple and the Complex*. Freeman and Company, New York
21. Ghil M (1994) Cryothermodynamics: the chaotic dynamics of paleoclimate. *Physica D* 77:130–159
22. Gorshkov A, Kossobokov V, Soloviev A (2003) Recognition of earthquake-prone areas. In: Keilis-Borok VI, Soloviev AA (eds) *Nonlinear Dynamics of the Lithosphere and Earthquake Prediction*. Springer, Berlin-Heidelberg, pp 239–310
23. Grotzinger J, Jordan TH, Press F, Siever R (2007) *Understanding Earth*, 5th edn. WH Freeman & Co, New York
24. Holland JH (1995) *Hidden Order: How Adaptation Builds Complexity*. Addison-Wesley, Reading
25. Huang Y, Saleur H, Sammis C, Sornette D (1998) Precursors, aftershocks, criticality and self-organized criticality. *Europhys Lett* 41:43–48
26. Kadanoff LP (1976) Scaling, universality and operator algebras. In: Domb C, Green MS (eds) *Phase Transitions and Critical Phenomena*, vol 5a. Academic Press, London, pp 1–34
27. Kantorovich LV, Keilis-Borok VI (1991) Earthquake prediction and decision-making: social, economic and civil protection aspects. In: *International Conference on Earthquake Prediction: State-of-the-Art. Scientific-Technical Contributions*, CSEM-EMSC, Strasbourg, pp 586–593
28. Kantorovich LV, Keilis-Borok VI, Molchan GM (1974) Seismic risk and principles of seismic zoning. In: *Seismic design decision analysis*. Internal Study Report 43, Department of Civil Engineering, MIT, Cambridge (Mass)
29. Keilis-Borok VI (1990) The lithosphere of the Earth as a non-linear system with implications for earthquake prediction. *Rev Geophys* 28:19–34
30. Keilis-Borok VI (ed) (1990) *Intermediate-Term Earthquake Prediction: Models, Algorithms, Worldwide Tests*. *Phys Earth Planet Inter*, special issue 61(1–2):1–139
31. Keilis-Borok VI (2002) Earthquake prediction: State-of-the-art and emerging possibilities. *Annu Rev Earth Planet Sci* 30:1–33
32. Keilis-Borok VI, Kossobokov VG (1990) Premonitory activation of earthquake flow: algorithm M8. *Phys Earth Planet Inter* 61(1–2):73–83
33. Keilis-Borok VI, Malinovskaya LN (1964) One regularity in the occurrence of strong earthquakes. *J Geophys Res* 69:3019–3024
34. Keilis-Borok VI, Press F (1980) On seismological applications of pattern recognition. In: Allegre CJ (ed) *Source Mechanism and Earthquake Prediction Applications*. Editions du Centre national de la recherche scientifique, Paris, pp 51–60
35. Keilis-Borok VI, Shebalin PN (eds) (1999) *Dynamics of Lithosphere and Earthquake Prediction*. *Phys Earth Planet Inter*, special issue 111(3–4):179–327
36. Keilis-Borok VI, Soloviev AA (eds) (2003) *Nonlinear Dynamics of the Lithosphere and Earthquake Prediction*. Springer, Berlin-Heidelberg
37. Keilis-Borok V, Shebalin P, Gabriellov A, Turcotte D (2004) Reverse tracing of short-term earthquake precursors. *Phys Earth Planet Inter* 145(1–4):75–85
38. Keilis-Borok V, Davis C, Molchan G, Shebalin P, Lahr P, Plumb C (2004) Earthquake prediction and disaster preparedness: Interactive algorithms. *EOS Trans AGU* 85(47), Fall Meet Suppl, Abstract S22B-02a
39. King G (1983) The accommodation of large strains in the upper lithosphere of the earth and other solids by self-similar fault systems: The geometrical origin of b-value. *Pure Appl Geophys* 121:761–815
40. King G (1986) Speculations on the geometry of the initiation and termination processes of earthquake rupture and its relation to morphology and geological structure. *Pure Appl Geophys* 124:567–583
41. Knopoff L (1969) The upper mantle of the Earth. *Science* 163:1277–1287
42. Kossobokov VG, Carlson JM (1995) Active zone size vs. activity:

- A study of different seismicity patterns in the context of the prediction algorithm M8. *J Geophys Res* 100:6431–6441
43. Kossobokov V, Shebalin P (2003) Earthquake Prediction. In: Keilis-Borok VI, Soloviev AA (eds) *Nonlinear Dynamics of the Lithosphere and Earthquake Prediction*. Springer, Berlin-Heidelberg, pp 141–207
  44. Kossobokov VG, Keilis-Borok VI, Smith SW (1990) Localization of intermediate-term earthquake prediction. *J Geophys Res* 95:19763–19772
  45. Kossobokov VG, Keilis-Borok VI, Cheng B (2000) Similarities of multiple fracturing on a neutron star and on the Earth. *Phys Rev E* 61(4):3529–3533
  46. Kravtsov YA (ed) (1993) *Limits of Predictability*. Springer, Berlin-Heidelberg
  47. Kuznetsov IV, Keilis-Borok VI (1997) The interrelation of earthquakes of the Pacific seismic belt. *Trans (Doklady) Russ Acad Sci, Earth Sci Sect 355A(6):869–873*
  48. Levshina T, Vorobieva I (1992) Application of algorithm for prediction of a strong repeated earthquake to the Joshua Tree and the Landers earthquakes' aftershock sequence. *EOS Trans AGU* 73(43), Fall Meet Suppl:382
  49. Sir Lighthill J (ed) (1996) *A Critical Review of VAN*. World Scientific, Singapore-New Jersey-London-Hong Kong
  50. Lomnitz-Adler J (1991) Model for steady state friction. *J Geophys Res* 96:6121–6131
  51. Lorenz EN (1963) Deterministic nonperiodic flow. *J Atmos Sci* 20:130–141
  52. Ma Z, Fu Z, Zhang Y, Wang C, Zhang G, Liu D (1990) *Earthquake Prediction: Nine Major Earthquakes in China*. Springer, New York
  53. Ma S-K (1976) *Modern Theory of Critical Phenomena*. WA Benjamin, Inc, Reading
  54. Mason IB (2003) Binary events. In: Jolliffe IT, Stephenson DB (eds) *Forecast Verification. A Practitioner's Guide in Atmospheric Science*. Wiley, Chichester, pp 37–76
  55. McKenzie DP, Morgan WJ (1969) The evolution of triple junctions. *Nature* 224:125–133
  56. McKenzie DP, Parker RL (1967) The North Pacific: An example of tectonics on a sphere. *Nature* 216:1276–1280
  57. Minster JB, Jordan TH (1984) In: Crouch JK, Bachman SB (eds) *Tectonics and Sedimentation Along the California Margin: Pacific Section*, vol 38. Academic, San Diego, pp 1–16
  58. Mogi K (1968) Migration of seismic activity. *Bull Earth Res Inst Univ Tokyo* 46(1):53–74
  59. Molchan GM (1990) Strategies in strong earthquake prediction. *Phys Earth Planet Inter* 61:84–98
  60. Molchan GM (1991) Structure of optimal strategies of earthquake prediction. *Tectonophysics* 193:267–276
  61. Molchan GM (1994) Models for optimization of earthquake prediction. In: Chowdhury DK (ed) *Computational Seismology and Geodynamics*, vol 1. Am Geophys Un, Washington, DC, pp 1–10
  62. Molchan GM (1997) Earthquake prediction as a decision-making problem. *Pure Appl Geophys* 149:233–237
  63. Molchan GM (2003) Earthquake Prediction Strategies: A Theoretical Analysis. In: Keilis-Borok VI, Soloviev AA (eds) *Nonlinear Dynamics of the Lithosphere and Earthquake Prediction*. Springer, Berlin-Heidelberg, pp 209–237
  64. Molchan G, Keilis-Borok V (2008) Earthquake prediction: Probabilistic aspect. *Geophys J Int* 173(3):1012–1017
  65. Molchan GM, Dmitrieva OE, Rotwain IM, Dewey J (1990) Statistical analysis of the results of earthquake prediction, based on burst of aftershocks. *Phys Earth Planet Inter* 61:128–139
  66. Newman W, Gabrielov A, Turcotte DL (eds) (1994) *Nonlinear Dynamics and Predictability of Geophysical Phenomena*. Am Geophys Un, Int Un Geodesy Geophys, Washington
  67. Newman WI, Turcotte DL, Gabrielov AM (1995) Log-periodic behavior of a hierarchical failure model with application to precursory seismic activation. *Phys Rev E* 52:4827–4835
  68. Pepke GF, Carlson JR, Shaw BE (1994) Prediction of large events on a dynamical model of fault. *J Geophys Res* 99:6769–6788
  69. Pollitz FF, Burgmann R, Romanowicz B (1998) Viscosity of oceanic asthenosphere inferred from remote triggering of earthquakes. *Science* 280:1245–1249
  70. Press F (ed) (1965) *Earthquake Prediction: A Proposal for a Ten Year Program of Research*. Ad Hoc Panel on Earthquake Prediction. White House Office of Science and Technology, Washington, DC, p 134
  71. Press F, Allen C (1995) Patterns of seismic release in the southern California region. *J Geophys Res* 100(B4):6421–6430
  72. Press F, Briggs P (1975) Chandler wobble, earthquakes, rotation and geomagnetic changes. *Nature (London)* 256:270–273
  73. Prozorov AG (1975) Changes of seismic activity connected to large earthquakes. In: Keilis-Borok VI (ed) *Interpretation of Data in Seismology and Neotectonics*. *Comput Seismol* vol 8. Nauka, Moscow, pp 71–82 (in Russian)
  74. Prozorov AG, Schreider SY (1990) Real time test of the long-range aftershock algorithm as a tool for mid-term earthquake prediction in Southern California. *Pure Appl Geophys* 133:329–347
  75. Richter C (1964) Comment on the paper "One Regularity in the Occurrence of Strong Earthquakes" by Keilis-Borok VI and Malinovskaya LN. *J Geophys Res* 69:3025
  76. Romanowicz B (1993) Spatiotemporal patterns in the energy-release of great earthquakes. *Science* 260:1923–1926
  77. Rundkvist DV, Rotwain IM (1996) Present-day geodynamics and seismicity of Asia minor. In: Chowdhury DK (ed) *Computational Seismology and Geodynamics*, vol 1. Am Geophys Un, Washington, DC, pp 130–149
  78. Rundle JB, Turcotte DL, Klein W (eds) (2000) *Geocomplexity and the Physics of Earthquakes*. Am Geophys Un, Washington, DC
  79. Sammis CG, Sornett D, Saleur H (1996) Complexity and earthquake forecasting. In: Rundle JB, Turcotte DL, Klein W (eds) *SFI Studies in the Science of Complexity*, vol XXV. Addison-Wesley, Reading
  80. Shebalin PN, Keilis-Borok VI (1999) Phenomenon of local "seismic reversal" before strong earthquakes. *Phys Earth Planet Inter* 111:215–227
  81. Shebalin P, Keilis-Borok V, Gabrielov A, Zaliapin I, Turcotte D (2006) Short-term earthquake prediction by reverse analysis of lithosphere dynamics. *Tectonophysics* 413:63–75
  82. Soloviev A, Ismail-Zadeh A (2003) Models of Dynamics of Block-and-Fault Systems. In: Keilis-Borok VI, Soloviev AA (eds) *Nonlinear Dynamics of the Lithosphere and Earthquake Prediction*. Springer, Berlin-Heidelberg, pp 71–139
  83. Sornette D (2004) *Critical Phenomena in Natural Sciences: Chaos, Fractals, Selforganization, and Disorder. Concept and Tools*, 2nd edn. Springer, Berlin-Heidelberg
  84. Sornette D, Sammis CG (1995) Complex critical exponents from renormalization group theory of earthquakes: Implications for earthquake predictions. *J Phys I France* 5:607–619

85. Traskin VY, Skvortsova ZN (2006) Rehbinder effect in geodynamical processes. In: Kissin IG, Rusinov VL (eds) *Fluids and Geodynamics*. Nauka, Moscow, pp 147–164 (in Russian)
86. Tukey JW (1977) *Exploratory Data Analysis*. Addison-Wesley Series in Behavioral Science: Quantitative Methods. Addison-Wesley, Reading
87. Turcotte DL (1997) *Fractals and Chaos in Geology and Geophysics*, 2nd edn. Cambridge University Press, Cambridge
88. Turcotte DL (1999) Seismicity and self-organized criticality. *Phys Earth Planet Inter* 111:275–294
89. Turcotte DL, Newman WI, Gabrielov A (2000) A statistical physics approach to earthquakes. In: *Geocomplexity and the Physics of Earthquakes*. Am Geophys Un, Washington, DC
90. Vil'kovich EV, Shnirman MG (1983) Epicenter migration waves: Examples and models. In: Keilis-Borok VI, Levshin AL (eds) *Mathematical models of the structure of the Earth and the earthquake prediction*. Comput Sismol, vol 14. Allerton Press, New York, pp 27–36
91. Vorobieva IA, Levshina TA (1994) Prediction of a second large earthquake based on aftershock sequence. In: Chowdhury DK (ed) *Computational Seismology and Geodynamics*, vol 1. Am Geophys Un, Washington, DC, pp 27–36
92. Yamashita T, Knopoff L (1992) Model for intermediate-term precursory clustering of earthquakes. *J Geophys Res* 97:19873–19879
93. Zaliapin I, Keilis-Borok V, Ghil M (2003) A Boolean delay model of colliding cascades. II: Prediction of critical transitions. *J Stat Phys* 111(3–4):839–861
- Aki K (1981) A probabilistic synthesis of precursory phenomena. In: Simpson DV, Richards PG (eds) *Earthquake Prediction*. An International Review. Manrice Ewing Ser 4. Am Geophys Un, Washington, DC, pp 566–574
- Bolt BA (1993) *Earthquakes – Newly Revised and Expanded*. Freeman, New York
- Bongard MM (1967) *The Problem of Recognition*. Nauka, Moscow (in Russian)
- Conference and School on Predictability of Natural Disasters for our Planet in Danger. A System View: Theory, Models, Data Analysis, 25 June–6 July 2007, the Abdus Salam International Centre for Theoretical Physics, Trieste, [http://cdsagenda5.ictp.it/full\\_display.php?ida=a06204](http://cdsagenda5.ictp.it/full_display.php?ida=a06204)
- Jaumé SC, Sykes LR (1999) Evolving towards a critical point: A review of accelerating seismic moment/energy release prior to large and great earthquakes. *Pure Appl Geophys* 155:279–306
- Kanamori H (1977) The energy release in great earthquakes. *J Geophys Res* 82(B20):2981–2988
- Mandelbrot B (1983) *The Fractal Geometry of Nature*. Freeman, New York
- Ranzman Ela (1979) *Places of Earthquakes and Morphostructures of Mountain Countries*. Nauka, Moscow (in Russian)
- Varnes DJ (1989) Predicting earthquakes by analyzing accelerating precursory seismic activity. *Pure Appl Geophys* 130:661–686
- Wyss M (ed) (1991) *Evaluation of Proposed Earthquake Precursors*. Am Geophys Un, Washington DC
- Wyss M (1997) Second round of evaluation of proposed earthquake precursors. *Pure Appl Geophys* 149:3–16
- Wyss M, Habermann R (1988) Precursory seismic quiescence. *Pure Appl Geophys* 126:319–332
- Zoller G, Hainzl S, Kurths J (2001) Observation of growing correlation length as an indicator for critical point behaviour prior to large earthquakes. *J Geophys Res* 106:2167–2176

### Books and Reviews

- Agnew DC, Ellsworth WL (1991) Earthquake prediction and long-term hazard assessment. *Rev Geophys Suppl* 29:877–889

## GPS: Applications in Crustal Deformation Monitoring

JESSICA MURRAY-MORALEDA  
US Geological Survey, Menlo Park, USA

### Article Outline

Glossary  
 Definition of the Subject  
 Introduction  
 Global Positioning System Measurements  
 Applications of GPS Data to the Study of Seismic  
 and Volcanic Hazards  
 Future Directions  
 Acknowledgments  
 Bibliography

### Glossary

**Confidence ellipse** As applied to a vector representing a displacement or velocity estimate, the confidence ellipse defines the region within which the value is estimated at or above a specified confidence level (e. g., 95%). Confidence ellipses are computed by propagation of errors when computing the position. The ellipse is usually plotted at the tip of a GPS vector (e. g., Fig. 9).

**GNSS** Global Navigation Satellite System. Although this acronym stands for the same phrase as GLONASS, GNSS is a generic term referring to space-based navigation systems like the Global Positioning System (GPS) operated by the U. S., Russia's GLONASS, and the Galileo system under development by the European Union.

**Kinematic GPS** A method of collecting GPS data in which the receiver is continuously or intermittently in motion. This receiver, called the rover, can receive corrections for ambiguity resolution and common errors from a nearby stationary receiver.

#### **Interferometric synthetic aperture radar (InSAR)**

A satellite-based imaging technique in which the satellite emits a radar signal and measures the phase of the returning signal after it has been scattered off the surface of the Earth. The difference in phase of the scattered waves measured during two passes of the same satellite can be used to produce a map of deformation, called an interferogram, that occurred during the time between the two satellite passes.

**International GNSS service (IGS)** An international consortium of agencies worldwide that provide data from

permanent GPS and GLONASS sites in order to generate precise orbital and satellite clock parameters.

**Ionosphere** The electrically charged portion of the atmosphere from ~60 km to ~400 km above sea level. The ionosphere is dispersive, meaning that the degree to which it delays signal propagation depends on the signal's frequency and the electron content of the ionosphere.

**Mega-thrust earthquake** A type of earthquake which causes rupture of a long portion of the interface between a subducting plate and the over-riding plate. These earthquakes involve slip on a huge surface area, making them among the largest on Earth.

**Moment magnitude ( $M_w$ )** A magnitude scale used to compare the energy released in earthquakes. The moment magnitude is computed from the seismic moment. Therefore, because  $M_w$  accounts for the full rupture length of the earthquake, the moment magnitude scale does not saturate for large events in the way that other magnitude scales do.

**Reference frame** A terrestrial reference frame is defined by a set of points on Earth whose coordinates are precisely determined in a coordinate system with a specified origin and orientation of the axes. In order to compare GPS site positions, displacements, or velocities they must all be transformed into the same reference frame. For GPS, the most commonly used reference frame is the International Terrestrial Reference Frame (ITRF) which is updated periodically.

**Rupture** The slip that occurs during an earthquake. This term is often used in discussing the way in which the slip progresses with time over the fault surface, as in "the rupture front propagated southeast."

**Satellite laser ranging (SLR)** A geodetic technique for measuring the position of points on the surface of the Earth. Observation stations emit pulses of light that bounce off retroreflectors on satellites and return to the stations. The stations record the travel time of the light which is used to calculate a range measurement.

**Slip** The distance that material on one side of a fault moves relative to that on the other side.

**Stable North America** The stable interior portion of the North American continent that is not affected by plate boundary deformation. Often this term is used in the context of a "stable North American" reference frame, meaning that GPS velocities are transformed so that the velocities at stations considered to be in the stable interior of the continent are essentially zero. Because of factors such as Glacial Isostatic Adjustment (GIA), even some GPS sites in the continental interior have nonzero velocities. These sites are typically omit-

ted when defining a stable North American reference frame.

**Strainmeter** An instrument that is capable of measuring change in distance over short baselines. These instruments typically come in two forms. The first is installed at the Earth's surface and uses a laser interferometer to measure the changes in distance over baseline lengths of 100s of meters. The second type is installed in a borehole 100s of meters deep to measure subtle changes in the diameter of the borehole. Some borehole strainmeters measure volumetric strain (e. g. the Sacks–Evertson strainmeter) and others measure three independent components of horizontal strain (e. g. the Gladwin tensor strainmeter).

**Strong motion seismograph** Seismic instrument designed to record high-amplitude shaking near an earthquake rupture. These instruments typically record acceleration, and are sometimes called accelerometers. Data recording is often triggered by the arrival of the first seismic waves, and these instruments can record acceleration several times that of gravity.

**Telesismic** Refers to seismic waves recorded at distances greater than 3000 km from the epicenter.

**Troposphere** The portion of the atmosphere from the Earth's surface to ~15 km which delays GPS signal propagation. The degree to which the GPS signal is delayed depends on the spatially and temporally varying atmospheric pressure and water vapor content.

**Very long baseline interferometry (VLBI)** A geodetic positioning technique in which radio signals from distant sources such as quasars received at an array of antennas are used to calculate precise positions.

## Definition of the Subject

The Global Positioning System (GPS) is a space-based Global Navigation Satellite System (GNSS). Using signals transmitted by GPS satellites, the positions of ground-based receivers can be calculated to high precision, making it possible to track the movement of points on the Earth's surface over time. Unlike older geodetic surveying methods which involved periodically measuring angles, distances, or elevations between points, GPS can provide three-component (latitude, longitude, and altitude) position information at a range of sampling rates and on a global scale. GPS equipment is easy to use and can be set up to collect data continuously. Since its early geophysical applications in the mid-1980s, this versatile tool, which can be used to track displacements over time periods of seconds to decades, has become indispensable for crustal

deformation studies, leading to many important insights and some surprising discoveries.

## Introduction

This article focuses on applications of GPS data to the study of tectonic, seismic, and volcanic processes. GPS has become a valuable tool for investigating other types of crustal deformation as well, including landslides (e. g., [25,50,100,128,151]), global sea-level change [150], and the ongoing rebound (termed Glacial Isostatic Adjustment or GIA) of the Earth's crust since the retreat of the ice sheets which covered much of North America and northern Europe during the last ice age (e. g., [19,80,88,95,117,141]), but these topics are beyond the scope of this article. The discussion presented here begins with an overview of how GPS works and how it is used to collect data for geophysical studies. The rest of the paper describes a variety of ways in which GPS data have been used to measure crustal deformation and investigate the underlying processes, as illustrated by examples from the literature. Since GPS is so widely used in geophysical studies, examples of many more applications exist, and the reader is encouraged to explore the literature for more information.

## Global Positioning System Measurements

### How GPS works

The US Department of Defense developed GPS to provide positioning and timing information, primarily for military purposes, that would be available any time of day, anywhere on Earth, regardless of weather conditions. The first GPS satellites were launched in 1978. Soon afterward the Soviet Union developed a similar system, called GLONASS (which, like the generic acronym GNSS, also stands for Global Navigation Satellite System), and more recently the European Space Agency has designed a satellite navigation system called Galileo which, unlike its predecessors, is dedicated to civilian and commercial, rather than military, use. The rest of this article will focus on GPS. The scope of this article permits only a brief overview. Dzurisin [40] gives a broader discussion with a focus on applications in volcanic investigations. Hofmann-Wellenhof et al. [56] give a thorough treatment of the technical details.

The GPS satellite constellation nominally consists of 24 satellites, as well as several spares. The satellites orbit 20,200 km above the Earth with orbital periods of nearly 12 hours, and each passes over a given point on the Earth's surface once per sidereal day (which is about four min-

utes shorter than a solar day). From any point on the Earth's surface, at any given time, from four to ten satellites are above the horizon (and thus potentially visible). Each satellite remains visible for approximately five out of every 12 hours [40].

The idea behind satellite positioning is that one can determine the distance between a receiver on the ground and an orbiting satellite from the time it takes a signal to travel from the satellite to the receiver. This calculation, therefore, requires a means for precise time-keeping according to a universally accepted standard. The US Naval Observatory defines "GPS time," and GPS specifications require GPS time to be within one microsecond of Coordinated Universal Time (UTC). The difference between a satellite's or receiver's internal clock and GPS time (due to clock drift) is termed "clock bias" and is accounted for in processing GPS data.

GPS satellites broadcast signals on two carrier frequencies termed L1 and L2 in the microwave band. The "coarse acquisition" (C/A) code is modulated on the L1 carrier, and the precise (P) code is modulated on both L1 and L2. A navigation message, containing information about the satellite orbits, clocks (the time given by the satellite's clock and information about the difference between that satellite's time and GPS time), and state of health, as well as ionospheric conditions, is modulated on both carriers.

The receiver "locks on" to a satellite by generating a replica of one or more of the codes modulated on the satellite signal it receives and continually cross-correlating the internally generated code with that received from the satellite until the two match. Once it has locked on to the satellite it can obtain the navigation message, determine the signal travel time, and measure the carrier signal. The apparent distance, or "pseudorange," between the GPS antenna and the satellite antenna is then calculated by multiplying the time it takes for the signal transmitted by the satellite to travel to the receiver by the speed of light. The term pseudorange emphasizes that the travel time used in this calculation includes the effects of satellite and receiver clock biases as well as a variety of other error sources, some of which may be mitigated during processing, and is therefore not equivalent to the true geometric range between the satellite and receiver. It is possible for the receiver to measure the code to a precision of about 1% of its length (293 meters for C/A-code and 29.3 meters for P-code), which results in 3 meter and 30 cm precision in calculated pseudorange. However, the P-code is generally encrypted by the military, called "anti-spoofing" or A-S, and therefore a civilian receiver cannot use this code for positioning.

The C/A code is modulated on L1, and thus this carrier can be easily measured once the receiver has locked on to the satellite using the C/A code. Because of A-S, in order to obtain the L2 carrier on which the encrypted P-code is modulated, civilian users must have receivers that apply more sophisticated signal processing techniques (see pp. 81–85 in [56]). The L1 carrier has a wavelength of 19 cm and the L2, 24.4 cm. Since the receiver can measure the carrier signal to 1% of a cycle length, much greater resolution can be achieved using the carrier phase measurements, rather than the code information, to calculate the pseudorange.

When the carrier signal is used, the satellite-receiver distance is calculated by multiplying the number of carrier cycles (which is generally not an integer) between the satellite and the receiver by the wavelength of the carrier signal. However, when a receiver locks on to a satellite, it only can measure the initial fraction of a carrier cycle that it receives. Although it measures the number of full cycles thereafter (which changes as the satellite moves overhead), the receiver has no way of knowing how many full cycles in addition to the initial fraction were between it and the satellite to begin with. This unknown number of cycles is often called the integer ambiguity and will be different for each satellite–receiver pair. In order to take advantage of the more precise positioning that can be achieved using the carrier signal, processing techniques have been developed to address the problem of integer ambiguities (see [40] for an overview and [56] for more detail).

The positions of the satellites at any given time are given by their orbital parameters. This information is transmitted as part of the navigation message, however more precise orbital information, available from the International GNSS Service (IGS), is used in scientific applications. With the satellite positions assumed known, once the distance from the receiver to at least four satellites is measured, the position in three coordinate dimensions (e. g., north, east, and vertical) of a GPS antenna on the ground can be found. Four satellites are necessary in order to solve for the three coordinate positions and the satellite and receiver clock bias. However, positioning accuracy is greatly improved with data from additional satellites as it is then possible to estimate some unknown noise sources.

Although GPS receivers are capable of determining a position in real time using internal software, for scientific applications the data are generally downloaded, and one of several processing software packages is used to obtain much more precise positions. This type of software allows the user more control over the way in which the data are processed, for instance by enabling the use of precise orbital parameters, the fixing of ambiguities, and the

application of sophisticated models for atmospheric delay and variations in the antenna phase center (the part of the antenna that actually receives the GPS signal).

It is possible to reduce or eliminate certain error sources by differencing data. These errors include atmospheric delays to signal propagation that affect neighboring GPS stations (e. g. within 10 s of km of each other) in a similar way, satellite orbital and clock errors that will be common to data from the same satellite recorded by more than one station at the same time, and receiver clock errors that are common to measurements to multiple satellites made by that receiver at the same time.

Other means of addressing error sources exist as well. For example, the ionosphere is dispersive, meaning the delay in signal propagation that it causes depends on the frequency of the signal. Because GPS uses two frequencies, the effect of this delay can be eliminated from the data. Tropospheric delay is addressed during data processing through a combination of models for the “dry” component (which is dependent on atmospheric pressure, temperature, and elevation), and treatment of the “wet” component (which shows large variability depending on water vapor content) as a stochastic parameter to be estimated. Multipath (when the GPS signal bounces off something before reaching the ground antenna) and set-up error (commonly, operator error in measuring the height of the antenna) are mitigated by antenna design, choice of station location, and surveying technique.

The military has used two means of limiting civilian access to GPS positioning capabilities. The first, called selective availability (SA), involved degrading the accuracy of the satellite orbit information broadcast by the satellites and introducing noise into the satellite clock information. This caused an approximately ten-fold increase in positioning error. Using information from a global network of continuously operating GPS stations it is possible to calculate more accurate satellite orbits for use in post-processing of GPS data. During processing, clock errors can be eliminated by differencing data from multiple stations or estimated along with station positions. Therefore, SA did not pose an insurmountable obstacle to scientific users of GPS. In May 2000 the US government discontinued SA, although it reserves the right to reinstate it if deemed necessary. The other means by which the military can reduce civilian access to GPS data is through A-S. A-S is implemented by encrypting the P-code such that only military users can decipher it, thus preventing outside parties from sending phony “GPS” signals that would prevent accurate positioning. A-S began in 1994 and continues to the present, but its effect on the accuracy of GPS positioning for scientific purposes is generally small.

## Methods of GPS Data Collection for Crustal Deformation Studies

For some geological and geophysical applications, typically those which require mapping or recording of sample locations, kinematic GPS methods are used. In these cases, a base station is used to generate corrections that are applied to measurements made by a “roving” antenna which is moved to each site of interest [40]. However, for the crustal deformation applications described in this article, the GPS antenna remains at one location for an extended period of time (usually at least eight hours). Measurements of this type are often classified as “campaign” or “continuous,” depending on the frequency of measurement and the way in which the receiver and antenna are installed. Here I give a brief description of each; Blewitt et al. [13] give a detailed discussion as well.

Before the advent of GPS as a tool for monitoring crustal deformation, other types of geodetic observations were made by periodically measuring triangulation, trilateration and leveling networks using traditional land surveying techniques. The objective was to see how the movement of the Earth’s surface changed over time, since this reflects a variety of geophysical processes. However, these techniques did not produce 3-component position data, but rather measurements of angles, distances, or elevation changes.

The surveying approach in the early years of GPS monitoring (e. g. mid-1980s to mid-1990s) was similar to that of its predecessors in that instruments were deployed temporarily, for several hours at a time, once a year or so. This approach is called “campaign” or “survey-mode” GPS (SGPS). At that time receivers were prohibitively expensive, limiting their wide-spread use. Also, there were fewer satellites in orbit, and one had to schedule surveying to coincide with the time of day for which there would be satellite coverage in the area of interest.

When collecting SGPS measurements one sets up a tripod which will hold the antenna over a marker, typically a benchmark set in the ground (Fig. 1a,b). The use of such a marker enables one to find the correct spot for future measurements months or years later. The antenna is centered using an optical plummet over a point imprinted on the marker to indicate its center (Fig. 1b). One must then measure the distance between the ground (where the benchmark is) and the antenna (where the GPS signal is actually received), and this distance will be used to convert the positions obtained from data processing to the positions of the benchmarks. As you might guess, there is considerable room for error when having to set up the antenna anew every time observations are made.



GPS: Applications in Crustal Deformation Monitoring, Figure 1

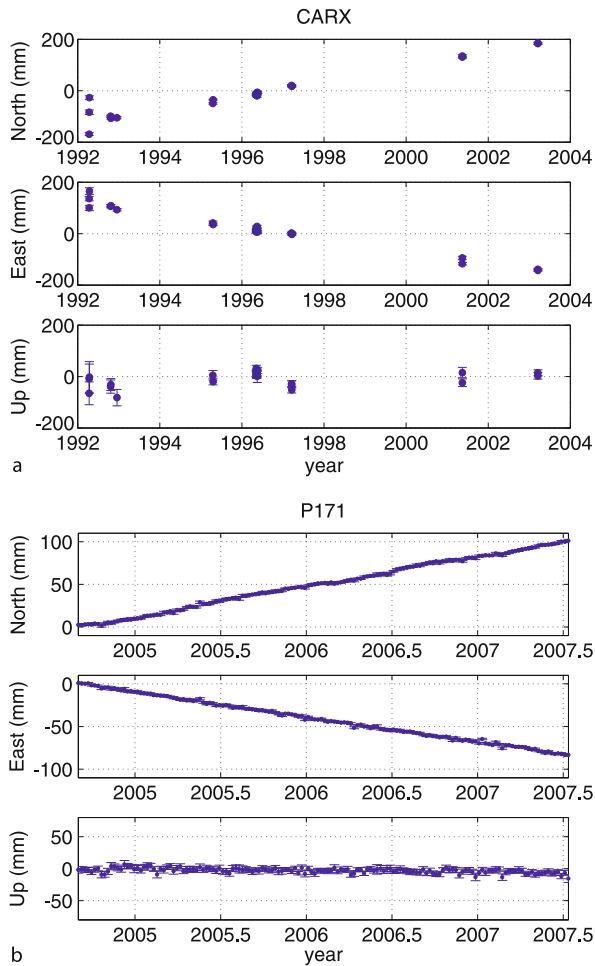
**a** Campaign GPS station located in western Nevada. In the foreground the tripod holds the antenna. The receiver will be stored in one of the boxes in the background while data are collected. Rocks are stacked on the feet of the tripod to prevent it from moving during data collection. (USGS photo) **b** Benchmark that is being observed in **a**. A benchmark often has a stamping telling which agency installed it and when. The cross in the center is the point over which the tripod is centered every time that benchmark is observed. (USGS photo) **c** Continuous GPS station located in Fremont, California. This station, P222, is part of the Plate Boundary Observatory. The antenna, protected from the elements by a domed cover, is supported by a monument with legs driven several meters into the ground. The antenna sends data to the receiver, located along with batteries in the box in the background, via a buried cable. Also visible in the background is the solar panel which provides supplemental power to the site. (UNAVCO photo, reprinted with permission)

Height measurement errors are particularly common. The repeated measurement of the same benchmark over time produces a time series like that in Fig. 2a.

As GPS receivers became less expensive, the possibility of permanently installing continuously recording GPS (CGPS) receivers became a real option. Some of the earliest continuous GPS stations for geophysical studies were installed in Japan [144]. The first continuous GPS sites in southern California were installed in 1990 at Pasadena, Pinyon Flat, and La Jolla [77]. Today several countries have extensive CGPS networks for geophysical monitoring purposes, most notably GEONET which consists of

over 1000 stations across Japan. Regional CGPS networks have existed in the US for many years covering southern California (Southern California Integrated GPS Network, or SCIGN [60]), the San Francisco Bay Area (Bay Area Regional Deformation, or BARD [72]), the Pacific Northwest (Pacific Northwest Geodetic Array, or PANGA [92]), and the Basin and Range (Basin and Range Geodetic Network, or BARGEN [10]; and Eastern Basin-Range and Yellowstone hotspot, or EBRY, <http://www.mines.utah.edu/~rbsmith/RESEARCH/UUGPS.html>). A much larger network consisting of ~850 sites is now underway. This new network, called the Plate Boundary Observatory (PBO),





GPS: Applications in Crustal Deformation Monitoring, Figure 2 Time series of changes in station positions. Both these sites are located in central California, west of the San Andreas fault. The position changes are plotted relative to the center of the North American continent. Because of the relative motion between the North American and Pacific tectonic plates, these sites are moving to the northwest over time, relative to the stable interior of the continent which is not affected by strain accumulation due to the interaction of the two plates. **a** Campaign GPS site CARX near Parkfield, California. **b** Continuous GPS station P171 of the Plate Boundary Observatory network. This station is located just south of Monterey Bay

will provide CGPS coverage for seismically and volcanically active areas throughout the western continental US and Alaska [147].

In addition to providing daily positions, CGPS sites have the advantage of permanent monumentation, thus eliminating set-up error (Fig. 1c). Having daily positions (e.g., Fig. 2b) enables much better estimates of site velocities, but the more frequent data and added precision

means it has become necessary to address additional error sources that were hidden in the noise of the less frequent measurements. These include seasonal signals and time-correlated noise likely due to monument instability, atmospheric effects, reference frame errors, or mismodeled orbits or antenna phase centers [15,33,74,86,166,170]. Despite the proliferation of CGPS stations in recent years, depending on factors such as site access and site condition, financial resources, and scientific goals, campaign-style GPS measurements are still frequently collected.

### Relative Precision

Daily repeatability (or scatter) in the time series for CGPS sites is typically 0.5 to 1.5 mm in the horizontal and  $\sim 3.5$  to 6.5 mm in the vertical. If noise that is common to all stations in the network (due, for example, to errors in satellite orbits or atmospheric delay) has been eliminated, short-term repeatabilities of  $\sim 0.5$  mm in the horizontal and  $\sim 2$  mm in the vertical can be achieved. With 1.5 to 2.5 years of CGPS data, velocity estimates with  $\sim 1$  mm/yr uncertainties are possible [13]. In contrast, SGPS position repeatability is  $\sim 3$  to 5 mm in the horizontal and  $\sim 10$  mm in the vertical [40], and  $\sim 10$  years of SGPS data would be required to reach velocity uncertainty of  $\sim 1$  mm/yr [166].

One way to improve the accuracy of GPS measurements is to extend the length of an observation session, for example from six hours to 24 hours, when estimating daily positions. The additional data mean that multipath noise (which varies throughout the day) averages out better, and it is easier to fix ambiguities and estimate tropospheric delay parameters. The accuracy of GPS measurements has also improved over time. A major reason for this has been the expansion of the global network of tracking stations that are used to calculate precise orbital information for the satellites and to define reference frames for GPS positions. Models for various noise sources have also been refined over time, and some noise sources have been mitigated, for instance by improving the stability of geodetic monuments. It should be noted that the vertical signal is generally much noisier than the horizontal because it is not possible to track satellites below the horizon, and therefore there is no position control from below.

### Applications of GPS Data to the Study of Seismic and Volcanic Hazards

GPS data provide important constraints on the underlying processes that lead to observed deformation, especially when used in combination with other data types. One of the major strengths of GPS is its ability to track positions over time spans of seconds to decades; another is that it en-

ables measurement of position changes over continental-scale baselines. This section will discuss several broad areas of study using GPS data and provide specific examples of each. Background on modeling methods will be included as necessary.

### Plate Motions

Earthquakes occur in response to stresses in the Earth's crust, and these stresses are largely due to the motion of the Earth's tectonic plates. Knowing the velocity at which the plates move gives us insight into the amount of deformation that must be accommodated on plate boundary faults, and what type of earthquakes might be expected. The Earth's plates are generally thought to be rigid, at least in their interiors (away from plate boundaries where plates interact with one another). One approach to measuring global plate motions combines rate information obtained from the age of magnetic reversals recorded in the basalt of the oceanic crust formed at mid-ocean ridges with information regarding the direction of plate motion gleaned from the orientation of transform faults and direction of slip in earthquakes. Because most such studies use a magnetic anomaly that is  $\sim 3$  million years old, the rates of plate motion inferred in this manner represent an average over the time since the Pliocene. The most commonly used plate motion model of this type is called NUVEL-1A [29,30]. This model gives relative velocities for pairs of plates. Under the assumption that the lithosphere has no net rotation with respect to the mantle below, Argus and Gordon [2] developed the NNR-NUVEL-1A model of absolute plate motions, where NNR refers to "no net rotation."

In contrast to approaches based on geologic data, geodetic techniques enable the estimation of present-day plate motions using essentially instantaneous measurements. Two such geodetic methods, Satellite Laser Ranging (SLR) and Very Long Baseline Interferometry (VLBI), were used in the 1980s to measure the positions of points on the Earth's surface. Given a good spatial distribution of sites world-wide, it is possible to use such data to estimate plate velocities. However, expense and practical considerations limited the number of sites that could be observed with SLR and VLBI. The advent of GPS provided a cost-effective alternative enabling precise three-component positioning with dense spatial coverage globally (e. g., Fig. 3).

Several studies have used GPS data, either alone or in combination with other observations, to estimate global plate motions (e. g., [4,79,124,140]), and numerous other studies have used such data in analyses focused on sub-

groups of plates. The plates are modeled as rigid, rotating, spherical caps. Their motions are expressed as Euler vectors, defined by a location of the Euler pole (point E in Fig. 4a) and the angular velocity of the plate around that pole. These parameters are related to the velocities of the GPS sites by

$$\mathbf{v} = \boldsymbol{\Omega} \times \mathbf{r} \quad (1)$$

where  $\mathbf{v}$  is the GPS velocity vector,  $\boldsymbol{\Omega}$  is the Euler vector for the plate in question, and  $\mathbf{r}$  is the position vector of the GPS site in Cartesian geocentric coordinates (e. g., [79]). The relationship given in (1) may be understood from Fig. 4. The Euler vector and position vector may be expressed as unit vectors ( $\boldsymbol{\Omega}_u$  and  $\mathbf{r}_u$ ) multiplied by their magnitudes ( $\omega$  and  $R$ )

$$\boldsymbol{\Omega} = \omega \boldsymbol{\Omega}_u \quad (2)$$

$$\mathbf{r} = R \mathbf{r}_u \quad (3)$$

where  $\omega$  is the velocity of rotation (generally in degrees or radians per million years), and  $R$  is the radius of the Earth (Fig. 4a). As shown in Fig. 4b, the distance,  $\mathbf{d}$ , that a GPS station travels during a time period  $t$  due to rotation about the Euler pole is given by

$$\mathbf{d} = \omega t R \sin \delta \quad (4)$$

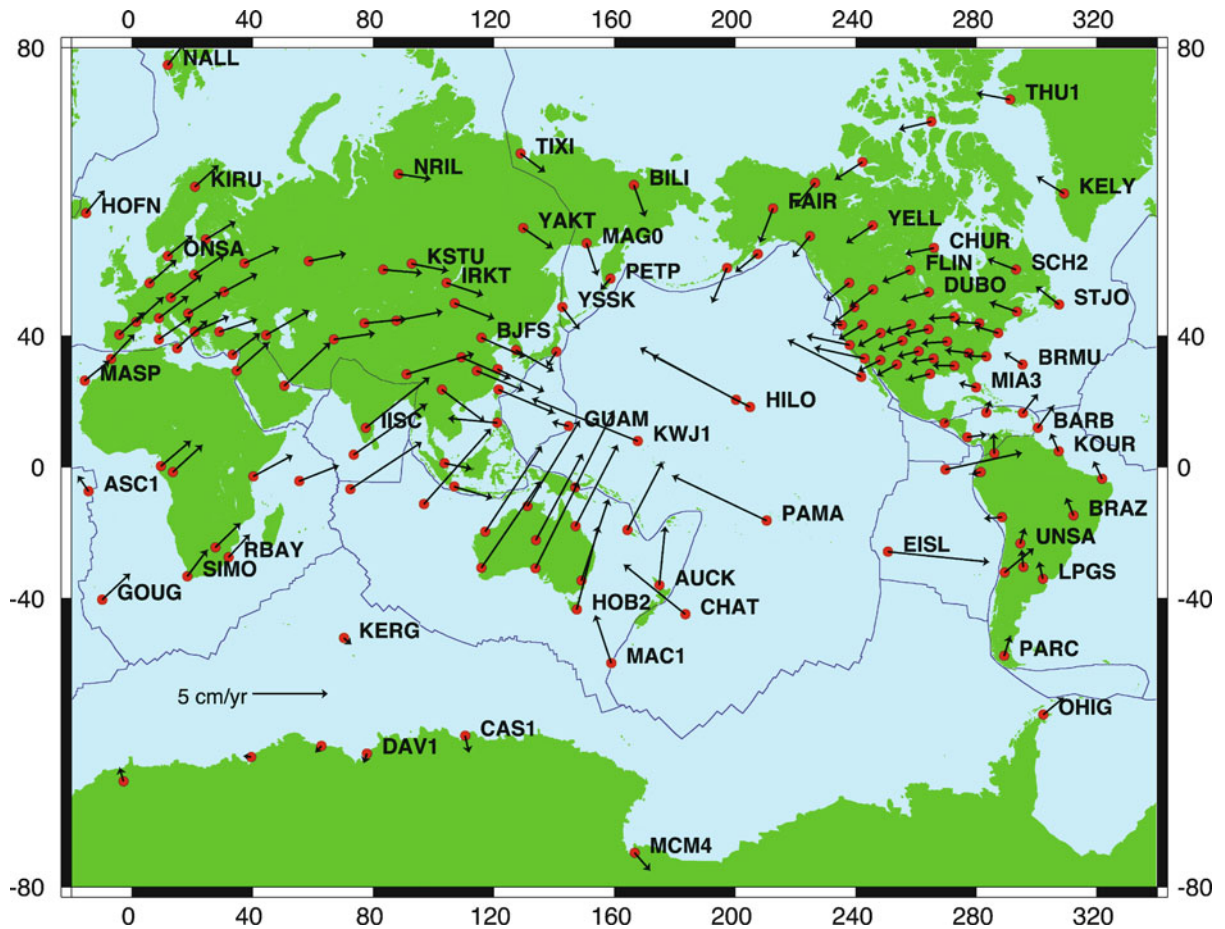
where  $\delta$  is the angular distance between  $\boldsymbol{\Omega}$  and  $\mathbf{r}$  (Fig. 4a). The quantity  $\sin \delta$  is simply the cross product of the unit vectors for the Euler vector and the position vector,

$$\sin \delta = \boldsymbol{\Omega}_u \times \mathbf{r}_u . \quad (5)$$

Therefore, plugging (2), (3), and (5) into (4), and dividing by time gives (1). The system of equations given in (1) can be solved to estimate the Euler vector that best fits the GPS data for each plate.

The early studies (e. g., [4,79]) were able to estimate velocities for only a handful of plates (e. g., six and eight, respectively). In recent years, however, the global distribution of CGPS sites has grown substantially such that Prawirodirdjo and Bock [124] estimated the velocities of 17 "major and minor tectonic plates" and Sella et al. [140] considered 19 "plates and continental blocks." The increasing precision of GPS measurements has made it possible to more rigorously test plate rigidity and the existence of purported plate boundaries, as well as address the potential systematic velocity error introduced by GIA in North America and Eurasia [124,140].

Velocities of GPS sites used in plate motion studies have repeatedly been found to be consistent with the as-

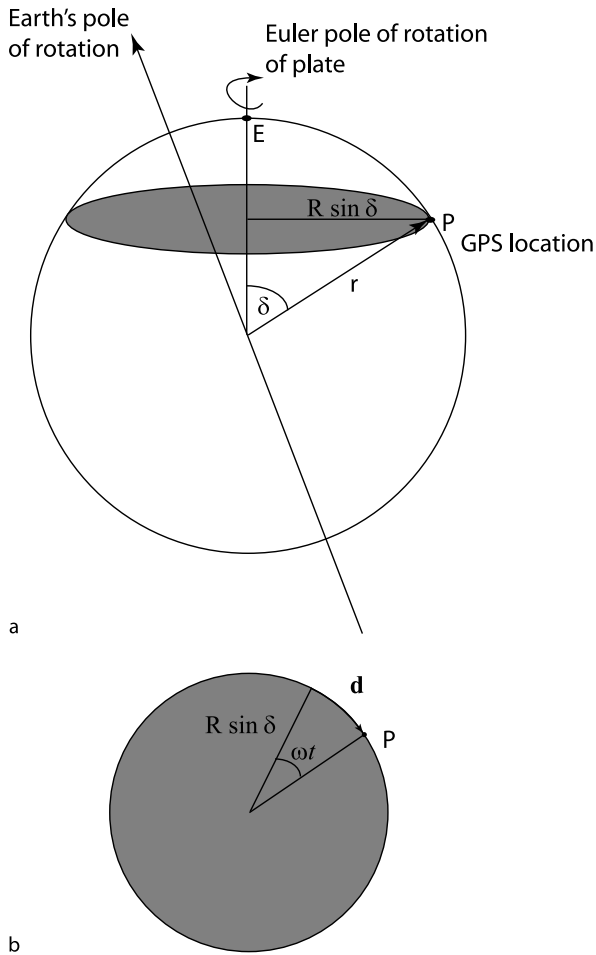


GPS: Applications in Crustal Deformation Monitoring, Figure 3  
 Velocities for a globally distributed selection of GPS sites. Figure courtesy NASA/JPL-Caltech

sumption that plates behave rigidly, as evidenced by the fact that in general a single Euler pole fits the GPS velocities for a given plate well. Moreover, estimates of global plate velocities inferred from these data have agreed with plate motions inferred from geologic and other data (e. g., NUVEL-1A). However, the improvement in size and quality of GPS datasets has highlighted some discrepancies between NUVEL-1A and the models arrived at geodetically (e. g., [140]). These differences may be due in some cases to systematic errors in the rates estimated from magnetic anomaly data in tectonically complex areas where relative plate motion is not localized at the spreading center. In other cases the difference may reflect an actual change in the rate of relative plate motion over the past 3 million years. Refining both the geologic and geodetic plate motion models continues to be an area of active research (e. g., [3,28]).

## Earthquake and Volcano Source Modeling

**Source Potency and Geometry** Although the Earth's plates behave rigidly, as evidenced by the velocities of GPS sites in the stable interiors of plates, at their edges neighboring plates interact, colliding, diverging, or sliding past each other. These processes, as well as others such as the movement of magma underground, impart stress to the Earth's crust, distorting the shape of a volume of crustal material. Some of this deformation is permanent, leading for example to mountain building. However, the brittle upper portion of the Earth's crust deforms elastically, meaning that a large portion of crustal deformation is recoverable; once the stress is relieved, the crustal material returns to its pre-stress shape. The discussion presented here will focus primarily on modeling which assumes an elastic crust, however several studies have considered



GPS: Applications in Crustal Deformation Monitoring, Figure 4 Relationship between the velocity of a GPS station and an Euler pole of rotation. Point P is the location of a GPS station, and point E is the location of the Euler pole of rotation,  $\Omega$ , that describes the motion of the plate on which point P sits. The Euler vector points from the center of the Earth to point E, and its magnitude is the rate of rotation,  $\omega$ . Point P is located at an angular distance of  $\delta$  from point E. The velocity recorded at point P due to rotation about the Euler pole is given by  $v = \Omega \times r$ , where  $r$  is the position vector of the GPS station, and its origin is the center of the earth. Therefore, the length of  $r$  is the radius of the earth,  $R$ . The gray circle that passes through the point P has radius  $R \sin \delta$ . **b** View of the gray circle in a looking along the direction of the Euler vector. The distance traveled by point P during a time period  $t$  due to the rotation given by  $\Omega$  is given by  $d$ . See text for details

the effects of nonelastic material properties (e.g., [37, 65, 135]).

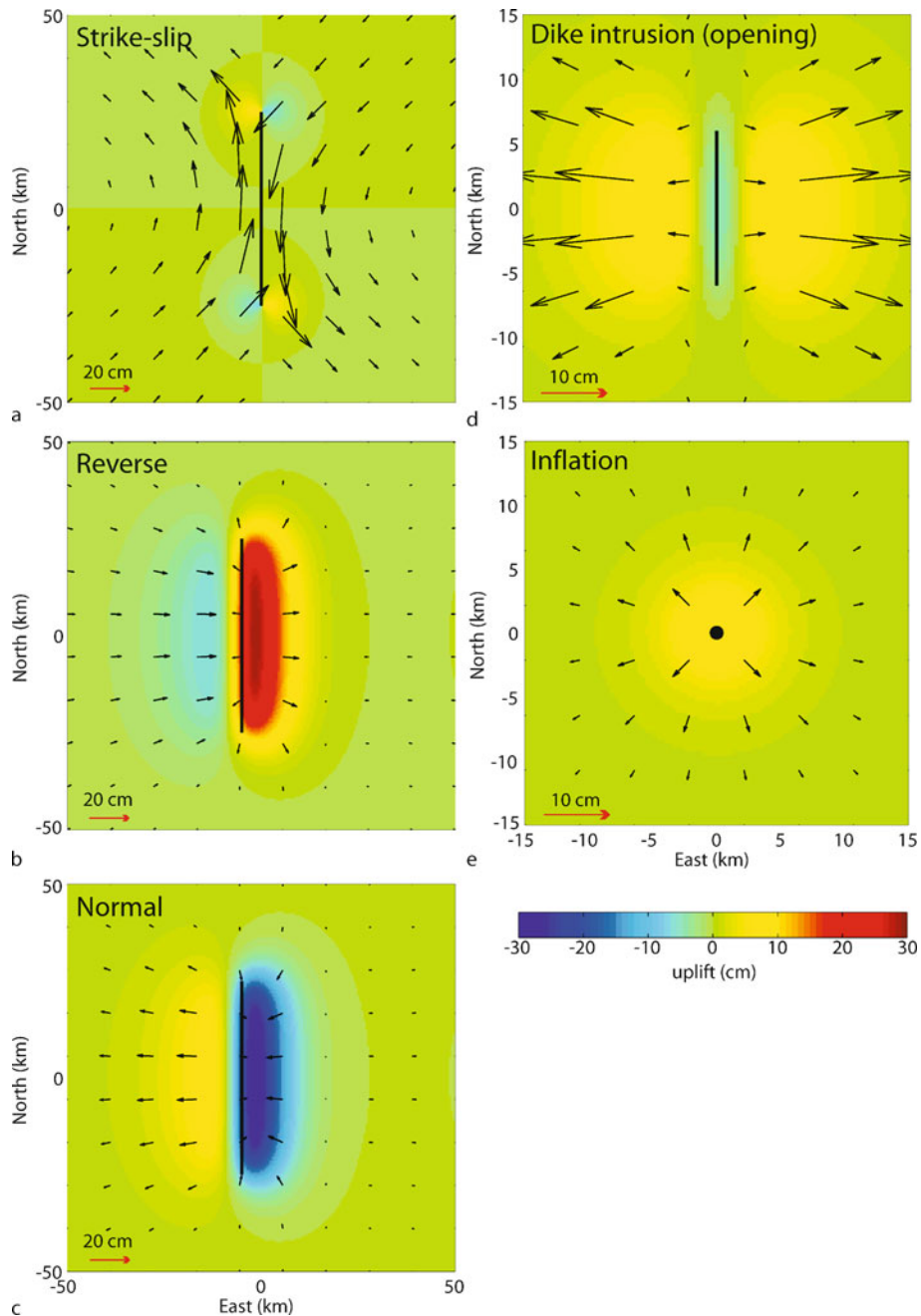
Mathematical expressions from continuum mechanics that describe the stress, strain, and displacement in an elastic solid due, for example, to a point source or to movement on a planar dislocation can be applied to the study

of crustal deformation sources using surface displacement data like those provided by GPS. Okada [109,110] presented concise analytic expressions that are used in many such studies today. Mogi [99] discussed the special case of surface displacement due to three orthogonal, equal-amplitude point sources of inflation which could represent an inflating or deflating magma body at depth. Dislocation sources are described by their dimensions, orientation, location in the Earth's crust, amount of movement (e.g. slip) which takes place across them, and the direction that the material on one side of the dislocation moves relative to that on the other (sometimes termed the "sense of slip"). Volcanic sources such as dikes and sills can be modeled by dislocations with opening rather than slip. Magma chambers are often represented by inflation sources defined by their locations, amount of inflation, and, in the case of more complicated geometries such as ellipsoidal sources, their shapes. Dzurisin [40] gives a good discussion of approaches for modeling a variety of volcanic deformation sources.

The term "source geometry" refers to all source characteristics except the amount of slip, opening, or inflation. These latter three parameters, which describe the strength of the source, are sometimes referred to as the "source potency." The surface displacement field produced in any deformation event reflects not only the source potency but also the source geometry and characteristics of the crustal material. Figure 5 presents the expected horizontal and vertical displacement due to different modes of shear slip and opening on a planar dislocation and to a point source of inflation. As can be seen, each deformation source produces characteristic surface displacement patterns. As with the estimation of plate motions described earlier, a system of equations can be written that relates a deformation source such as a dislocation in the crust to the displacements measured with GPS at the Earth's surface. In simplified terms this system of equations can be written as

$$\mathbf{d} = \mathbf{G}\mathbf{s} \quad (6)$$

where  $\mathbf{d}$  is a vector of station displacements measured by GPS,  $\mathbf{s}$  is a vector of source potency (e.g., fault slip), and  $\mathbf{G}$  is a matrix which embodies the mathematical expressions relating potency to displacements for an assumed fault geometry and elastic properties. This system of equations can be solved (or "inverted") to estimate the unknown potency that best fits the known displacements. The displacements at the Earth's surface are nonlinearly related to the source geometry, but are linearly related to the source potency. Therefore, when the source geometry is known, inverting for the potency is a linear inverse problem. In the simplest case the potency can be assumed to be uniform

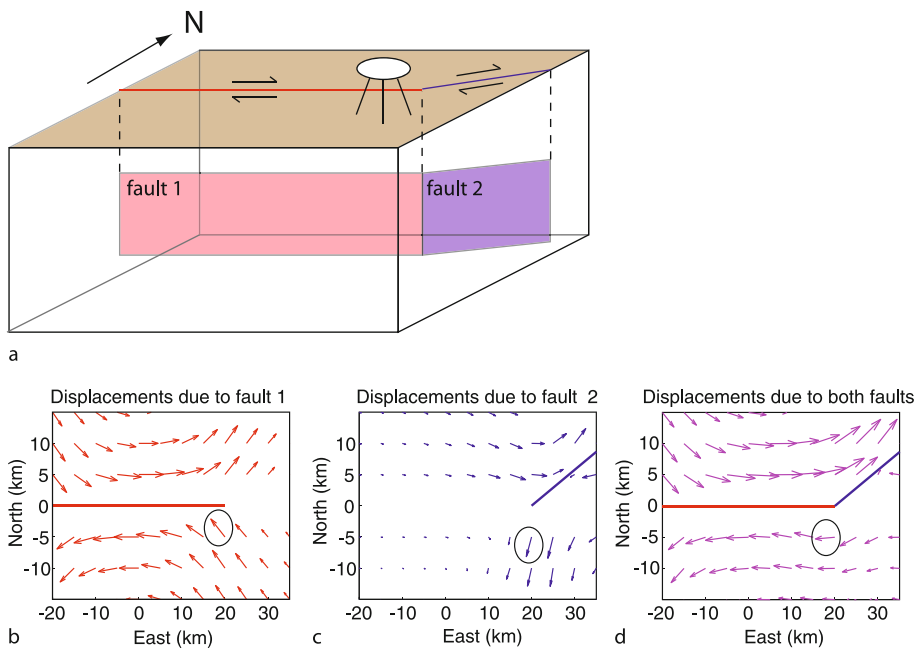


GPS: Applications in Crustal Deformation Monitoring, Figure 5

Predicted displacement due to dislocation and inflation sources assuming a homogeneous elastic half space. Vectors show horizontal displacement and *colored background* shows vertical motion. Note change of vector and spatial scales from a – c to d, e. *Heavy black line* in a – d is surface projection of upper edge of source dislocation. *Black circle* in e is surface projection of point source of inflation. See Table 1 for source parameters

GPS: Applications in Crustal Deformation Monitoring, Table 1  
Source parameters for deformation sources depicted in Fig. 5

Panel in Figure 5	Description	Length (km)	Width (km)	Depth to top (km)	Dip (degrees)	Slip or opening (m)	Sense of slip	Inflation ( $10^6 \text{ m}^3$ )
a	Dislocation	50	14	0	90	1	Right lateral, strike slip	
b	Dislocation	50	10	6.34	60	1	Reverse	
c	Dislocation	50	10	6.34	60	1	Normal	
d	Dislocation	12	5	3	90	1	Opening	
e	Inflation point source			5				10



GPS: Applications in Crustal Deformation Monitoring, Figure 6

Predicted displacements due to two sources. **a** Source geometry consisting of two vertical strike-slip faults and location of a GPS receiver. **b** Map view of displacement due to slip on the fault shown in red in **a**. **c** Map view of displacement due to slip on the fault shown in blue in **a**. **d** Map view of displacement due to slip on both faults. Note how the displacement at the circled locations in **b** and **c** due to the individual faults is very different from that which would be recorded by GPS in **d** due to slip on both faults

for the deformation source. In the case of a fault, this implies that the same amount of slip occurred everywhere on the fault, and the vector  $s$  would have just one element. However, in the presence of multiple sources, the total displacement at the surface is simply the sum of the contributions from all the sources (e. g., Fig. 6). This means that spatially variable fault slip can be estimated by dividing the model fault into a grid of subfaults, each of which contribute to the observed displacements, and estimating the slip on each subfault. In this case the length of the vector  $s$  would be the number of subfaults. The number of subfaults used is generally dictated by how much data are available.

Inversion of geodetic data for characteristics of deformation sources is underdetermined, meaning that a large number of source models can fit the data within errors. Having vertical and horizontal displacement measurements improves the ability to distinguish among different possible source geometries. For instance, several types of volcanic sources will produce similar patterns of vertical deformation, but with the inclusion of horizontal displacement measurements it is possible to differentiate among them. When using GPS data to estimate the spatial distribution of slip on a fault (by dividing the model fault into subfaults and estimating the slip on each), spatial smoothing is often used to provide added constraints in the inver-

sion. The justification for this is that abrupt changes in the amplitude of slip would result in high stresses on the fault surface, which is physically unlikely. The relative weight given to fitting the data and to spatial smoothness is often determined empirically (e. g., [161]). Non-negativity may also be applied, for instance to include prior knowledge about the sense of slip on a fault as a constraint on fault slip estimates.

Regardless of these means for regularizing inversions, because geodetic measurements are collected at the surface, their sensitivity to the details of a deformation source decreases with depth. GPS measurements will be most sensitive to source processes occurring in the upper few kilometers of the Earth's crust near the GPS receiver's location. Deeper sources will affect GPS sites over a broader region, but the recorded deformation signal will lack detail about the source. This can be best understood if one thinks of an earthquake that causes rupture of the Earth's surface. A GPS receiver near the fault will record data which primarily reflect the shallow slip and surface rupture close to that receiver. A receiver 15 km away from the fault will not be sensitive to the shallow, near-fault deformation, and the recorded signal will be due to large-scale features of slip on deeper parts of the fault. Likewise, a receiver at the summit of a volcano can record the movement of magma that is collecting near the crater, whereas receivers lower on the flanks of the volcano will likely not record that signal but could be expected to track deformation due to movement of magma at greater depths.

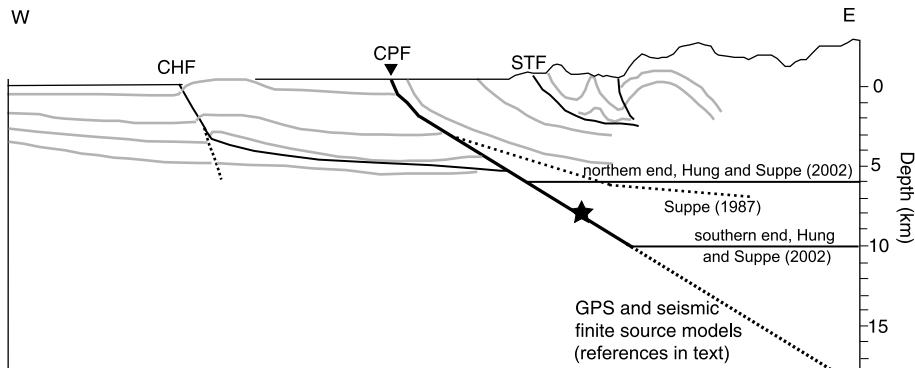
The structure of major faults such as the San Andreas has been studied extensively by mapping, imaging the spatial distribution of background seismicity, and applying geophysical techniques that use, for example, seismic reflection and refraction, gravity, or magnetic data to highlight contrasts in rock properties. Therefore, if an earthquake hypocenter is found to be located on a major fault, the source geometry may be well-known a priori. In the absence of such information, or to refine the source geometry used in inversions, the spatial distribution of aftershocks and the location and extent of any surface rupture are also used.

Traditionally, spatially sparse geodetic measurements were assumed to be insensitive to the details of the model fault geometry used in inversions. With the recent growth of spatially dense GPS networks, however, more physically realistic fault geometries have been required. Maerten et al. [84] showed that using non-planar fault geometries that better represented independent information, for example from surface rupture mapping, led to significant improvement in fits to the GPS displacements for the 1999  $M_w$  7.1 Hector Mine earthquake. Methods for precisely re-

locating seismicity (e. g., [163]) have illuminated fine-scale fault structures that were previously obscured in less-precise catalog locations. This has enabled the development of more realistic model fault geometries.

For example, Murray and Langbein [101] used displacements measured with GPS to estimate the slip distribution of the 2004  $M_6$  Parkfield, California earthquake. This event took place on the well-studied San Andreas fault in central California. Earlier work (e. g., [41,149]) had suggested that the fault was essentially vertical, with a strike of  $149^\circ$ . During the 2004 event, the coseismic displacement recorded for one GPS station, CARH, located within about 500 meters of the fault was in the opposite direction to what is predicted for a right lateral strike-slip fault (e. g., Fig. 5a). In order to fit the data for this station, an additional fault structure was needed. A sub-parallel fault called the Southwest Fracture Zone (SWFZ) which had exhibited movement in a previous earthquake at this locale was a likely candidate. Using relocated aftershocks of the 2004 event [159] as a guide, Murray and Langbein [101] developed a non-planar fault geometry that consisted of the primary San Andreas fault and a subsidiary SWFZ. These two model fault surfaces passed through the relocated seismicity and intersected the mapped surface traces of the faults at the Earth's surface. Using this fault geometry, Murray and Langbein [101] inverted the GPS data to image the coseismic and postseismic slip associated with this earthquake.

Unlike in the case of Parkfield described above, often very little is known a priori about the geometry of a deformation source. This is especially true for sources that lie completely underground such as magmatic intrusions and earthquakes that do not cause any surface rupture (e. g., "blind" thrust events). However, as discussed earlier (Figs. 5 and 6), the surface displacement field produced in any deformation event reflects not only the source potency but also the source geometry. Therefore, by observing the spatial pattern of displacement using GPS, it is possible to infer what type of source lies underground by finding the source model that best predicts the observed data, for instance by using the expressions for deformation in an elastic material. Since surface displacements are nonlinearly related to source geometry, parameters describing the geometry cannot be estimated using linear inversion techniques but rather must be found through nonlinear optimization. Cervelli et al. [21] give a good overview of several approaches to this type of problem. When estimating the geometry and potency of an earthquake or opening source, often a two-step approach is employed: first the source geometry is estimated assuming uniform fault slip or dike opening, and then the inferred geometry is held



GPS: Applications in Crustal Deformation Monitoring, Figure 7

Cross section of the Chelungpu fault in Taiwan. The *solid, near-horizontal black lines* and the *upper dotted line* represent the ramp and décollement structure of the “thin-skinned” model, while the *steeper dotted line* below 10 km is the deeper extension of the Chelungpu fault envisioned by the “thick-skinned” model. CHF, Changhua fault; CPF, Chelungpu fault; STF, Shuangtung fault; *star*, hypocenter of 1999 Chi-Chi earthquake. Reprinted from [65] with permission from Elsevier

fixed and the spatial distribution of slip or opening is estimated. As described in the following examples, the ability to infer the source geometry can help answer important questions about the underlying processes driving deformation such as whether a fault terminates in a décollement or how magma sources interact.

(a) *Deep Structure of the Chelungpu Fault, Taiwan* Johnson and Segall [65] used GPS-measured displacements caused by the 1999  $M_w$  7.6 Chi-Chi Taiwan earthquake to constrain the geometry of its rupture surface. They then went on to answer fundamental questions about the seismotectonics of Taiwan, which has formed due to the collision between the Philippine Sea plate and the Eurasian plate.

The extensive surface rupture that accompanied the Chi-Chi earthquake suggested that oblique reverse / left-lateral slip occurred on the previously known Chelungpu thrust fault which strikes north-south and dips  $\sim 30^\circ$  east. A nearly horizontal décollement structure has been interpreted to exist at depths of 6 to 10 km beneath much of Taiwan based on geologic mapping, seismicity locations, and seismic reflection profiles. In one proposed deformation model, termed the “thin-skinned” model, thrust faults like the Chelungpu fault intersect the décollement at depths of  $< 10$  km (Fig. 7). The alternative “thick-skinned” model interprets seismic and gravity data to suggest that, although a décollement may have controlled the long-term tectonic evolution of the island, more recent deformation has taken place on down-dip extensions of thrust faults with the same dip as their shallower portions (Fig. 7). In the case of the Chi-Chi earthquake, aftershocks occurred both at depths which might coincide with a décollement

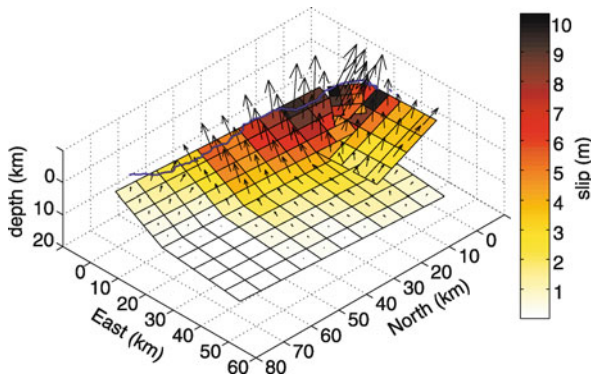
and considerably deeper, potentially on the down-dip extension of the Chelungpu thrust fault.

Johnson and Segall [65] optimized the source geometry of the Chi-Chi earthquake using a half-space with laterally and vertically varying shear modulus. They showed that the GPS data required slip on a thrust plane that transitions into an essentially horizontal décollement structure at  $\sim 8$  km depth. Moreover, they showed that displacements at the north end of the fault, where the surface rupture changed to a more east-west orientation, could only be fit by an additional thrust fault, which they term a “lateral ramp,” extending to the depth of the décollement. The estimated slip distribution of the earthquake using this geometry is shown in Fig. 8. These authors use the results for the Chi-Chi event as the basis for a conceptual model in which deformation follows the thin-skinned model, and lateral ramps form north of the Chelungpu fault due to this fault’s orientation oblique to the direction of plate convergence.

(b) *Magma Plumbing System at Kilauea Volcano, Hawaii* Kilauea volcano, on the Big Island of Hawaii, consists of a summit crater as well as two rift zones, themselves consisting of several craters, extending from the summit down the flanks of the volcano (Fig. 9). Since 1983 Pu’u O’o, a collection of volcanic vents in the East Rift Zone, has been the center of eruptive activity, apparently fed by magma flowing through lava tubes from Kilauea summit.

In January 1997 a fissure eruption occurred on the East Rift Zone at Napau crater,  $\sim 3$  km closer to the summit than Pu’u O’o. In addition to uplift, the GPS instruments recorded horizontal displacements around the eruptive fissure at Napau Crater that were directed outward from





GPS: Applications in Crustal Deformation Monitoring, Figure 8  
**Slip distribution of the 1999  $M_w$  7.6 Chi-Chi Taiwan earthquake and optimized fault geometry inferred from GPS data. Colors indicate magnitude of slip and vectors show the direction that the hanging wall moved relative to the foot wall. The blue curve shows the trace of the earthquake rupture at the Earth's surface. Reprinted from [65] with permission from Elsevier**

the rift zone except at the ends of the fissure where they pointed inwards, parallel to the rift (Fig. 9). The magnitude of the displacements died off quickly with distance from the fissure. This displacement pattern is characteristic of a shallow dike intrusion within the rift zone. Prior to the eruption, sites at Kilauea summit showed subsidence and a radially inward pattern due to horizontal shortening across the summit, and this accelerated drastically during the eruption. This pattern suggests deflation, and ultimately emptying, of a magmatic source beneath the summit.

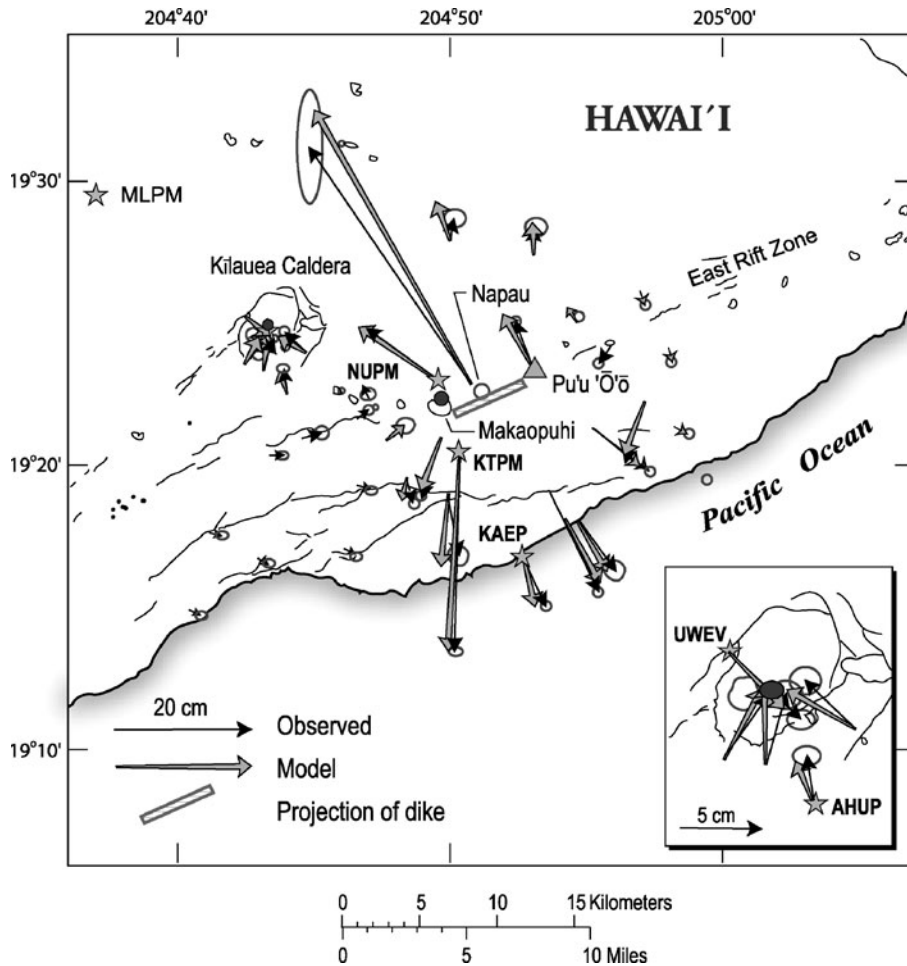
Using nonlinear optimization, Owen et al. [112] found that a source geometry consisting of a steeply dipping dike aligned with the rift and the fissures, combined with deflation both at the summit and at Makaopuhi crater, best fit the GPS observations (Fig. 9). The volume change at the two deflation sources was an order of magnitude less than the volume of the inferred dike intrusion. However, this discrepancy can be remedied if magma previously stored in a lava lake at Pu'u O'o which was seen to drain during this event, along with magma in a conduit thought to connect the summit magma chamber to Pu'u O'o, also contributed to the intrusion. That the period of time leading up to the eruption had been characterized by steady deflation of the summit indicates that dike formation was not in response to magma overpressurization at the summit, but rather some other process. Owen et al. [112] suggest that ongoing southeastward movement of the south flank of the volcano (e. g., [111]) created tensile stresses that encouraged the dike intrusion and fissure eruption at the rift. Following the eruption, inflation resumed at the summit

and the lava lake refilled, further evidence of the connection between the two magma reservoirs.

**Combined Use of Multiple Data Types** Wherever possible, multiple data types are used together to infer crustal deformation source characteristics. For example, GPS data are frequently used in combination with other geodetic measurements such as Interferometric Synthetic Aperture Radar (InSAR) [18] and leveling [38] data. GPS measurements are also often used in combination with seismic records to estimate fault slip.

(a) *GPS and InSAR* GPS data are commonly used in combination with InSAR data because of the complementary nature of these two data types. GPS observations provide three-component displacements, good horizontal precision, and (in the case of CGPS) good temporal coverage. InSAR, on the other hand, has exceptional spatial coverage, is more sensitive to vertical deformation than GPS, and does not require the deployment of instruments on the ground (thus enabling data collection from otherwise hazardous areas such as volcanoes). InSAR, jointly with GPS where possible, has been widely used to study volcanic deformation (see for example [39,40,119]). Likewise, GPS and InSAR observations, in some cases in combination with seismic data, have been used to infer the slip distribution and rupture history of numerous earthquakes including the 1992 Landers earthquake [54], the 1995 Kobe earthquake [113], the 1999 Hector Mine earthquake (e. g., [70,132,148]), and the 1999 Izmit earthquake [16,27]. Studies have explored the relative constraints on source parameters provided by each data-type, as well as approaches for optimally weighting different data types in inversions, especially when one method, like InSAR, produces many more data points than another, such as GPS (e. g., [71,126,148]). Wright et al. [168] used a combination of InSAR and GPS data to infer the slip distribution of the 2002 Denali earthquake. Because of the remote location of this event, GPS sites were clustered along roads [57,58], thus resulting in poor spatial coverage along some portions of the fault. In this situation InSAR observations helped reduce uncertainty in the slip estimates compared to estimates obtained from GPS data alone.

(b) *GPS and Seismic Data* It has long been recognized that GPS data are a useful complement to seismic records for estimating fault slip [162]. While seismic data are sensitive to the rupture process of an earthquake (the amount and temporal progression of slip), when using these data to characterize the rupture process, trade-offs exist between the time-history of slip and its spatial distribution.



GPS: Applications in Crustal Deformation Monitoring, Figure 9

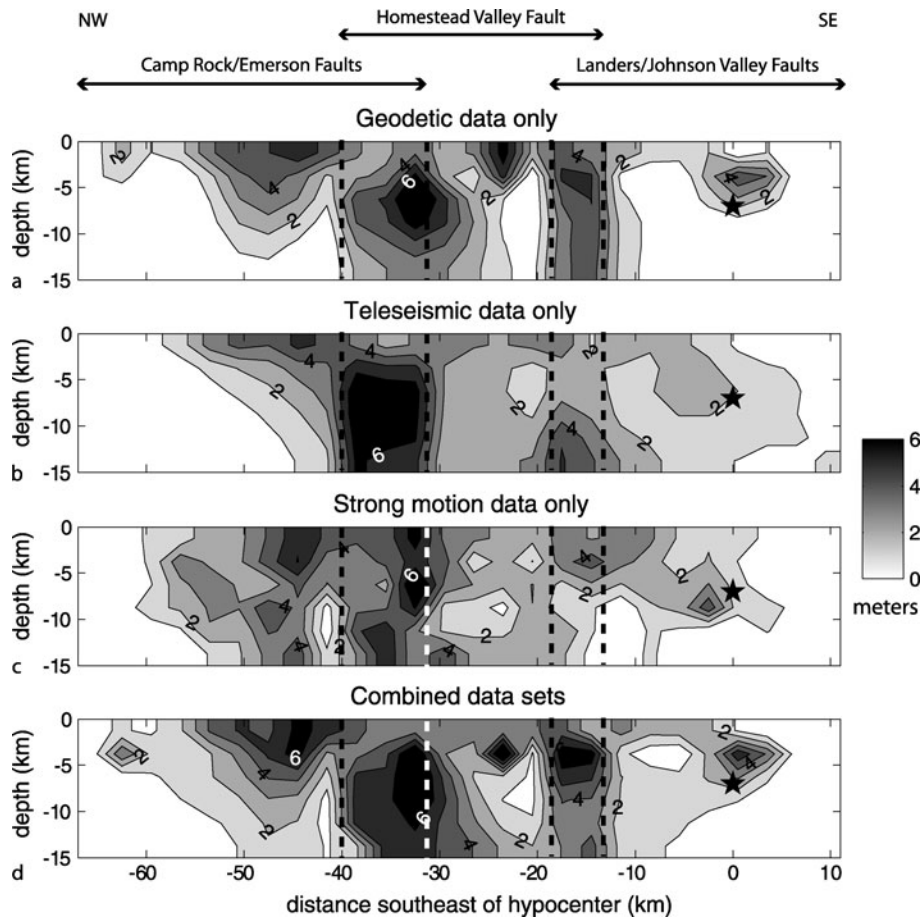
Displacements measured using GPS during the 1997 fissure eruption at Kilauea volcano. On the basis of the observations (black vectors with 95% confidence ellipses) Owen et al. [112] inferred the source of the deformation to be a combination of deflation at the summit of Kilauea (inward pointing arrows; see inset) and the intrusion of a dike along the east rift zone culminating in the eruption. The displacements predicted by this source geometry are shown by the gray arrows. CGPS sites are indicated by stars with four-character station codes. The deflation sources are shown as gray circles and the dike location by the hatched rectangle. Thin black lines are faults, fractures, and fissures. Adapted with permission from [112] (copyright 2000, American Geophysical Union)

GPS offsets due to an earthquake reflect only the final (or “static”) slip distribution, and thus the slip history inferred from the seismic data can be constrained to produce a static slip distribution that fits the geodetic displacements. The combined use of these two data types tends to have the added advantage of improved instrumental coverage over the study area.

Wald and Heaton [162] conducted a comparison of slip distributions for the 1992  $M_w$  7.2 Landers earthquake inferred from strong motion, teleseismic, and geodetic data (GPS displacements and displacements calculated from trilateration measurements) individually and in

a combined inversion. For all inversions they used a consistent fault geometry parametrization comprised of three fault segments based on the aftershock locations and the extensive ground surface rupture that was mapped following the event. They corrected the geodetic data for the effects of the  $M_w$  6.2 Big Bear event and assumed that post-Landers measurements were made soon enough to avoid contamination by postseismic signals.

On inspection of the final slip distributions (Fig. 10) obtained from inversions of each of the three datasets independently, the authors identified several features that were common to all three and thus appeared to be robust



GPS: Applications in Crustal Deformation Monitoring, Figure 10

Slip distributions estimated by Wald and Heaton [162] for the 1992  $M_w$  7.2 Landers earthquake using different data sets. Contour interval is 1 meter. Star marks hypocenter of earthquake. Dashed lines indicate along-strike boundaries of faults named at top. a Geodetic data only. b Teleseismic data only. c Strong motion data only. d Combined inversion of all three data sets. Adapted from [162]

regardless of dataset. For example, slip at the hypocenter was moderate and limited to a small depth range. Peak slip at depth occurred along the central portion of the fault, while slip became shallower at the ends (to the northwest and southeast). The greatest near-surface slip was on the Camp Rock / Emerson faults at the northwest end of the rupture. The slip distribution obtained through combined inversion of the three datasets was most similar to that from the geodetic data alone because the timing of slip is an additional degree of freedom in the inversion of teleseismic and strong motion waveform data that is not available in the inversion of static offsets. In addition to the final slip distribution, these authors imaged the temporal progression of slip on the fault surface. The rupture appears to slow as it nears the surface, as well as when it approaches the two step-over regions between fault seg-

ments. Furthermore, the authors infer that although the rupture generally propagated unilaterally northwest, each time the rupture jumped northwest to a subsequent segment it also propagated backwards down the abandoned portion of this segment southeast of the fault intersection. Thus, the combined use of geodetic and seismic data provided a more complete and robust understanding of the rupture dynamics of this earthquake, which could then be used to investigate, for example, the spatial patterns in the strength of ground shaking due to this event.

Another example of the combined use of geodetic and seismic data comes from the Dec. 26, 2004 Sumatra–Andaman earthquake, a subduction zone megathrust event which ruptured a  $\sim 1200$  km length of the plate boundary between the Indo-Australian and Eurasian

plates [11,81]. This earthquake, which produced peak-to-peak surface wave motions greater than 1 cm worldwide [115] and measurable static offsets at GPS sites at least 4500 km away [6] and probably farther [73], resulted in more than 283,000 deaths, largely due to its triggering of a major tsunami. This is the largest earthquake to have been recorded since the establishment of digital seismic networks and GPS, and both seismic and geodetic data have been critical in describing the rupture process of this earthquake.

The moment magnitude of an earthquake,  $M_w$ , is a measure of its size and may be derived from the seismic moment ( $M_o$ ), which is defined as

$$M_o = \mu sA \quad (7)$$

where  $\mu$  is the shear modulus of the faulted rock (in units of Pascals),  $s$  is the amount of fault slip during the earthquake (meters),  $A$  is the area of the surface that slipped in the earthquake (meters<sup>2</sup>), and  $M_o$  has units of Newtons  $\times$  meters. “Tsunami earthquakes” [120] are a subclass of subduction zone earthquakes defined as producing tsunamis larger than would have been predicted from their moment magnitude. Tsunami earthquakes have been observed to have slow rupture velocity with relatively little seismic energy release at high frequencies. Although the moment magnitude should represent the net amount of static slip that occurred in the earthquake, it is typically estimated from seismic data at periods of 100 to 300 seconds [81]. If significant seismic energy is released at longer periods, this method will underestimate moment magnitude. The Harvard CMT (Centroid Moment Tensor) solution for the Sumatra–Andaman event, computed using surface wave data with periods of 300 to 500 seconds, had  $M_o = 4.0 \times 10^{22}$  Nm which corresponds to  $M_w$  9.0 [81]. However, researchers quickly began to see evidence that this earthquake was in fact larger.

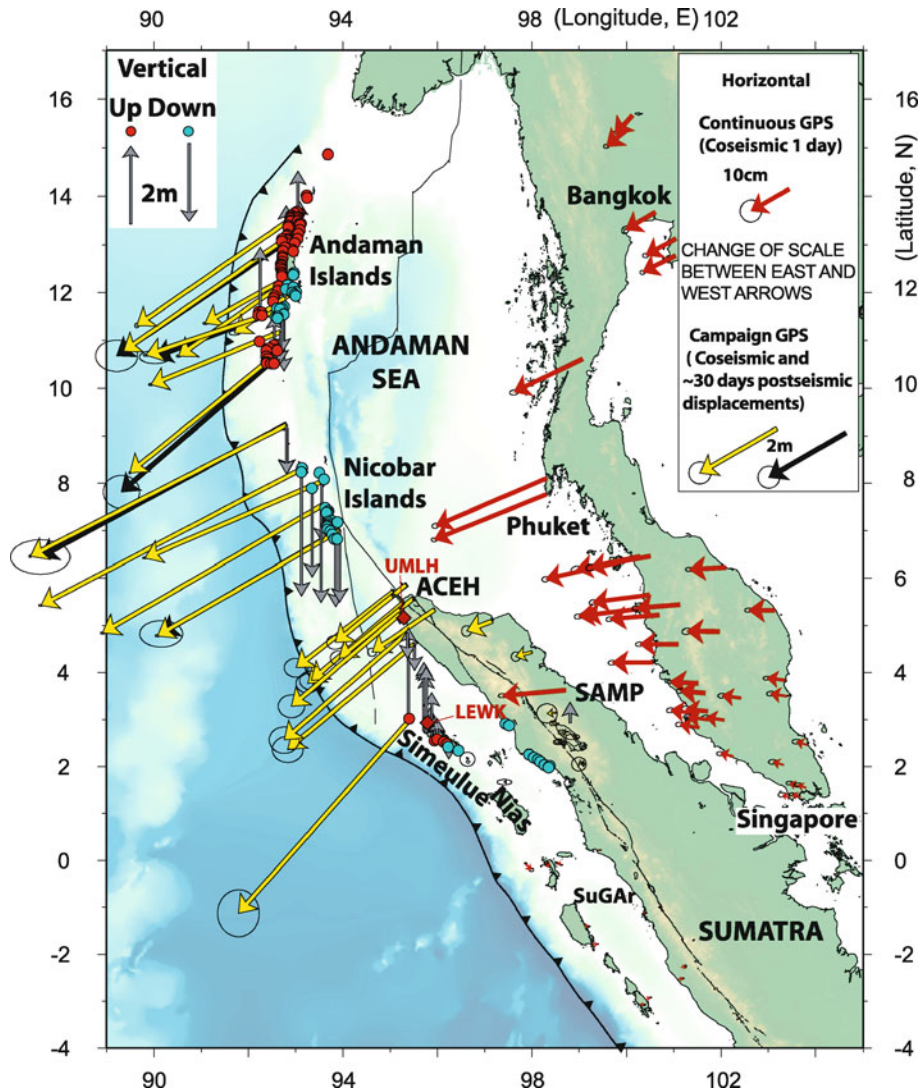
Ammon et al. [1] analyzed seismic data from a broad frequency range (including data with periods up to 54 minutes) and concluded that these observations could be fit by a model in which the majority of the rupture occurred over a 10 minute time span and slip was concentrated south of 8° N (on the southernmost ~800 km of the fault surface, Fig. 11). However, they note that the large GPS displacements reported in the Nicobar and Andaman islands would require 2- to 3-times more slip north of 8° N. Studies of the Earth’s seismic free oscillations [116,152] used these very low-frequency data (at periods up to one hour) to estimate the moment of the earthquake and found that moment increased with the period of the oscillations. These data suggest that the rup-

ture required ~10 minutes to travel from south to north, had more slip in the Nicobar and Andaman islands region than originally thought, and led to  $M_w$  estimates of 9.13 to 9.3. Park et al. [116] also point out that even slower slip (e.g. over a time span of ~1 hour) could have occurred and would be difficult to detect in the free oscillation data.

Banerjee et al. [7] compiled GPS data from several studies and used these observations to infer the static slip distribution on the rupture surface. The total moment release associated with their slip estimate is  $7.62 \times 10^{22}$  Nm (corresponding to an  $M_w$  9.22), smaller than that estimated by Stein and Okal [152], but greater than the  $6.11 \times 10^{22}$  Nm ( $M_w$  9.13) inferred from GPS data by Kreemer et al. [73]. Banerjee et al. [7] conclude that from ~2 to ~16 meters of slip on portions of the fault north of 8° N is needed to fit the GPS data from the Andaman and Nicobar islands. However, they argue that this slip did not occur slowly over a time span of an hour or more because the continuous GPS site at Phuket, Thailand showed little movement after ~10 minutes following the earthquake [160], and other sites in Thailand which should be particularly sensitive to slip on the Andaman segment given their location also do not show movement more than 10 to 20 minutes following the earthquake [53]. Although some slip estimates based on geodetic data alone have higher moment than those from seismic data alone, suggesting the occurrence of aseismic slip, Chlieh et al. [24] and Rhie et al. [129] conducted joint inversions of GPS and seismic data and were able to fit both datasets satisfactorily with a single rupture model.

Thus, although seismic data provide information on the details of fault rupture that cannot be obtained from GPS, in the case of the Sumatra–Andaman event the GPS observations provided needed constraints on the extent and duration of fault rupture, both of which had important implications for tsunami generation.

(c) *GPS and Gravity Data* GPS data have become widely used to study volcanic deformation processes, including the long-term uplift observed at calderas such as Yellowstone (Wyoming), Campi Flegrei (Italy), and Long Valley (California). Geodetic observations can constrain the source geometry and volume change [8]. They cannot, however, discriminate if the deformation is due to an influx of hydrothermal fluids or the intrusion of magma. Battaglia et al. [8,9] address this problem through the combined use of geodetic and gravity data recorded during a period of uplift in Long Valley caldera. Modeling the observed uplift using a point source will produce biased results if the true source does not possess spherical sym-



GPS: Applications in Crustal Deformation Monitoring, Figure 11

Summary of geodetic data recorded in the vicinity of the 2004 Sumatra–Andaman earthquake. The campaign GPS (yellow and black) vectors are compiled from [49,153] and contain approximately one month of postseismic deformation. The continuous GPS data come from Vigny et al. [160]. Note that the near-field vectors (those in the western part of the mapped region) and the far-field vectors use a different scale. Dots represent measurements of vertical deformation from satellite imagery [91]. The gray arrows indicate uplift and subsidence from GPS data, measurements of the vertical movement of coral heads, and mapping of shoreline changes [12,49,153]. Figure adapted from [24]

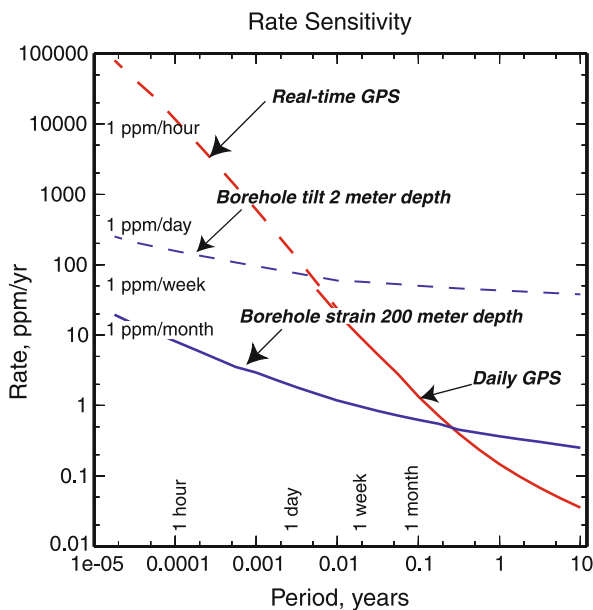
metry. Furthermore, the uplift signal of a range of source geometries can be similar, but the horizontal deformation signal can help in distinguishing among different models (e. g., [32]). Battaglia et al. [8] use a combination of vertical and horizontal geodetic measurements (GPS, leveling, and line-length data) to find the best-fitting source geometry, in their case a vertical prolate ellipsoid. Then, with that source geometry uniquely determined, they perform

a joint inversion of the uplift and gravity data to infer the volume and mass of the source, from which they obtain a density range of 1180 to 2330 kg/m<sup>3</sup> [9]. Since this density range is too high for hydrothermal fluids to be the sole source of uplift at Long Valley, these authors conclude that a silicic magma body or combination of magma and hydrothermal fluids is required to produce the observed deformation.

### Deformation over Time Scales of Decades to Seconds

We have seen that GPS data are useful for constraining models of deformation that happens rapidly, such as slip in an earthquake. However, perhaps the greatest strength of GPS is its ability to record deformation that occurs over a wide range of time periods. GPS can measure fault slip that happens too slowly to generate seismic waves (termed aseismic slip), volcanic deformation that occurs over several days or several years, and long-term interseismic strain accumulation. GPS can provide much more temporally dense measurements than InSAR and, although less sensitive, is stable to longer time periods than strainmeter data as shown in Fig. 12 [68].

**Interseismic Deformation** Interseismic deformation refers to the gradual straining of the Earth's crust that



GPS: Applications in Crustal Deformation Monitoring, Figure 12 Comparison of rate sensitivity for GPS, borehole strainmeters at 200 meters depth, and borehole tiltmeters (which measure the gradient in vertical deformation) at 2 meters depth. The x-axis indicates the time period that may be spanned by the different data types. For example, daily GPS measurements may span one day or longer. The y-axis indicates the strain rate that can be resolved as a function of the period. Strain is the change in length (area or volume) per a unit length (area or volume), and thus is unitless and can be expressed as parts per million (ppm). As can be seen from the plot, the borehole tiltmeters and strainmeters are more sensitive than GPS at shorter periods, but at periods longer than a few days and a few months, respectively, GPS measurements provide better resolution of strain rates. Figure courtesy of John Langbein

occurs during the time between moderate to large earthquakes. This strain can be caused by the build up of stress that will eventually be released during earthquakes, as well as reflect the broader-scale patterns of deformation in response to tectonic plate motion.

*(a) Block Versus Continuum Models for Deformation* As one moves from the global scale of tectonic plates to the continental scale, a major question is whether continents deform through the movement of many small rigid blocks (like plates on a smaller scale) or through more continuous deformation [157]. GPS data have been used in numerous studies to try to elucidate this hotly debated issue, and one region that has been a focus of study is the Tibetan plateau which accommodates strain due to the collision of India with Eurasia. GPS velocities for a profile of stations spanning central Tibet roughly parallel to the direction of maximum convergence between India and Eurasia show a linear gradient in the component of velocity parallel to the convergence direction. This observation has been cited as evidence for continuously distributed deformation across the region (e. g. [171]) or distributed strain combined with the movement of a small number of crustal blocks [23]. In contrast, Thatcher [158] showed that the velocities predicted by a model in which the Tibetan plateau was divided into a set of rigid, rotating crustal blocks defined by faults and other geologic features could also fit the data since the difference between the observed GPS velocities and those predicted by the block model were relatively small and did not exhibit widespread systematic spatial patterns. Although the rigid block model does not predict a linear velocity gradient across central Tibet, Thatcher [158] notes that both the block and distributed models fit the velocity profile within errors. The spatial sampling provided by GPS remains sparse across large portions of the Tibetan plateau, and it is likely that additional data will be necessary to resolve the outstanding question as to whether continental deformation in this region is primarily block-like or continuous.

*(b) Estimating Interseismic Fault Slip Rates* Following the 1906 San Francisco earthquake, H. F. Reid [127] used his observations of deformation in that event to formulate a description of the earthquake cycle which he termed “elastic rebound.” Reid recognized that strain builds up in the Earth's crust around faults during the interseismic period, and that strain is eventually released in earthquakes. Reid's hypothesis predated the theory of plate tectonics, but we now understand that the source of the ongoing stress affecting the faults is the motion of the Earth's plates. The strain build-up and release causes measurable

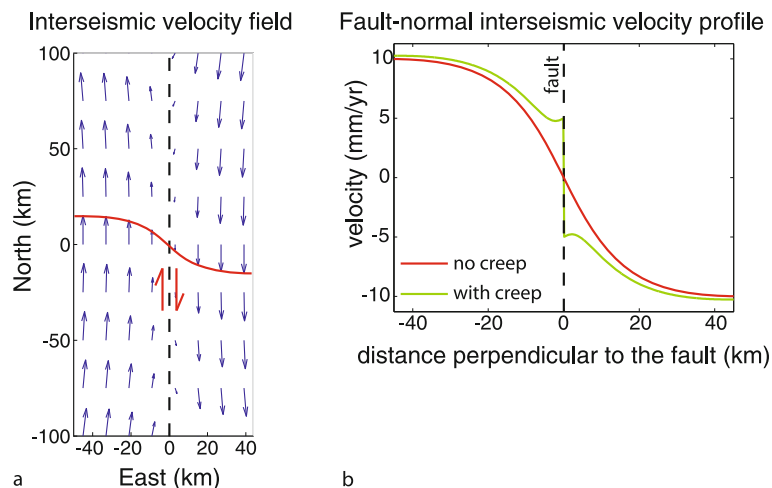
deformation of the Earth's surface. Areas where the crustal strain measured using GPS is large are likely to have earthquakes to relieve that strain. These events may take place on faults that are not visible on the earth's surface, and in this case the geodetic data can provide an important clue to the existence of seismic hazard. GPS enables measurement of deformation during all phases of the earthquake cycle.

One of the most important pieces of information needed to characterize a region's seismic hazard is an estimate of the slip rates on the major active faults that could affect that region. Most seismically active regions are at plate boundaries. The Earth's rigid plates are always moving at rates that are essentially constant over time periods comparable to earthquake recurrence intervals (e.g., hundreds to thousands of years). At plate boundaries the relative motion between two neighboring plates is accommodated on faults. The portion of the rate of relative motion that is accommodated across a given fault is that fault's slip rate. Slip rates are often estimated from geologic data, for example by dating samples collected from a location in which a measurable offset of a stream channel has occurred due to movement on a fault that crosses the stream [146]. However, slip rates can be inferred from geodetic data as well.

Below a certain depth in the Earth's crust (e.g., ~15 km for many strike-slip faults in the San Andreas system) the temperature and pressure are sufficiently high for

the crustal rock to behave plastically in response to stress, rather than experiencing brittle failure. Earthquakes occur above this depth, which is termed the "brittle-ductile transition," the "locking depth," or the "transition depth," but not below. In the time between moderate to large earthquakes most faults are largely locked above the transition depth, meaning no movement occurs across them. (It is true that very small earthquakes occur frequently on most faults, but these events affect a relatively small portion of the fault's surface area and release a tiny fraction of the energy, or moment, released in moderate and large events.) The material below the transition depth deforms gradually and continuously in response to plate motion, and a fault that is locked above the transition depth may exist as a zone of distributed shear below that depth. In the vicinity of a locked fault, the constant movement of the material below the transition depth strains the elastic crust above. This is manifest by a characteristic pattern of interseismic velocities for points on the Earth's surface near the fault. For example, the interseismic velocity profile perpendicular to a strike slip fault like the San Andreas will have a characteristic sigmoidal shape as shown in Fig. 13, the details of which reflect the locking depth and slip rate of the fault.

In regions dominated by a small number of major faults of known geometry (e.g., the San Andreas fault system in northern California, or subduction zones of Cascadia and Japan), interseismic slip rates can be inferred



GPS: Applications in Crustal Deformation Monitoring, Figure 13

**a** Expected velocity field due to interseismic slip below 15 kilometers on a vertical strike slip fault. The red curve highlights the difference in velocity near and far from the fault. **b** Expected velocity profiles for a locked fault and one that exhibits shallow creep. The locked fault is as in a. The creeping fault slips below 15 km depth and creeps in the uppermost 4 km but is locked from 4 to 15 km depth. Note that in the case of a creeping fault, there is an offset in the velocity profile close to the fault. The near-fault inflection in the green curve shows the combined effect of the strain due to slip below the locked zone and creep that reaches the Earth's surface

in much the same way as was done for earthquake slip (Eq. (6)) using relatively simple dislocation models in which the fault is prescribed to be locked above the transition depth and freely slipping below that (simulating the motion of the tectonic plates). A number of studies (e. g. [47,125]) have taken this approach to model the GPS velocities of sites in northern California as the superposition of interseismic velocity signals due to slip on the multiple sub-parallel faults that make up the San Andreas system. When the slip rate on a fault is estimated from interseismic GPS velocities, the resulting value is generally called the interseismic slip rate to emphasize that it has been estimated from data collected over a time period entirely within the interval between two earthquakes on the fault in question. When slip rates are estimated from geologic data they are often called “long-term average” rates to emphasize that they represent the rate over many earthquake cycles. In the absence of post-seismic effects or other transient deformation (described in more detail in a later section) these two slip rate estimates for a given fault should be the same. In subduction zones, interseismic velocities are often used to estimate the degree of plate “coupling,” which reflects the size of the locked zone that may rupture in a large earthquake [17,44,93,108,154,164].

As discussed earlier in the context of the Tibetan plateau, one interpretation of continental deformation patterns is that they arise from the rotation of fault-bounded blocks, and the GPS data can be used to estimate the Euler poles of rotation for each block. Data from sites near known faults may be discarded from the analysis because these observations will likely reflect the interseismic elastic strain accumulation due to the faults rather than the long-term rigid behavior of the blocks. Slip rates on the faults that bound the blocks can be calculated from the relative rates of block rotation (e. g. [158]). An alternative block modeling approach retains all the data, and the estimated block rotation rates must result in slip rates on block-bounding faults that are compatible with the patterns of strain accumulation recorded in the GPS velocity field. As with the dislocation models, a transition depth is assumed for each fault. The portion of each fault above the transition depth is treated as locked, and slip on the portion below the transition depth drives the observed strain accumulation.

Block modeling of this type lends itself to the many seismically active regions, such as southern California, that are characterized by numerous faults with complex geometries. Unlike models which represent individual faults by separate dislocations, block models are required to be self-consistent in that the rotation rates must be compatible for all blocks, slip rates have to be consistent at fault in-

tersections, and the total slip rate across the region is made to match the relative plate rate. A drawback of this approach, however, is that it is difficult to accommodate dipping faults and to constrain fault-perpendicular motion in a realistic way.

Both block models and dislocation models suffer from trade-offs between slip rate estimates on different faults and sensitivity to poor data coverage. Furthermore, it is difficult to resolve the contribution to the observed GPS velocity from strain accumulation on closely spaced faults (e. g. within two locking depths of each other), and thus slip rate estimates on neighboring faults tend to trade-off with each other and with the assumed locking depth. It will always be difficult to resolve slip rates on faults that are close together given that displacement measurements are confined to the Earth’s surface.

Several studies have applied the technique of block modeling with inclusion of elastic strain accumulation to the western United States [26,89,90]. For many faults the slip rates estimated by these studies agree to within errors with those estimated from geologic studies. However, there are some discrepancies. For example, the  $\sim 5$  mm/yr slip rate Meade and Hager [90] estimated for the San Bernardino segment of the San Andreas fault is considerably lower than the geologic estimate of  $\sim 25$  mm/yr. Discrepancies between geodetic and geologic slip rates have been found in several locations world wide, but it is not the case that one data type tends to produce consistently higher rate estimates than the other. The differences likely result from a combination of factors, including assumptions made in interpreting the data, the localized nature of geologic estimates, and the different time periods spanned by the two data types. Consistency between geodetic and geologic slip rates is an area of ongoing study.

Time varying deformation is often observed following large earthquakes. One source of this signal in many cases is the viscoelastic response of the material below the transition depth. When a large earthquake occurs it imparts stress to this material, which then deforms slowly, restressing the elastic crust above. The rate of the resulting strain would be expected to be high directly after the earthquake and die off with time. Likewise, the strain accumulation rate on the affected fault would vary throughout the earthquake cycle. These processes have been incorporated into another category of interseismic deformation models called “viscoelastic cycle” models [66,67,133,137,156], which can be used to estimate fault slip rates and earthquake recurrence times. These models account for the strain rate maxima on faults during the interseismic period as well as temporal variations in the months to years following a large earthquake.



(c) *Fault Creep* An interesting interseismic phenomenon, termed “fault creep,” is observed on several faults of the San Andreas system in central California and the San Francisco Bay Area, as well as faults in Taiwan, the Philippines, and Turkey. Faults exhibiting creep slip steadily or episodically at low average rates (e. g. 10 mm/yr). Some creep events are confined to the uppermost ~500 meters of faults, but in other cases creep occurs deeper at what are generally considered to be seismogenic depths (e. g. to a depth of ~15 km on faults of the San Andreas system). However, this fault slip is too slow to generate seismic waves. The ongoing slip of creeping faults is a nuisance, offsetting cultural features like curbs and buildings. However, ongoing creep constantly relieves stress which would otherwise be released in an earthquake and thus reduces the seismic hazard due to that fault. For example, a 2003 study of earthquake probabilities for the San Francisco Bay area found that accounting for creep on the Calaveras fault reduced the predicted rate of  $M \geq 6.7$  earthquakes on this fault by a factor of three [167]. Therefore, knowing the extent of creep is valuable in hazard assessments.

Fault creep will have a different interseismic signature than that for a fault that is locked in the seismogenic zone (Fig. 13b). Instead of the smooth transition seen across a locked fault, there will be a step because there is fault offset near the surface. Therefore, GPS data can be used to infer the depth-extent of fault creep. Several studies have used GPS data, in some cases in combination with InSAR, creepmeter, and microseismicity observations, to estimate the spatial distribution of fault creep (e. g. [64,85,103,136]).

**Temporally Varying Deformation** Continuous GPS measurements are particularly useful for observing transient, or time-varying, deformation. Sources of transient deformation include slow slip events, postseismic response, and volcanic processes.

(a) *Slow Slip Events* Slow slip events, sometimes also referred to as “slow earthquakes” or “silent earthquakes” (the latter emphasizing the lack of a seismic signature to the event), are a phenomenon in which fault slip occurs at too slow a rate to generate seismic waves. The duration of slow slip events that have been observed geodetically world-wide ranges from days (e. g. [22,106]) to years (e. g. [98,102]). Precursory transient slip with duration of minutes has been observed prior to earthquakes on mid-oceanic ridge transform faults using seismic data (e. g. [62]).

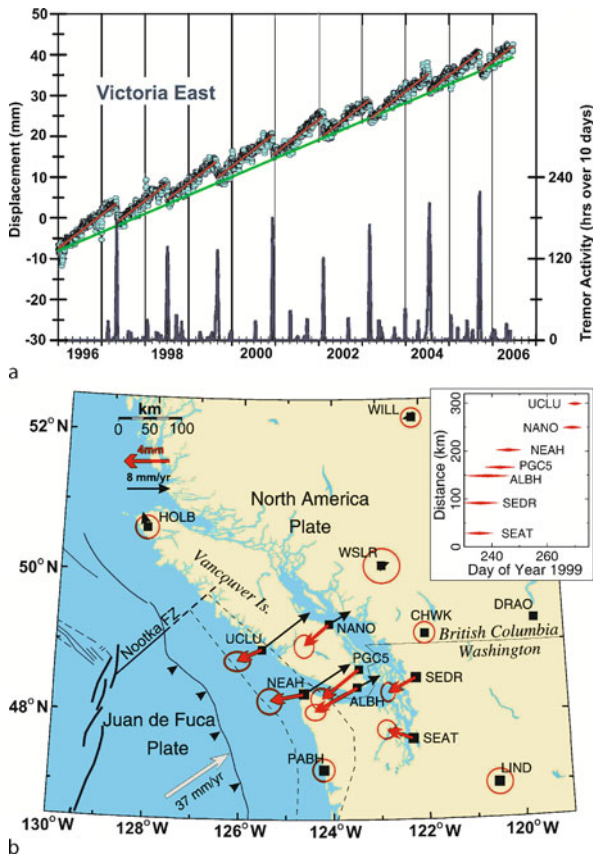
Shallow creep events and a multi-year transient increase in slip rate have been observed along the San

Andreas fault where continuous or frequent monitoring using creepmeters, strainmeters, and two-color electronic distance measuring instruments provided temporally dense measurements [51,52,75,82]. In 1996, a couple of years after the establishment of the Japanese CGPS network, evidence for slow slip in the Boso Peninsula region near Tokyo became apparent in the data [131]. As the spatial coverage of CGPS networks worldwide improved it became clear that slow slip events were much more frequent than previously thought, occurring over a variety of spatial and temporal timescales and tectonic settings.

The majority of large slow slip events observed to date have occurred in subduction zones, including those of Japan, the Pacific Northwest of the United States, Mexico, New Zealand, and Alaska. In 2001 a surprising pattern was recognized in time series for several CGPS sites in the Cascadia region in the Pacific Northwest of the United States (Fig. 14a). The subduction interface is thought to be locked near the Earth’s surface and freely slipping at greater depths. The ongoing deep slip results in interseismic motion of GPS sites toward the over-riding plate (Fig. 14b). In the case of the Cascadia sites, this means that the long-term average interseismic movement is eastward. However, it was observed that occasionally the sites briefly moved in the opposite direction, causing a step-like pattern in the position time series. An intriguing feature of the observed reversals in station velocities is that the pattern was found to repeat approximately every 14 months [94].

That several CGPS sites in the region showed a coherent reversal in the direction of motion at the same time (Fig. 14b) suggested that the source of this signal could be slip on the interface between the down-going slab and over-riding plate of the subduction zone. Modeling of the GPS data [35] indicated that the source region was a portion of the subduction interface between 30 km and 40 km depth that is a transitional zone linking the shallow fully locked and deeper fully slipping parts of the interface.

In 2002 Obara [105] reported a very low frequency seismic signal called tremor, typically observed at active volcanoes, emanating from ~30 km depth in the subduction zone of southwest Japan. A similar signal was soon discovered in Cascadia, and researchers quickly realized that the tremor occurred simultaneously with slow slip events in those locales (Fig. 14a). Furthermore, they identified the source regions of the tremor and found that it coincides spatially with the inferred source region of the slow slip [106,130]. In volcanic settings tremor is thought to be caused by the movement of fluids through conduits underground. Recent studies have found that slow slip events in subduction zones are accompanied by very low frequency earthquakes and that the tremor in these lo-



GPS: Applications in Crustal Deformation Monitoring, Figure 14  
**a** Time series for station ALBH located in Victoria British Columbia, Canada relative to stable North America. The blue dots are station positions. Although the overall interseismic movement of this site is eastward due to the ongoing strain caused by subduction (green line), every  $\sim 14$  months this site moves westward (steps in time series). The red line represents the average velocity in the time between slow slip events, which is a higher rate than the long-term interseismic movement (green line). The blue curve represents the time series of nonvolcanic seismic tremor. Periods of increased tremor activity coincide with the times at which the GPS site shows anomalous westward movement [130]. **b** Portion of the Cascadia subduction zone. Black vectors are interseismic velocities of continuous GPS sites relative to stable North America, and red vectors are anomalous displacements during the 1999 slow slip event. The inset shows relative timing of transient displacements among different sites. From [35]. Copyright, Her Majesty the Queen in right of Canada (2001)

cales is actually made up of many low amplitude low frequency events [142,143]. One interpretation is that shear slip on the subduction interface, rather than fluid flow, causes nonvolcanic tremor, implying that tremor and slow slip both arise from the same underlying process of shear slip. Although the tremor may not be directly caused by

fluid flow, the low frequency earthquakes and slow slip appear to coincide spatially with areas of high fluid pressure in the pore spaces of subduction zone rocks. The high pore pressure, perhaps resulting from metamorphic reactions that release fluid, may encourage shear slip [143].

Like Cascadia, other subduction zones, for example the Guerrero region of Mexico [83] and the Shikoku [106] and Tokai [55] regions of Japan, have also experienced, to varying degrees, quasi-periodic slow slip events. Slow slip, in some cases multiple events, has been observed in other subduction zones such as Alaska [107], New Zealand [34], and the Tokai [98] and Bungo channel [114] regions of Japan, but it remains to be seen if the transient slip is periodic. Even in Cascadia, which shows clear periodicity, there is variation in the periodicity along the strike of the subduction zone. For example, slow slip events have been observed in both northern and southern Vancouver Island with a  $\sim 14$  month periodicity, but the events in these two locations are 6 months out of phase with each other. Slow slip events in northern California, also part of the Cascadia subduction zone, have been found to have an  $\sim 11$  month recurrence interval [155]. In southwest Japan, the Shikoku region experiences short duration (on the order of a week), small amplitude slow slip events (in fact only detectable in the tiltmeter data) at six-month intervals coincident with tremor [106]. In contrast, the subduction zone beneath the Bungo channel region, which abuts the Shikoku region directly to the southwest, experiences infrequent large slow slip events lasting  $\sim 1.5$  years [114]. Similarly, in the Tokai region, short-term and long-term slow slip events seem to occur on nearly overlapping parts of the subduction zone [55,98].

Similar to fault creep, the cumulative effect of several slow slip events may be to relieve stress on the transition zone along the whole length of the subduction interface without a large earthquake. However, slow earthquakes in the subduction zones of Cascadia, Japan, and elsewhere may impart stress to the locked subduction interface up dip and thus increase the likelihood of a large earthquake [35,123]. Understanding both the mechanism of these events and what it means for seismic hazard continues to be a focus of intense study.

Transient slip has also been observed in non-subduction zone settings. GPS data have shown that the south flank of Kilauea volcano in Hawaii moves seaward at a rate of several cm/yr [111], perhaps a manifestation of the instability of the volcanic edifice. However, in 2001 it was observed that this motion sped up for a few days. Cervelli et al. [22] modeled the GPS observables during the period of increased station velocity and concluded that the source was a  $\sim M 6$  slow earthquake lasting 36 hours and gener-

ating an average of almost 9 cm of slip on a nearly horizontal thrust fault about 4.5 km underground. This observed transient signal occurred nine days after a storm caused 1 meter of rainfall on this part of the Big Island. Cervelli et al. [22] calculated that given reasonable values for the porosity and permeability of the rocks in the study area, the slow slip could have been triggered by an increase in pore fluid pressure as the rain water penetrated into faults on the volcanic edifice. Subsequently several more very similar transient events have been observed in this locale by GPS [139], not accompanied by anomalously high rainfall. Moreover, it has been recognized that these events are accompanied by increased seismicity in an adjacent area following the onset of the deformation signal. This lends credence to the interpretation that the displacement signal is due to fault slip and suggests that the transient slip triggers the seismicity. In order for slow slip to trigger the observed seismicity, the slip must occur deeper than originally thought. A model in which slip occurs at the depth of the interface between the volcano and the underlying ocean floor,  $\sim 8$  km, fits the GPS displacements as well as one in which slip occurs at more shallow depths ( $\sim 4.5$  km) [139].

*(b) Postseismic Deformation* Often following a moderate or large earthquake, continued aseismic deformation is observed. This may be due to several sources, including continued slip on the fault surface, diffusion of pore fluid, and the viscoelastic response of the lower crust (below the transition depth) and upper mantle. All of these processes are triggered by the stress changes imparted to the surrounding crust and mantle by the earthquake. More than one source can be active simultaneously, and the effects change and decay with time. The spatial and temporal evolution of postseismic signals can provide important insights into the frictional, hydrological, and rheological characteristics of the crust and upper mantle, and enable a better understanding of the way stress is redistributed in the crust, which is an important consideration in assessing seismic hazard.

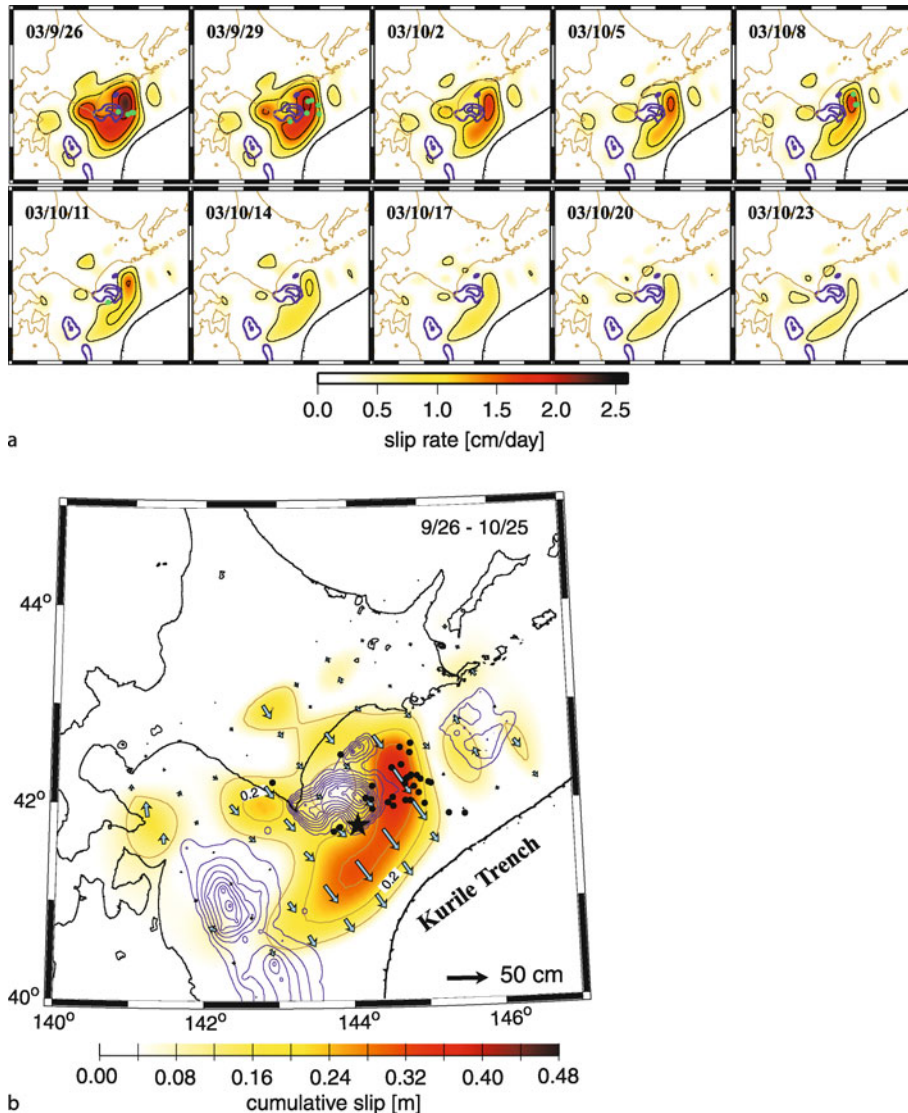
Afterslip is continued slip on the fault plane after the rapid slip which generates seismic waves has ceased. The effects can start immediately after the event (e. g. [76]) and can last for months or years, typically showing a logarithmic decay with time (e. g. [134]). Afterslip is thought to arise from the stress changes imparted by the earthquake to parts of the fault which have frictional properties that allow slow slip. Typically the total afterslip following an event is a fraction of the amount of slip that occurred coseismically. However, some events such as the Sanriku-haruka-oki [169] and Tokachi-Oki [97] earthquakes in

Japan have had afterslip with moment release approaching or, in the case of the 2004  $M_w$  6 Parkfield earthquake [76], exceeding the coseismic moment. Although the factors controlling the amount of afterslip are not fully understood, the fact that the San Andreas fault near Parkfield, as well as many subduction zone faults, are known to exhibit fault creep may be a factor.

The spatial and temporal evolution of afterslip following the 2003  $M_w$  8.0 Tokachi-Oki earthquake in Japan was estimated from GPS data using a Kalman-filtering technique [97] (Fig. 15a). The rate of afterslip started quite high and died off gradually. These authors noted that the highest afterslip rates tended to localize around the area which slipped in the earthquake. This is not unexpected as the coseismic slip area would have just experienced a reduction in shear stress which would discourage further slip here. Interestingly, the afterslip seems to avoid areas inferred to have slipped in other earthquakes that occurred in the decades prior to the Tokachi-Oki event as evidenced in the plot of cumulative afterslip in Fig. 15b. This interpretation provides support for the idea that rapid coseismic and gradual aseismic slip may occur on different parts of the fault because of variations in frictional properties. Similar results were found from an analysis of afterslip following the 2005 Nias–Simeulue thrust event [59].

As described earlier in regards to viscoelastic cycle models of interseismic deformation, the stress changes imparted by a moderate or large earthquake to the material below the elastic crust causes time-varying deformation. The geodetically recorded postseismic deformation following several large events has been interpreted to reflect viscoelastic processes (e. g. [156]). The rate at which viscoelastic postseismic deformation decays depends in part on the viscosity of the lower crust and upper mantle. The temporal decay of displacements measured with GPS have been used to infer viscosity values and thus the relative strength of these two layers (e. g. [5,31,45,46,121,122]). Estimates for the viscosity of the lower crust range from  $10^{19}$  to  $10^{21}$  Pa s, and for the upper mantle range from  $10^{17}$  to  $10^{19}$  Pa s.

Stress changes in the crust due to fault slip in an earthquake compress the pore space of rocks in some areas and cause dilation of the pore space elsewhere, depending on the orientation of the fault and sense and distribution of slip. The resulting pore pressure gradients cause fluid to flow from areas of high pressure to those of low pressure. This fluid flow causes further time-dependent strain. Such effects have been observed following several earthquakes including a pair of moderate earthquakes that occurred in 2000 in Iceland [5,69] and the  $M_w$  7.3 Landers event in California [43,118]. If poroelastic effects are confined



GPS: Applications in Crustal Deformation Monitoring, Figure 15

**a** Temporal evolution of the rate of afterslip (indicated by *colored shading*) on the subduction interface following the 2003  $M_w$  8 Tokachi-Oki earthquake. Dates given in upper left of each frame. *Blue contours* are the areas that slipped in previous earthquakes in this region. The *blue contours* in the center of the mapped area are those of the Tokachi-Oki earthquake. The *green dots* are aftershocks. **b** Cumulative afterslip in the first 30 days following the earthquake estimated from GPS data. Slip magnitude is given by the *colored shading*; the estimated amount and direction that the upper plate (the area northwest of the Kurile trench) moved relative to the lower plate is shown by the *arrows*. Afterslip tends to surround the areas inferred to have slipped in the Tokachi-Oki earthquake (epicenter given by the *black star*) and other events, shown by the *blue contours*. The *black dots* are aftershocks. Adapted with permission from [97] (copyright 2006, American Geophysical Union)

to the upper few kilometers of the fault zone, the spatial extent of the resulting surface deformation will be localized near the fault. Also, vertical surface displacement is a large component of the poroelastic signal. Because GPS measurements of vertical displacement are noisier than the horizontal data, and because the distribution of GPS sta-

tions may be limited near the causative fault, much of the insight into poroelastic deformation following events like the Landers earthquake and the earthquakes in Iceland has come from InSAR data.

It is unlikely that postseismic deformation associated with a given earthquake can be explained by a single pro-

cess. Given the often limited spatial and temporal data coverage and, in the case of GPS data, the sometimes poor vertical displacement control, it can be difficult to differentiate among different potential sources, e. g., afterslip, viscoelastic, and poroelastic deformation. However, the surface displacements due to each process can exhibit diagnostic patterns in time and space, which, if observed, make it possible to discern distinct causative processes. For example, due to its deeper source viscoelastic deformation should affect a broader geographic region than poroelastic deformation or shallow afterslip. Because viscoelastic deformation involves the response of material with viscosities on the order of  $10^{18}$  to  $10^{19}$  Pa s, while poroelastic processes involve the flow of aqueous fluid through the ground, the signal due to the former is expected to last considerably longer (e. g., several years) than that of the latter (e. g. several months). Several studies have analyzed postseismic deformation, in some cases observed using multiple data types, and attributed the observed deformation to a combination of two or more effects [5,43,46].

*(c) Volcano Deformation* Volcanic deformation is often characterized by transient signals. Magma or hydrothermal fluids migrate beneath the volcanic edifice, causing inflation or deflation and sometimes culminating in an intrusion or eruption. The steep flanks of volcanoes are often unstable, leading to landslides and in some cases collapse of large sections of the edifice. GPS is very well-suited to monitoring these types of deformation signals, and many of the world's volcanoes have CGPS receivers installed for this purpose. Some of the earliest GPS observations of volcanic deformation come from a submarine volcano near the Izu peninsula in Japan. Here GPS data from two receivers recorded deformation several days before volcanic tremor or visible signs of eruption were apparent [48,145].

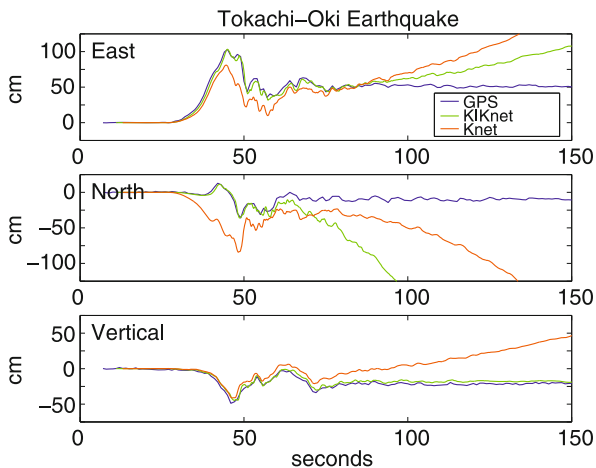
Loss of instruments in a volcanic eruption is a real and costly risk. InSAR, which does not require any equipment on the ground, is a widely used method for monitoring volcano deformation. However, InSAR does not provide three-component deformation measurements and does not have the temporal resolution that CGPS does. InSAR also suffers from decorrelation when vegetation, snowfall, or lava flows change the land surface during the time between two image acquisitions, however it still generally provides better spatial coverage than GPS. Dzurisin [40] presents a good overview of InSAR as applied to volcano deformation.

During a recent eruption of Augustine, a stratovolcano in Alaska, GPS data proved to be a valuable complement to other monitoring systems [20]. In the summer of 2005 CGPS stations on the volcano began to record an infla-

tion signal, following an increase in microseismicity below the volcano that had begun in May 2005 or perhaps earlier. Beginning in November 2005 the rate of inflation increased rapidly, but then died off somewhat in early January 2006. The inflation signal has been interpreted as evidence for a dike intrusion into the volcanic edifice which nearly reached the surface [20]. On January 11, 2006 a series of explosive eruptions began at Augustine, destroying two of the six CGPS receivers. By January 17th, though, the eruptions had died down, as had seismicity and gas emissions. Only the GPS data from two of the remaining stations showed continued inflation. The quiescence lasted 10 days before another explosive eruption and effusive lava flows took place. This is an example in which GPS data provided early corroboration in the summer of 2005 that increased seismicity was due to magma movement into the volcanic edifice. Moreover, during the ten days of quiescence when other indicators such as seismicity and gas measurements showed little activity, the GPS data showed that further eruptive activity was likely [20].

**High-Rate GPS** Continuous GPS networks typically record data at 15- or 30-second sampling rates, which is more than adequate for obtaining daily positions. One such CGPS network, the Southern California Integrated GPS Network (SCIGN), recorded data at 30-second intervals for the 1999  $M_w$  7.1 Hector Mine earthquake. Nikolaidis et al. [104] used these data to obtain positions at every observation epoch for the time spanning the earthquake and demonstrated that these observations, although aliased, agreed with those from nearby strong ground motion instruments. However, GPS receivers are capable of recording data at much higher rates, e. g. 1 Hz or greater. In recent years, for scientific as well as surveying and navigation purposes, an increasing number of CGPS sites have been set to record high-rate data and in some cases transmit them in real-time.

Unlike seismic data, GPS receivers provide a direct measure of displacement, and the instrument stays on scale even during the strong shaking of an earthquake. This makes it possible to obtain high-rate displacement time histories without the error introduced by integrating velocity or acceleration records from seismic instruments and without the data loss that occurs when the shaking exceeds the dynamic range of a seismic instrument (Fig. 16). Although the results of Nikolaidis et al. [104] hinted at a potential application of high-rate data, this kind of observation was not yet available for an earthquake. Such an event occurred in 2002 with the  $M_w$  7.9 Denali earthquake in Alaska. Larson et al. [78], obtained displacement-time histories of seismic waves from 1 Hz GPS data recorded



GPS: Applications in Crustal Deformation Monitoring, Figure 16 Comparison of displacement time series from a GPS station and doubly integrated accelerometer data from two nearby stations of the KIKnet and Knet networks in Japan for the 2003  $M_w$  8 Tokachi-Oki earthquake. The GPS data and that from the KIKnet site (blue and green curves) agree well for the first  $\sim 50$  seconds following the arrival of the seismic signal (at  $\sim 30$  seconds). The deviation of the accelerometer data after several seconds is likely due to low frequency noise that is amplified by the double integration required to produce displacements. Additionally, the Knet site (red curve) may be affected by tilting during shaking or other problems with sensor orientation, leading to the greater deviation between the red curve and those of the other two instruments. Reprinted with permission from [96] (copyright 2006, American Geophysical Union)

during this event and showed they were in good agreement with those obtained by doubly integrating accelerometer data.

Wang et al. [165] demonstrated that GPS receivers operating at 1 Hz provide a faithful recording of signals with periods of 2 seconds or greater. However, GPS receivers will not replace seismometers for recording earthquakes since they cannot capture the higher-frequency components of the seismic waves. Rather, they provide a valuable complement to seismic data. For example, Emore et al. [42] show that high-rate GPS data can be used as a constraint when integrating accelerograms. Moreover, the combined use of seismic and high-rate GPS data extends the observable frequency and amplitude range of seismic waves considerably [78]. Miyazaki et al. [96] demonstrated that GPS seismograms measured by 1 Hz GPS could be used in a similar way to seismic data to infer the spatial and temporal progression of fault slip during the earthquake. High-rate GPS observations can be especially useful in this capacity when used in combination with seismic data (e.g. [63]).

High-rate GPS also vastly improves our ability to track early postseismic deformation. Seismic data do not record any processes, like afterslip, that do not generate seismic waves, so postseismic deformation measurements must be made using other instruments including GPS, strainmeters, and InSAR. Until CGPS networks became common, it was generally not possible to deploy instruments for post-earthquake geodetic measurements until field crews could reach the affected area, days or weeks after the event. As a result, any change in measured position from the last pre-earthquake survey until the first post-earthquake survey contained the full coseismic signal and some portion of postseismic displacement. Even the “coseismic” displacements obtained from differencing daily positions from CGPS data collected the day before and the day after the earthquake can be contaminated by postseismic signals that began immediately after the event. High-rate GPS data, on the other hand, provide a means to record the deformation signal continuously, starting in the seconds after an earthquake. An example of this was the 2004  $M_w$  6 Parkfield earthquake. Langbein et al. [76] showed that rapid postseismic displacement began immediately after the event, and these authors were able to use the high-rate time series to separate the coseismic and postseismic displacements. For other events, including the 1966  $M_w \sim 6$  Parkfield earthquake, the moment release estimated from seismic data has often been substantially lower than that estimated using geodetic data (e.g. [138]). However in the case of the 2004 Parkfield event the slip estimated from the coseismic portion of the GPS displacement signal had moment release in good agreement with that estimated from seismic data [76].

The availability and use of high-rate GPS data in real-time for geophysical applications is not yet widespread, but it is growing as more CGPS sites are installed, telemetry of high bandwidth data becomes more feasible, and approaches to processing the data are refined. These data have great potential for aiding real-time deformation monitoring on volcanoes, for early warning of major earthquakes, and for tsunami warning as well.

For example, Mattia et al. [87] highlight the value of high-rate real-time GPS data in a time-critical situation to quickly differentiate localized volcanic processes (in their case a new volcanic vent opening) from more far-reaching dangers (e.g., a feared flank failure that could have caused a local tsunami).

Real-time high-rate GPS data may soon also contribute to rapid earthquake warning and response. Hudnut et al. [61] envision a system consisting of a pair of GPS receivers straddling a fault such as the San Andreas which record data at high rate and transmit these mea-

surements in real time. In the case of a major earthquake which, within seconds, caused surface offset on the fault of more than a few centimeters, the “GPS slip-sensor” would indicate that the event was large significantly sooner than could be expected from seismic data alone. This rapid information could be used to trigger preventive measures for critical infrastructure such as transportation systems (e. g., slowing trains to help prevent derailment) and factories (e. g. halting processes involving hazardous chemicals).

Blewitt et al. [14], using data from the 2004 Sumatra earthquake, demonstrated that GPS data available in real time, even if recorded at the traditional 30-second sampling rate, can be used to reliably estimate displacements larger than about 10 mm for great earthquakes. Such information would significantly improve the robustness of tsunami warning systems, at least for oceanwide tsunamis, by giving a better indication of the true moment magnitude of the earthquake more quickly (e. g. within 15 minutes) than can be achieved with seismic data.

The patterns of strong shaking that occur during an earthquake provide a good indication as to where damage will be the most severe. The U. S. Geological Survey has implemented a system called ShakeMap which rapidly generates a map of shaking intensity based on instrumental recordings. This is a valuable tool for emergency responders and scientists in the aftermath of an earthquake, and is available to the public as well (<http://earthquake.usgs.gov/eqcenter/shakemap/>). However, in many areas seismic instrument coverage is sparse, hindering the generation of accurate and detailed maps. Dreger et al. [36] demonstrate that seismic data can be used to rapidly generate a model of fault slip during an earthquake and that such a model can significantly improve the quality of ShakeMap produced for that event. This is especially true when the direction of earthquake rupture away from the hypocenter shows a preferred orientation along strike, since shaking will be stronger at locations that coincide with this directivity. Current work is focused on incorporation of real-time high-rate GPS data into the fault slip analysis with the ultimate goal of further refining ShakeMaps.

### Future Directions

As evidenced by the applications described in this article, GPS has become an indispensable tool for monitoring crustal deformation hazards and investigating the underlying processes. It provides an affordable means of obtaining surface displacement measurements with millimeter-level precision at any time of day anywhere on earth in all weather conditions with no line of sight requirement. GPS

data complement other observations such as seismic and InSAR data by providing a direct measure of displacement in three dimensions that stays on scale. GPS can record deformation over temporal scales of seconds to decades, thus making it possible to track surface waves generated by an earthquake, postseismic deformation, slow slip events, volcanic unrest, fault creep, interseismic strain accumulation, and plate motions.

For the first decade or so, applications of GPS in crustal deformation tended toward periodic measurements in the style of earlier geodetic surveys. However, the size, power consumption, and cost of receivers have decreased steadily, while the data storage capacity has increased. This has fostered the growth of large CGPS networks, accompanied by more centralized and uniform processing and availability of results. This trend is likely to continue, and many future applications of GPS will focus on further exploiting the ability of this tool to provide temporally dense measurements for tracking deformation.

High-rate data show promise for hazard monitoring and response as well as for providing insight into the physical processes underlying time-varying deformation, especially as more sophisticated methods become available for processing these data in real-time and mitigating error sources such as multipath which are particularly problematic for high rate measurements.

In areas with few CGPS sites or where spatially dense coverage is needed to address specific scientific questions, an alternative to traditional campaign GPS measurements has recently been developed [13]. Termed “semi-permanent” or “semi-continuous” GPS, it involves rotating a pool of GPS receivers through several subsets of GPS sites such that each subset is observed for periods of multiple weeks several times a year. Each site to be observed is outfitted with a specially designed antenna mount set in a rock outcrop so that the antenna is attached in the same location and orientation every time the site is occupied, thus eliminating the set-up error inherent in many SGPS surveys. Blewitt et al. [13] have demonstrated that the scatter in the position time series as well as the uncertainties on velocities estimated using data from a semi-permanent network in Nevada are nearly as low as for CGPS networks over a comparable length of time (e. g., 1.5 years). Although this method requires sites that have rock outcrops and that are secure enough for equipment to be left unattended for extended periods of time, its use is likely to grow as it is arguably more cost-effective than either CGPS or SGPS for many applications.

Finally, a series of changes are being made to modernize the GPS signal. For instance, beginning with a generation of satellites called Block IIR-M, first launched in the

fall of 2005, a civilian code is now modulated on the L2 carrier in addition to the C/A code that has always been available on the L1 carrier. This makes it easier to obtain the L2 carrier signal and enables elimination of ionospheric delay when positioning with the code measurements alone. A new series of GPS satellites, the Block IIF, set to launch in 2007 will transmit a third carrier frequency called L5 which will aid in ambiguity resolution. The L5 observable, modulated with a civilian-accessible code, will be broadcast at a higher power than the L1 or L2, making it easier to acquire. Future satellites will transmit a second civilian code on the L1 frequency that will be more robust than the C/A code and will be interoperable with a civilian accessible code to be transmitted by the planned Galileo satellites.

## Acknowledgments

John Langbein, Margaret Boettcher, Maurizio Battaglia, Emily Desmarais, Tom Hanks, and Fred Pollitz provided helpful comments which significantly improved this manuscript.

## Bibliography

### Primary Literature

- Ammon C, Ji C, Thio HK, Robinson D, Ni S, Hjorleifsdottir V, Kanamori H, Lay T, Das S, Helmlinger D, Ichinose G, Polet J, Wald D (2005) Rupture process of the 2004 Sumatra-Andaman earthquake. *Science* 308:1133–1139
- Argus D, Gordon R (1991) No-net-rotation model of current plate velocities incorporating plate motion model NUVEL-1. *Geophys Res Lett* 18:2039–2042
- Argus D, Gordon R, Ma C, Eanes R, Heflin M, Owen S, Willis P (2006) GEODVEL: Plate motions from space geodesy. *Eos Trans AGU* 87(52) Fall Meet Suppl:Abstract G41A-06
- Argus D, Heflin M (1995) Plate motion and crustal deformation estimated with geodetic data from the Global Positioning System. *Geophys Res Lett* 22:1973–1976
- Árnadóttir T, Jónsson S, Pollitz F, Jiang W, Feigl K (2005) Post-seismic deformation following the June 2000 earthquake sequence in the south Iceland seismic zone. *J Geophys Res* 110:B12308; doi:10.1029/2005JB003701
- Banerjee P, Pollitz F, Bürgmann R (2005) The size and duration of the Sumatra-Andaman earthquake from far-field static offsets. *Science* 308:1769–1772
- Banerjee P, Pollitz F, Nagarajan B, Bürgmann R (2007) Co-seismic slip distributions of the 26 December 2004 Sumatra-Andaman and 28 March 2005 Nias earthquakes from GPS static offsets. *Bull Seismol Soc Amer* 97:586–5102
- Battaglia M, Segall P, Murray J, Cervelli P, Langbein J (2003) The mechanics of unrest at Long Valley caldera, California: 1. Modeling the geometry of the source using GPS, leveling and two-color EDM data. *J Volc Geotherm Res* 127:195–217
- Battaglia M, Segall P, Roberts C (2003) The mechanics of unrest at Long Valley caldera, California: 2. Constraining the nature of the source using geodetic and micro-gravity data. *J Volc Geotherm Res* 127:219–245
- Bennett R, Davis J, Wernicke B (1996) First results from the northern Basin and Range continuous GPS network. *Eos Trans AGU* 77(46) Fall Meet Suppl:150
- Bilek S, Satake K, Sieh K (2007) Introduction to the special issue on the 2004 Sumatra-Andaman earthquake and the Indian Ocean tsunami. *Bull Seis Soc Amer* 97:S1–S5
- Bilham R, Engdahl R, Feldl N, Satyabala S (2005) Partial and complete rupture of the Indo-Andaman plate boundary 1847–2004. *Seismol Res Lett* 76:299–311
- Blewitt G, Hammond WC, Kreemer C (2009) Geodetic observation of contemporary strain in the northern Walker Lane: 1, Semi-permanent GPS strategy. In: Oldow JS, Cashman PH (eds) Late Cenozoic Structure and Evolution of the Great Basin – Sierra Nevada Transition. *Geol Soc Amer* (in press) doi:10.1130/2009.2447(1)
- Blewitt G, Kreemer C, Hammond W, Plag HP, Stein S, Okal E (2006) Rapid determination of earthquake magnitude using GPS for tsunami warning systems. *Geophys Res Lett* 33:L11309; doi:10.1029/2006GL026145
- Bock Y, Nikolaidis R, de Jonge P, Bevis M (2000) Instantaneous geodetic positioning at medium distances with the Global Positioning System. *J Geophys Res* 105:28223–28253
- Bos A, Usai S, Spakman W (2004) A joint analysis of GPS motions and InSAR to infer the coseismic surface deformation of the Izmit, Turkey earthquake. *Geophys J Int* 158:849–863
- Bürgmann R, Kogan M, Steblov G, Hilley G, Levin V, Apel E (2005) Interseismic coupling and asperity distribution along the Kamchatka subduction zone. *J Geophys Res* 110:B07405; doi:10.1029/2005JB003648
- Bürgmann R, Rosen P, Fielding E (2000) Synthetic aperture radar interferometry to measure Earth's surface topography and its deformation. *Ann Rev Earth Planet Sci* 28:169–209
- Calais E, Han JY, DeMets C, Nocquet JM (2006) Deformation of the North American plate interior from a decade of continuous GPS measurements. *J Geophys Res* 111:B06402; doi:10.1029/2005JB004253
- Cervelli P, Fournier T, Freymueller J, Power J (2006) Ground deformation associated with the precursory unrest and early phases of the January 2006 eruption of Augustine Volcano, Alaska. *Geophys Res Lett* 33:L18304; doi:10.1029/2006GL027219
- Cervelli P, Murray M, Segall P, Aoki Y, Kato T (2001) Estimating source parameters from deformation data, with an application to the March 1997 earthquake swarm off the Izu Peninsula, Japan. *J Geophys Res* 106:11217–11237
- Cervelli P, Segall P, Johnson K, Lisowski M, Miklius A (2002) Sudden aseismic fault slip on the south flank of Kilauea volcano. *Nature* 415:1014–1018
- Chen Q, Freymueller J, Wang Q, Yang Z, Xu C, Liu J (2004) A deforming block model for the present-day tectonics of Tibet. *J Geophys Res* 109:B01403; doi:10.1029/2002JB002151
- Chlieh M, Avouac JP, Hjorleifsdottir V, Song TR, Ji C, Sieh K, Sladen A, Hebert H, Prawirodirdjo L, Bock Y, Galetzka J (2007) Coseismic slip and afterslip of the great  $M_w$  9.15 Sumatra-Andaman earthquake of 2004. *Bull Seis Soc Amer* 97:S152–S173
- Coe JA, Ellis WL, Godt JW, Savage WZ, Savage JE, Michael JA, Kibler JD, Powers PS, Lidke DJ, Debray S (2003) Seasonal movement of the Slumgullion landslide determined from



- Global Positioning System surveys and field instrumentation, July 1998 – March 2002. *Eng Geol* 68:67–101
26. d'Alessio M, Johanson I, Bürgmann R, Schmidt D, Murray M (2005) Slicing up the San Francisco Bay Area: Block kinematics and fault slip rates from GPS-derived surface velocities. *J Geophys Res* 110:B06403; doi:10.1029/2004JB003496
  27. Delouis B, Giardini D, Lundgren P, Salichon J (2002) Joint inversion of InSAR, GPS, teleseismic, and strong-motion data for the spatial and temporal distribution of earthquake slip; application to the 1999 Izmit mainshock. *Bull Seis Soc Amer* 92:278–299
  28. DeMets C, Gordon R, Argus D (2006) Moving beyond NUVEL-1A: The MORVEL estimates of geologically recent global plate motions. *Eos Trans AGU* 87(52) Fall Meet Suppl:Abstract G41A-05
  29. DeMets C, Gordon R, Argus D, Stein S (1990) Current plate motions. *Geophys J Int* 101:425–478
  30. DeMets C, Gordon R, Argus D, Stein S (1994) Effect of recent revisions to the geomagnetic reversal time scale on estimates of current plate motions. *Geophys Res Lett* 21:2191–2194
  31. Deng J, Gurnis M, Kanamori H, Hauksson E (1998) Viscoelastic flow in the lower crust after the 1992 Landers, California, earthquake. *Science* 282:1689–1692
  32. Dieterich J, Decker R (1975) Finite element modeling of surface deformation associated with volcanism. *J Geophys Res* 80:4094–4102
  33. Dong D, Fang P, Bock Y, Cheng MK, Miyazaki S (2002) Anatomy of apparent seasonal variations from GPS-derived site position time series. *J Geophys Res* 107:2075; doi:10.1029/2001JB000573
  34. Douglas A, Beavan J, Wallace L, Townend J (2005) Slow slip on the northern Hikurangi subduction interface, New Zealand. *Geophys Res Lett* 32:L16305; doi:10.1029/2005GL023607
  35. Dragert H, Wang K, James TS (2001) A silent slip event on the deeper Cascadia subduction interface. *Science* 292:1525–1528
  36. Dreger D, Gee L, Lombard P, Murray M, Romanowicz B (2005) Rapid finite-source analysis and near-fault strong ground motions: Application to the 2003  $M_w$  6.5 San Simeon and 2004  $M_w$  6.0 Parkfield earthquakes. *Seismol Res Lett* 76:40–48
  37. Du Y, Segall P, Gao H (1997) Quasi-static dislocations in three dimensional inhomogeneous media. *Geophys Res Lett* 24:2347–2350
  38. Dzurisin D (1992) Geodetic leveling as a tool for studying restless volcanoes. In: Ewert J, Swanson D (eds) *Monitoring volcanoes: Techniques and strategies used by the staff of the Cascades Volcano Observatory, 1980-1990*, USGS Bull. 1966, US Geological Survey, Reston, VA, pp 125–134
  39. Dzurisin D (2003) A comprehensive approach to monitoring volcano deformation as a window on the eruption cycle. *Rev Geophys* 41; doi:10.1029/2001RG000107
  40. Dzurisin D (2007) *Volcano deformation: Geodetic monitoring techniques*. Springer, New York
  41. Eberhart-Phillips D, Michael AJ (1993) Three-dimensional velocity structure, seismicity, and fault structure in the Parkfield region, central California. *J Geophys Res* 98:15737–15758
  42. Emore G, Haase J, Choi K, Larson K, Yamagiwa A (2007) Recovering seismic displacements through combined use of 1-Hz GPS and strong-motion accelerometers. *Bull Seismol Soc Amer* 97:357–378; doi:10.1785/0120060153
  43. Fialko Y (2004) Evidence of fluid-filled upper crust from observations of postseismic deformation due to the 1992  $M_w$  7.3 Landers earthquake. *J Geophys Res* 109:B08401; doi:10.1029/2004JB002985
  44. Fletcher H, Beavan J, Freymueller J, Gilbert L (2001) High interseismic coupling of the Alaska subduction zone SW of Kodiak island inferred from GPS data. *Geophys Res Lett* 28:443–446
  45. Freed A, Bürgmann R (2004) Evidence of power-law flow in the Mojave desert mantle. *Nature* 430:548–551
  46. Freed A, Bürgmann R, Calais E, Freymueller J, Hreinsdóttir S (2006) Implications of deformation following the 2002 Denali, Alaska, earthquake for postseismic relaxation processes and lithospheric rheology. *J Geophys Res* 111:B01401; doi:10.1029/2005JB003894
  47. Freymueller J, Murray M, Segall P, Castillo D (1999) Kinematics of the Pacific-North America plate boundary zone, northern California. *J Geophys Res* 104:7419–7441
  48. Fujinawa Y, Shimada S, Ohmi S, Sekiguchi S, Eguchi T, Okada Y (1991) Fixed point GPS observation of crustal movement associated with the 1989 seismic swarm and submarine volcanic activities of Ito, central Japan. *J Phys Earth* 39:141–153
  49. Gahalaut VK, Nagarajan B, Catherine JK, Kumar S (2006) Constraints on 2004 Sumatra–Andaman earthquake rupture from GPS measurements in Andaman-Nicobar Islands. *Earth Planet Sci Lett* 242:365–374
  50. Gili JA, Corominas J, Rius J (2000) Using Global Positioning System techniques in landslide monitoring. *Engineering Geology* 55:167–192
  51. Gladwin M, Gwyther R, Hart R, Breckenridge K (1994) Measurements of the strain field associated with episodic creep events on the San Andreas fault near San Juan Bautista, California. *J Geophys Res* 99:4559–4565
  52. Gwyther RL, Gladwin MT, Mee M, Hart RHG (1996) Anomalous shear strain at Parkfield during 1993–94. *Geophys Res Lett* 23:2425–2428
  53. Hashimoto M, Hashizume M, Takemoto S, Fukada Y, Fujimori K, Takiguchi H, Satomura M, Otsuka Y, Saito S (2006) Postseismic deformations following the Sumatra–Andaman and Nias earthquakes detected by continuous GPS observation in SE Asia. *Seism Res Lett* 77:289
  54. Hernandez B, Cotton F, Campillo M (1999) Contribution of radar interferometry to a two-step inversion of the kinematic process of the 1992 Landers earthquake. *J Geophys Res* 104:13083–13099
  55. Hirose H, Obara K (2006) Short-term slow slip and correlated tremor episodes in the Tokai region, central Japan. *Geophys Res Lett* 33:L17311; doi:10.1029/2006GL026579
  56. Hofmann-Wellenhof B, Lichtenegger H, Collins J (2001) *Global Positioning System theory and practice*, 5th edn. Springer, New York
  57. Hreinsdóttir S, Freymueller JT, Bürgmann R, Mitchell J (2006) Coseismic deformation of the 2002 Denali Fault earthquake: Insights from GPS measurements. *J Geophys Res* 111:B03308; doi:10.1029/2005JB003676
  58. Hreinsdóttir S, Freymueller JT, Fletcher HJ, Larsen CF, Bürgmann R (2003) Coseismic slip distribution of the 2002  $M_w$  7.9 Denali Fault earthquake, Alaska, determined from GPS measurements. *Geophys Res Lett* 30:1670; doi:10.1029/2003GL017447
  59. Hsu Y, Simons M, Avouac J-P, Galetzka J, Sieh K, Chlieh M, Natawidjaja D, Prawirodirdjo L, Bock Y (2006) Frictional after-

- slip following the 2005 Nias-Simeulue earthquake, Sumatra. *Science* 312:1921–1926
60. Hudnut K (1997) The Southern California Integrated GPS Network (SCIGN). Open-File Report, US Geological Survey, Report: OF 97-0467:10-13
  61. Hudnut K, Anderson G, Aspiotes A, King N, Moffitt R, Stark K (2002) GPS fault slip sensors. APEC Symposium on Confronting Urban Earthquakes/Seismic Early Warning. Academia Sinica, Taipei, pp 93–96
  62. Ihlmlé PF, Jordan TH (1994) Teleseismic search for slow precursors to large earthquakes. *Science* 266:1547–1551
  63. Ji C, Larson K, Tan Y, Hudnut K, Choi K (2004) Slip history of the 2003 San Simeon earthquake constrained by combining 1-Hz GPS, strong motion, and teleseismic data. *Geophys Res Lett* 31:L17608; doi:10.1029/2004GL020448
  64. Johanson I, Bürgmann R (2005) Creep and quakes on the northern transition zone of the San Andreas fault from GPS and InSAR data. *J Geophys Res* 32:L14306; doi:10.1029/2005GL023150
  65. Johnson K, Segall P (2004) Imaging the ramp-décollement geometry of the Chelungpu fault using coseismic GPS displacements from the 1999 Chi-Chi, Taiwan earthquake. *Tectonophysics* 378:123–139
  66. Johnson K, Segall P (2004) Viscoelastic cycle models of deep stress driven creep along the San Andreas Fault. *J Geophys Res* 109; doi:10.1029/2004JB003096
  67. Johnson K, Segall P (2005) A viscoelastic earthquake cycle model for Taiwan. *J Geophys Res* 110:B10404; doi:10.1029/2004JB003516
  68. Johnston M, Linde A (2002) Implications of crustal strain during convolitional, slow, and silent earthquakes. *Handbook of Earthquake and Engineering Seismology* 81A:589–605
  69. Jónsson S, Segall P, Pedersen R, Björnsson G (2003) Post-earthquake ground movements correlated to pore-pressure transients. *Nature* 424:179–183
  70. Jónsson S, Zebker H, Segall P, Amelung F (2002) Fault Slip Distribution of the 1999  $M_w$  7.1 Hector Mine, California, Earthquake, estimated from Satellite Radar and GPS Measurements. *Bull Seis Soc Amer* 92:1377–1389
  71. Kaverina A, Dreger D, Price E (2002) The combined inversion of seismic and geodetic data for the source process of the 16 October 1999  $M_w$  7.1 Hector Mine, California, earthquake. *Bull Seis Soc Amer* 92:1266–1280
  72. King N, Murray M, Prescott W, Clymer R, Romanowicz B (1994) The Bay Area Regional Deformation (BARD) permanent GPS array. *Eos Trans AGU* 75(44) Fall Meet Suppl:470
  73. Kreemer C, Blewitt G, Hammond W, Plag HP (2006) Global deformation from the great 2004 Sumatra–Andaman earthquake observed by GPS: Implications for rupture process and global reference frame. *Earth Planets Space* 58:141–148
  74. Langbein J (2004) Noise in two-color electronic distance meter measurements revisited. *J Geophys Res* 109:B04406; doi:10.1029/2003JB002819
  75. Langbein J, Gwyther RL, Hart RHG, Gladwin MT (1999) Slip-rate increase at Parkfield in 1993 detected by high-precision EDM and borehole tensor strainmeters. *Geophys Res Lett* 26:2529–2532
  76. Langbein J, Murray J, Snyder HA (2006) Coseismic and initial postseismic deformation from the 2004 Parkfield, California, earthquake, observed by Global Positioning System, electronic distance meter, creepmeters, and borehole strainmeters. *Bull Seismol Soc Amer* 96:S304–S320; doi:10.1785/0120050823
  77. Larson K (1995) Crustal deformation. *Rev Geophys* 33:371–378; doi:10.1029/95RG00439
  78. Larson K, Bodin P, Gombert J (2003) Using 1-Hz GPS data to measure deformations caused by the Denali fault earthquake. *Science* 300:1421–1424; doi:10.1126/science.1084531
  79. Larson K, Freymueller J, Philipson S (1997) Global plate velocities from the Global Positioning System. *J Geophys Res* 102:9961–9981
  80. Larson K, van Dam T (2000) Measuring postglacial rebound with GPS and absolute gravity. *Geophys Res Lett* 27:3925–3928
  81. Lay T, Kanamori H, Ammon C, Nettles M, Ward S, Aster R, Beck S, Bilek S, Brudzinski M, Butler R, DeShon H, Ekstrom G, Satake K, Sipkin S (2005) The great Sumatra–Andaman earthquake of 26 December 2004. *Science* 308:1127–1133
  82. Linde A, Gladwin M, Johnston M, Gwyther R, Bilham R (1996) A slow earthquake sequence on the San Andreas fault. *Nature* 383:65–68
  83. Lowry A (2006) Resonant slow fault slip in subduction zones forced by climatic load stress. *Nature*, 442:802–805
  84. Maerten F, Resor P, Pollard D, Maerten L (2005) Inverting for slip on three-dimensional fault surfaces using angular dislocations. *Bull Seismol Soc Amer* 95:1654–1665
  85. Manaker D, Bürgmann R, Prescott W, Langbein J (2003) Distribution of interseismic slip rates and the potential for significant earthquakes on the Calaveras fault, central California. *J Geophys Res* 108:B62287; doi:10.1029/2002JB001749
  86. Mao A, Harrison K, Dixon T (1999) Noise in GPS coordinate time series. *J Geophys Res* 104:2797–2816
  87. Mattia M, Rossi M, Guglielmino F, Aloisi M, Bock Y (2004) The shallow plumbing system of Stromboli Island as imaged from 1 Hz instantaneous GPS positions. *Geophys Res Lett* 31:L24610; doi:10.1029/2004GL021281
  88. Mazzotti S, James TS, Henton J, Adams J (2005) GPS crustal strain, postglacial rebound, and seismic hazard in eastern North America: The Saint Lawrence valley example. *J Geophys Res* 110:B11301; doi:10.1029/2004JB003590
  89. McCaffrey R (2005) Block kinematics of the Pacific–North America plate boundary in southwestern United States from inversion of GPS, seismological, and geologic data. *J Geophys Res* 110:B07401; doi:10.1029/2004JB003307
  90. Meade B, Hager B (2005) Block models of crustal motion in southern California constrained by GPS measurements. *J Geophys Res* 110:B03403; doi:10.1029/2004JB003209
  91. Meltzner A, Sieh K, Abrams M, Agnew D, Hudnut K, Avouac JP, Natawidjaja DH (2006) Uplift and subsidence associated with the great Aceh–Andaman earthquake of 2004. *J Geophys Res* 111:B02407; doi:10.1029/2005JB003891
  92. Miller M, Johnson D, Rubin C, Dragert H, Endo E, Humphreys E, Nabelek J, Qamar A (1997) GPS Monitoring of the Cascadia Margin: The Pacific Northwest Geodetic Array (PANGA). *Eos Trans AGU* 78(46) Fall Meet Suppl:167
  93. Miller M, Johnson D, Rubin C, Dragert H, Wang K, Qamar A, Goldfinger C (2001) GPS-determination of along-strike variation in Cascadia margin kinematics: Implications for relative plate motion, subduction zone coupling, and permanent deformation. *Tectonics* 20:161–176
  94. Miller M, Melbourne T, Johnson D, Sumner W (2002) Periodic

- slow earthquakes from the Cascadia subduction zone. *Science* 295:2423
95. Milne GA, Davis JL, Mitrovica JX, Scherneck HG, Johanson JM, Vermeer M, Koivula H (2001) Space-geodetic constraints on glacial isostatic adjustment in Fennoscandia. *Science* 291:2381–2385
  96. Miyazaki S, Larson K, Choi K, Hikima K, Koketsu K, Bodin P, Haase J, Emore G, Yamagiwa A (2004) Modeling the rupture process of the 2003 September 25 Tokachi-Oki (Hokkaido) earthquake using 1-Hz GPS data. *Geophys Res Lett* 31:L21603; doi:10.1029/2004GL021457
  97. Miyazaki S, Segall P, Fukuda J, Kato T (2004) Space time distribution of afterslip following the 2003 Tokachi-oki earthquake: Implications for variations in fault zone frictional properties. *Geophys Res Lett* 31:L06623; doi:10.1029/2003GL019410
  98. Miyazaki S, Segall P, McGuire J, Kato T, Hatanaka Y (2006) Spatial and temporal evolution of stress and slip rate during the 2000 Tokai slow earthquake. *J Geophys Res* 111:B03409; doi:10.1029/2004JB003426
  99. Mogi K (1958) Relations between the eruptions of various volcanoes and the deformations of the ground surfaces around them. *Bull Seismol Soc Amer* 36:111–123
  100. Mora P, Baldi P, Casula G, Fabris M, Ghirotti M, Mazzini E, Pesci A (2003) Global Positioning Systems and digital photogrammetry for the monitoring of mass movements: Application to the Ca' di Malta landslide (northern Apennines, Italy). *Eng Geol* 68:103–121
  101. Murray J, Langbein J (2006) Slip on the San Andreas fault at Parkfield, California, over two earthquake cycles, and the implications for seismic hazard. *Bull Seismol Soc Amer* 96:S283–S303
  102. Murray J, Segall P (2005) Spatiotemporal evolution of a slip-rate increase on the San Andreas fault near Parkfield, CA. *J Geophys Res* 110:B09407; doi:10.1029/2005JB003651
  103. Murray J, Segall P, Cervelli P, Prescott W, Svarc J (2001) Inversion of GPS data for spatially variable slip-rate on the San Andreas Fault near Parkfield, CA. *Geophys Res Lett* 28:359–362
  104. Nikolaidis R, Bock Y, de Jonge P, Shearer P, Agnew D, Domelaar M (2001) Seismic wave observations with the Global Positioning System. *J Geophys Res* 106:21897–21916
  105. Obara K (2002) Nonvolcanic deep tremor associated with subduction in southwest Japan. *Science* 296:1679–1681
  106. Obara K, Hirose H, Yamamizu F, Kasahara K (2004) Episodic slow slip events accompanied by non-volcanic tremors in southwest Japan subduction zone. *Geophys Res Lett* 31; doi:10.1029/2004GL020848
  107. Ohta Y, Freymueller J, Hreinsdóttir S, Suito H (2006) A large slow slip event and the depth of the seismogenic zone in the south central Alaska subduction zone. *Earth Plan Sci Lett* 247:108–116
  108. Ohta Y, Kimata F, Sagiya T (2004) Reexamination of the interplate coupling in the Tokai region, central Japan, based on the GPS data in 1997–2002. *Geophys Res Lett* 31:L24604; doi:10.1029/2004GL021404
  109. Okada Y (1985) Surface deformation due to shear and tensile faults in a half-space. *Bull Seismol Soc Amer* 75:1135–1154
  110. Okada Y (1992) Internal deformation due to shear and tensile faults in a half-space. *Bull Seismol Soc Amer* 82:1018–1040
  111. Owen S, Segall P, Lisowski M, Miklius A, Denlinger R, Sako M (2000) Rapid deformation of Kilauea volcano: GPS measurements between 1990 and 1996. *J Geophys Res* 105:18983–18998
  112. Owen S, Segall P, Lisowski M, Murray M, Bevis M, Foster J (2000) The January 30, 1997 eruptive event on Kilauea Volcano, Hawaii, as monitored by continuous GPS. *Geophys Res Lett* 27:2757–2760
  113. Ozawa S, Murakami M, Fujiwara S, Tobita M (1997) Synthetic aperture radar interferogram of the 1995 Kobe earthquake and its geodetic inversion. *Geophys Res Lett* 24:2327–2330
  114. Ozawa S, Suito H, Imakiire T, Murakami M (2007) Spatiotemporal evolution of aseismic interplate slip between 1996 and 1998 and between 2002 and 2004, in Bungo channel, southwest Japan. *J Geophys Res* 112:B05409; doi:10.1029/2006JB004643
  115. Park J, Anderson K, Aster R, Butler R, Lay T, Simpson D (2005) Global seismographic network records the great Sumatra–Andaman earthquake. *Eos Trans AGU* 86:57, 60–61
  116. Park J, Song TR, Tromp J, Okal E, Stein S, Roullet G, Clevede E, Laske G, Kanamori H, Davis P, Berger J, Braitenberg C, Van Camp M, Lei X, Sun H, Xu H, Rosat S (2005) Earth's free oscillations excited by the 26 December 2004 Sumatra–Andaman earthquake. *Science* 308:1139–1144
  117. Park K, Nerem RS, Davis JL, Schenewerk MS, Milne GA, Mitrovica JX (2002) Investigation of glacial isostatic adjustment in the northeast US using GPS measurements. *Geophys Res Lett* 29:1509; doi:10.1029/2001GL013782
  118. Peltzer G, Rosen P, Rogez F, Hudnut K (1998) Poroelastic rebound along the Landers 1992 earthquake surface rupture. *J Geophys Res* 103:30131–30145
  119. Poland M, Hamburger M, Newman A (2006) The changing shapes of active volcanoes: History, evolution, and future challenges for volcano geodesy. *J Volc Geotherm Res* 150:1–13
  120. Polet J, Kanamori H (2000) Shallow subduction zone earthquakes and their tsunamigenic potential. *Geophys J Int* 142:684–702
  121. Pollitz F (2005) Transient rheology of the upper mantle beneath central Alaska inferred from the crustal velocity field following the 2002 Denali earthquake. *J Geophys Res* 110:B08407; doi:10.1029/2005JB003672
  122. Pollitz F, Wicks C, Thatcher W (2001) Mantle flow beneath a continental strike-slip fault: postseismic deformation after the 1999 Hector Mine earthquake. *Science* 293:1814–1818
  123. Pratt T (2006) Do Episodic Tremor and Slip (ETS) Events Affect Seismicity in the Northern Cascadia Subduction Zone? *Eos Trans AGU* 87(52), Fall Meet Suppl, Abstract T54A-04
  124. Prawirodirdjo L, Bock Y (2004) Instantaneous global plate motion model from 12 years of continuous GPS observations. *J Geophys Res* 109:B08405; doi:10.1029/2003JB002944
  125. Prescott W, Savage J, Svarc J, Manaker D (2001) Deformation across the Pacific–North America plate boundary near San Francisco, California. *J Geophys Res* 106:6673–6682
  126. Pritchard M, Norabuena E, Ji C, Boroschek R, Comte D, Simons M, Dixon T, Rosen P (2007) Geodetic, teleseismic, and strong motion constraints on slip from recent southern Peru subduction zone earthquakes. *J Geophys Res* 112:B03307; doi:10.1029/2006JB004294
  127. Reid HF (1910) The California Earthquake of April 18, 1906. In: Report of the state earthquake investigation commission, vol 2, Carnegie Institute, Washington DC

128. Reid M, LaHusen R, Schmidt K (2004) Capturing 3-D displacements in active landslides using GPS. *Abstracts with Programs. Geol Soc Amer* 36:331
129. Rhie J, Dreger D, Bürgmann R, Romanowicz B (2007) Slip of the 2004 Sumatra–Andaman earthquake from joint inversion of long-period global seismic waveforms and GPS static offsets. *Bull Seismol Soc Amer* 97:S115–S127
130. Rogers G, Dragert H (2003) Episodic tremor and slip on the Cascadia subduction zone: the chatter of silent slip. *Science* 300:1942–1943
131. Sagiya T (2004) Interplate coupling in the Kanto district, central Japan, and the Boso Peninsula silent earthquake in May 1996. *Pure Appl Geophys* 161:2327–2342; doi:10.1007/s00024-004-2566-6
132. Salichon J, Lundgren P, Delouis B, Giardini D (2004) Slip history of the 16 October 1999  $M_w$  7.1 Hector Mine earthquake (California) from the inversion of InSAR, GPS, and teleseismic data. *Bull Seismol Soc Amer* 94:2015–2027
133. Savage J, Prescott W (1978) Asthenosphere readjustment and the earthquake cycle. *J Geophys Res* 83:3369–3376
134. Savage J, Svarc J, Yu SB (2005) Postseismic relaxation and transient creep. *J Geophys Res* 110:B11402; doi:10.1029/2005JB003687
135. Schmalzle G, Dixon T, Malservisi R, Govers R (2006) Strain accumulation across the Carrizo segment of the San Andreas fault, California: Impact of laterally varying crustal properties. *J Geophys Res* 111:B05403; doi:10.1029/2005JB003843
136. Schmidt D, Bürgmann R, Nadeau R, d'Alessio M (2005) Distribution of aseismic slip rate on the Hayward fault inferred from seismic and geodetic data. *J Geophys Res* 110:B08406; doi:10.1029/2004JB003397
137. Segall P (2002) Integrating geologic and geodetic estimates of slip rate on the San Andreas Fault system. *Int Geol Rev* 44:62–82
138. Segall P, Harris R (1986) Slip deficit on the San Andreas fault at Parkfield, California, as revealed by inversion of geodetic data. *Science* 233:1409–1413
139. Segall P, Desmarais E, Shelly D, Miklius A, Cervelli P (2006) Earthquakes triggered by silent slip events on Kilauea volcano, Hawaii. *Nature* 442:71–74; doi:10.1038/nature04938
140. Sella G, Dixon T, Mao A (2002) REVEL: A model for recent plate velocities from space geodesy. *J Geophys Res* 107:B42081; doi:10.1029/2000JB000033
141. Sella G, Stein S, Dixon T, Craymer M, James T, Mazzotti S, Dokka R (2007) Observation of glacial isostatic adjustment in “stable” North America with GPS. *Geophys Res Lett* 34:L02306; doi:10.1029/2006GL027081
142. Shelly D, Beroza G, Ide S (2007) Non-volcanic tremor and low-frequency earthquake swarms. *Nature* 446:305–307
143. Shelly D, Beroza G, Ide S, Nakamura S (2006) Low-frequency earthquakes in Shikoku, Japan, and their relationship to episodic tremor and slip. *Nature* 442:188–191
144. Shimada S, Bock Y (1992) Crustal deformation measurements in central Japan determined by a Global Positioning System fixed point network. *J Geophys Res* 97:12437–12455
145. Shimada S, Fujinawa Y, Sekiguchi S, Ohmi S, Eguchi T, Okada Y (1990) Detection of a volcanic fracture in Japan using Global Positioning System measurements. *Nature* 343:631–633
146. Sieh K, Jahns RH (1984) Holocene activity of the San Andreas fault at Wallace Creek, California. *Geol Soc Amer Bull* 95:883–896
147. Silver P, Bock Y, Agnew D, Henry T, Linde A, McEvelly T, Minster JB, Romanowicz B, Sachs I, Smith R, Solomon S, Stein S (1999) A Plate Boundary Observatory. *IRIS Newsletter XVI*:3–9
148. Simons M, Fialko Y, Rivera L (2002) Co seismic deformation from the 1999  $M_w$  7.1 Hector Mine, California, earthquake as inferred from InSAR and GPS observations. *Bull Seismol Soc Amer* 92:1390–1402
149. Sims JD (1990) Geologic map of the San Andreas fault in the Parkfield 7.5-minute quadrangle, Monterey and Fresno counties, California. *US Geol Surv Misc Field Studies Map MF-2115*
150. Snay R, Cline M, Dillinger W, Foote R, Hilla S, Kass W, Ray J, Rohde J, Sella G, Soler T (2007) Using global positioning system-derived crustal velocities to estimate rates of absolute sea level change from North American tide gauge records. *J Geophys Res* 112:B04409; doi:10.1029/2006JB004606
151. Squarzonni C, Delacourt C, Allemand P (2005) Differential single-frequency GPS monitoring of the La Valette landslide (French Alps). *Eng Geol* 79:215–229
152. Stein S, Okal E (2005) Speed and size of the Sumatra earthquake. *Nature* 434:581–582
153. Subarya C, Chlieh M, Prawirodirdjo L, Avouac JP, Bock Y, Sieh K, Meltzner AJ, Natawidjaja DH, McCaffrey R (2006) Plate-boundary deformation associated with the great Sumatra–Andaman earthquake. *Nature* 440:46–51
154. Suwa Y, Miura S, Hasegawa A, Sato T, Tachibana K (2006) Interplate coupling beneath NE Japan inferred from three-dimensional displacement field. *J Geophys Res* 111:B04402; doi:10.1029/2004JB003203
155. Szeliga W, Melbourne T, Miller M, Santillan V (2004) Southern Cascadia episodic slow earthquakes. *Geophys Res Lett* 31:L16602; doi:10.1029/2004GL020824
156. Thatcher W (1983) Nonlinear strain buildup and the earthquake cycle on the San Andreas fault. *J Geophys Res* 88:5893–5902
157. Thatcher W (1995) Microplate versus continuum descriptions of active tectonic deformation. *J Geophys Res* 100:3885–3894
158. Thatcher W (2007) Microplate model for the present-day deformation of Tibet. *J Geophys Res* 112:B01401; doi:10.1029/2005JB004244
159. Thurber C, Zhang H, Waldhauser F, Hardebeck J, Michael A, Eberhart-Phillips D (2006) Three-dimensional compressional wavespeed model, earthquake relocations, and focal mechanisms for the Parkfield, California, region. *Bull Seism Soc Amer* 96:538–549
160. Vigny C, Simons WJF, Abu S, Bamphenyu R, Satirapod C, Choosakul N, Subarya C, Socquet A, Omar K, Abidin HZ, Ambrosius BAC (2005) Insight into the 2004 Sumatra–Andaman earthquake from GPS measurements in southeast Asia. *Nature* 436:201–206
161. Wahba G (1990) *Spline Models for Observational Data*. SIAM, Philadelphia PA
162. Wald D, Heaton T (1994) Spatial and temporal distribution of slip for the 1992 Landers, California, earthquake. *Bull Seismol Soc Amer* 84:668–691
163. Waldhauser F, Ellsworth W (2000) A double-difference earthquake location algorithm: Method and application to the northern Hayward fault, California. *Bull Seismol Soc Amer* 90:1353–1368
164. Wallace L, Beavan J, McCaffrey R, Darby D (2004) Subduction zone coupling and tectonic block rotations in the North Is-

- land, New Zealand. *J Geophys Res* 109:B12406; doi:10.1029/2004JB003241
165. Wang GQ, Boore D, Tang G, Zhou X (2007) Comparisons of ground motions from collocated and closely spaced one-sample-per-second Global Positioning System and accelerometer recordings of the 2003 M 6.5 San Simeon, California, earthquake in the Parkfield region. *Bull Seismol Soc Amer* 97:76–90
166. Williams S, Bock Y, Fang P, Jamason P, Nikolaidis R, Prawirodirdjo L, Miller M, Johnson D (2004) Error analysis of continuous GPS position time series. *J Geophys Res* 109. doi:10.1029/2003JB002741
167. Working Group on California Earthquake Probabilities (2003) Earthquake probabilities in the San Francisco Bay Region: 2002-2031. US Geol Surv Open File Report 03-214
168. Wright T, Lu Z, Wicks C (2004) Constraining the slip distribution and fault geometry of the  $M_w$  7.9, 3 November 2002, Denali Fault earthquake with interferometric synthetic aperture radar and Global Positioning System data. *Bull Seismol Soc Amer* 94:175–189
169. Yagi Y, Kikuchi M, Nishimura T (2003) Co-seismic slip, post-seismic slip, and largest aftershock associated with the 1994 Sanriku-haruka-oki, Japan, earthquake. *Geophys Res Lett* 30:2177; doi:10.1029/2003GL018189
170. Zhang J, Bock Y, Johnson H, Fang P, Williams S, Genrich J, Wdowinski S, Behr J (1997) Southern California permanent GPS geodetic array: Error analysis of daily position estimates and site velocities. *J Geophys Res* 102:18035–18055
171. Zhang PZ, Shen Z, Wang M, Gan W, Bürgmann R, Molnar P, Wang Q, Niu Z, Sun J, Wu J, Hanrong S, Xinzhaoy Y (2004) Continuous deformation of the Tibetan Plateau from global positioning system data. *Geol* 32:809–812

### Books and Reviews

- Bolt B (1999) *Earthquakes*, 4th edn. W H Freeman and Company, New York
- Menke W (1989) *Geophysical data analysis: Discrete inverse theory*, rev edn. In: Dmowska R, Holton J (eds) *International geophysics series* 45. Academic Press, San Diego
- Misra P, Enge P (2001) *Global Positioning System: Signals, measurements, and performance*. Ganga-Jamuna Press, Lincoln, MA
- Schwartz S, Rokosky J (2007) Slow slip events and seismic tremor at circum-pacific subduction zones. *Rev Geophys* 45:RG3004
- Shearer P (1999) *Introduction to seismology*. Cambridge University Press, New York
- Strang G, Borre K (1997) *Linear algebra, geodesy, and GPS*. Wellesley-Cambridge Press, Wellesley, MA

## Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures

P. MARTIN MAI

Swiss Seismological Service, Institute of Geophysics,  
ETH, Zürich, Switzerland

### Article Outline

Glossary

Definition of the Subject

Introduction

Characterizing Earthquake Source Complexity

Wave Propagation in Complex Media:

Path and Site Effects

Ground-Motion Scaling Relations

Future Directions

Acknowledgments

Bibliography

### Glossary

**Attenuation relation (ground-motion prediction equation)** The term “attenuation relation” is a former shorthand notation in earthquake engineering for “empirical ground-motion attenuation relationship”, now referred to as “ground-motion prediction equation” (GMPE). Attenuation relations represent empirical scaling equations that relate observed *ground-motion intensity measures* to parameters of the earthquake source, the wave propagation from the source to the observer and the site response at the observer location.

**Dynamic rupture model** Dynamic rupture models build a physical understanding of the earthquake rupture based on the material properties around the source volume, and the initial and boundary conditions for the forces/stresses acting on the fault plane. The distribution of on-fault *slip-rate* vectors and the temporal rupture evolution is obtained by solving the elastodynamic equations of motion under an assumed constitutive law (friction model), considering the energy balance at the crack tip (Chap. 11 in [6]). See also *kinematic rupture model*.

**Ground motion intensity measures** Earthquake shaking due to seismic waves, observed at recording sites or experienced by people and structures, is commonly reported in terms of various scalar intensity measures that capture parts of the transient wave-field. Seismogram-based ground-motion intensity measures are,

for instance, peak ground acceleration (*PGA*) and peak ground velocity (*PGV*), while the modified Mercalli intensity (*MMI*) is a damage-related measure. In earthquake engineering, ground-motion intensities are often reported as *response spectra*: the response of an idealized building (modeled as a single-degree-of-freedom oscillator) of given eigenperiod  $T$  and damping  $\zeta$  (usually 5%) to a given ground-motion time series. Spectral acceleration ( $S_A$ ), spectral velocity ( $S_V$ ) and spectral displacement ( $S_D$ ) are analyzed considering the period of the structure.

**Ground motion uncertainty** In ground-motion prediction for engineering purposes, random (aleatory) variability and scientific (epistemic) uncertainty are distinguished. The latter is due to incomplete knowledge and/or limited data, and is captured by alternative empirical attenuation relations or different ground-motion simulation strategies. Aleatory variability is quantified in terms of a standard deviation of an attenuation relation or by a large number of model realizations within a particular simulation method. The distinction between aleatory variability and epistemic uncertainty is particularly useful in probabilistic seismic hazard analysis (PSHA).

**Kinematic rupture model** A kinematic rupture model characterizes the time-dependent displacement field on the rupture plane without considering the forces or stresses acting on the fault and causing its motion. The rupture process is completely specified by the spatio-temporal distribution of the slip vector, the local *slip-velocity* function on the fault, and the *rupture velocity* with which the rupture propagates over the fault plane. See also *dynamic rupture model*.

**Path effects** Seismic waves propagating through the Earth are sensitive to the detailed geologic structure along the wave path, generating pronounced waveform complexities. Considering crust and upper-mantle structure (relevant for near-field ground motions) three major elements to path effects are distinguished in practice: (a) waves in a flat-layered attenuating Earth; (b) basins and other deterministic deviations from a flat-layered model; (c) random heterogeneities in the three-dimensional velocity-density structure. The distinction between path effects and *site effects* is often ambiguous.

**Rise time** The rise time (or slip duration)  $\tau_r$  measures how long each point on the fault moves during the rupture process, and must not be confused with the *rupture duration*. The rise time is related to *the slip-velocity function*, and is usually measured as the time it takes to attain 5–95% of the final slip at each point. For

simple parametric slip-functions (e. g. boxcar, isosceles triangle or combinations thereof), the rise time is generally given by the width of this function.

**Rupture duration** The rupture duration characterizes the total time for the earthquake rupture process to complete, starting at the nucleation point (hypocenter) and lasting until the last point on the rupture plane stops sliding. Rupture duration therefore depends on **rupture velocity** and scales with the size (source dimension) of the earthquake.

**Rupture velocity** Earthquake ruptures, either modeled as propagating cracks or slip-pulses, expand over the fault plane at rupture speeds ( $v_r$ ) close to the local shear-wave velocity ( $v_s$ ), typically in the range  $0.5 \cdot v_s \leq v_r \leq 0.9 \cdot v_s$ , or about 1.0–3.5 km/sec for crustal earthquakes. However, the crack tip, the transition region from unbroken, intact rock to the currently slipping zone, does not necessarily travel at constant rupture speed. Rupture velocity may locally slow down or accelerate, even to super-shear velocities (in which case the crack front travels at speeds faster than the local shear-wave velocity), depending on the initial and boundary conditions that govern the dynamic rupture process.

**Site effects** Site effects refer to wave-propagation effects in the immediate proximity to the observation point; they are distinguished from **path effects** which comprise the complete path from the source to the receiver (although the boundary between these two is often ill defined). The local sedimentary cover, topography, strong geologic contrasts or water-table variations may contribute to site effects that modify the incoming “bedrock” seismic motions.

**Slip distribution** The slip distribution represents the cumulative slip on each point on the fault acquired during the co-seismic rupture process (i. e. small contributions from post-seismic slip episodes are ignored). A slip distribution for an earthquake is computed from the space-time integration of **slip-velocity functions** on the rupture plane.

**Slip-velocity function (Slip-rate function)** Each point participating in the rupture process experiences a time-dependent slip history during which the two sides of the fault go through a stage of acceleration, stable sliding, deceleration, and final stopping. This local displacement trajectory is often represented in terms of a slip-velocity function (or slip function) whose details depend on the dynamic rupture process and the constitutive behavior of the host rock. Slip-rate functions are often modeled using simple parametric functions.

**Source effects** Amplitudes and waveform character of seismic waves are strongly affected by source effects, i. e. by the details of the earthquake rupture process. Far-field signals carry the signature of the overall “point-source” earthquake source mechanism; near-field recordings are very sensitive to the spatio-temporal details of the rupture process, characterized in a finite-fault source model either as **kinematic** or **dynamic rupture model**.

**Static stress drop** The static stress drop  $\Delta\sigma$  represents the difference between the initial and final stress across the fault before and after an earthquake, and is related to slip on the fault. It is defined, based on a shear crack with uniform stress drop, as  $\Delta\sigma = C \cdot \mu \cdot D/L_c$ , where  $\mu$  is the shear-modulus,  $D$  the mean slip over the fault,  $L_c$  a characteristic length scale (usually the smallest dimension of the rupturing fault), and  $C$  a constant of order unity which depends on the source geometry. Using the fault width  $W$  as characteristic length, static stress drop is related to the seismic moment,  $M_o = \mu \cdot L \cdot W \cdot D = C \cdot \Delta\sigma \cdot A^{3/2}$ , where  $A$  is fault area, and  $L$  is fault length. Inferred values of static stress drop are in the range of 0.1–10 MPa, independent of seismic moment, leading to the generally assumed self-similar constant stress-drop scaling (see ► **Earthquake Scaling Laws**). The static stress drop must not be confused with the dynamic stress drop [164] which captures the time-dependent stress change on a point of the fault during the dynamic faulting event, and may be significantly higher or lower than the static stress drop.

## Definition of the Subject

The accurate prediction of the level and variability of (potentially damaging) near-source strong-ground motions in future earthquakes is one of the key challenges for seismologists and earthquake engineers. The increasing number of near-source recordings collected by dense strong-motion networks exemplifies the inherent complexity of near-field ground shaking, governed by a number of (partially interacting) physical processes. Characterizing, quantifying, and modeling (either by means of empirical scaling relations or by numerical simulations) ground-motion complexity requires the joint investigation of three dominant ingredients: (I) the physics of earthquake rupture; (II) the details of wave-propagation in heterogeneous media; (III) the effects of local site conditions.

This article discusses briefly the beginnings of strong-motion seismology and the recognition of ground-motion complexity. Using two well recorded recent earthquakes, I introduce observational aspects of near-field ground shak-

ing and the basic mathematical description for computing ground motion. The article proceeds by describing each of the three “ground-motion ingredients” in some detail, but does not attempt to provide an in-depth review of all the scientific advancements in these fields. Rather, I explain the key elements for characterizing and modeling ground-motion complexity, supplemented with a concise overview of the underlying physical processes. Current research increasingly incorporates advanced physical concepts into standard practice, therefore leading to improved strong-motion simulation approaches to accurately predict intensity and variability of near-source shaking.

## Introduction

As early as 1910 Reid [148] identified strong spatial variations of seismic shaking due to the 1906 San Francisco earthquake, which he correctly attributed to the specific geologic conditions at the locations from which shaking intensities were reported. The space-time complexity of the seismic wavefield in the vicinity of the causative fault became more widely recognized since the beginnings of strong-motion seismology in the 1930s and the establishment of strong-motion networks in the 1960s [9]. Observed peak accelerations frequently reach 1 g, occasionally even exceed 2 g, where nearby stations often show not only much lower peak amplitudes but also very different waveform character. Since the advent of modern digital instruments in the 1990s and corresponding online near-real-time databases, seismologists and earthquake engineers have access to high-quality recordings which irrefutably show the complexity of earthquake shaking. Given a particular location, near-source ground motions vary strongly for different earthquakes, and so do the motions for a single event when recorded at different sites.

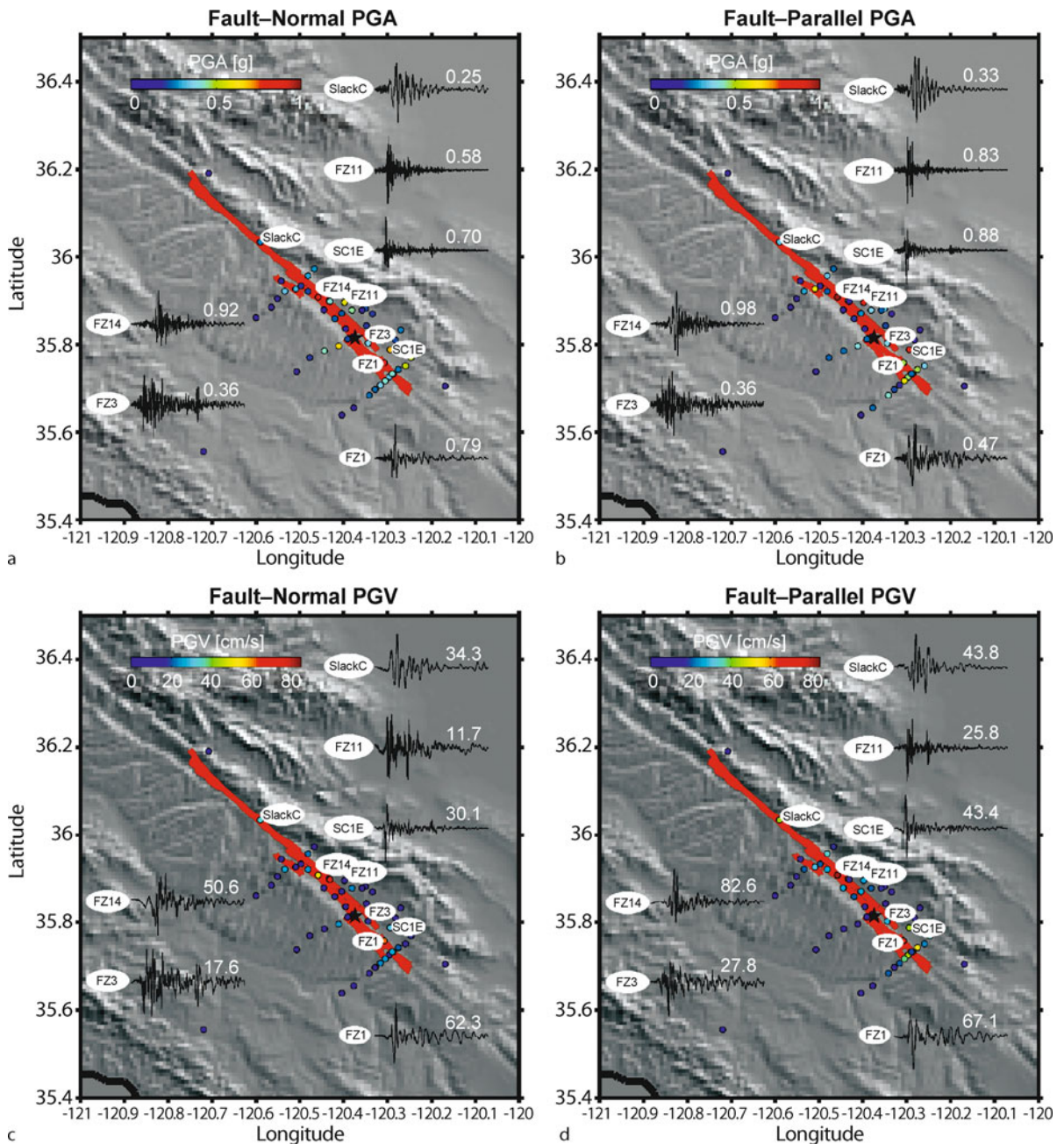
While hundreds of earthquakes happen annually in the magnitude range  $6.0 \leq M \leq 6.9$ , only about 15 events occur in the range  $7.0 \leq M \leq 7.9$ , most of which happening in remote areas and do not cause large damage. Those events are typically recorded at rather large distances (so called teleseismic events) and have weak (low amplitude) motions. Their ground-motion variability is due to differences in the large-scale wave paths through the Earth and due to the properties of the point-source representation for far-distant earthquakes. Weak-motion recordings are also obtained from small nearby earthquakes. In this article, however, we will be concerned with strong-motion observation in the near-field of (large) earthquakes. Such observations are much harder to obtain than teleseismic recordings because we cannot anticipate in detail where future earthquakes may happen (in order to opti-

mize the instrumentation) and also because large crustal earthquakes are rather infrequent.

Investigating the characteristics of near-source ground-motions is of great importance for earthquake engineers who are concerned with seismically safe design, and for seismologists who study the physical processes leading to ground-motion complexity. To illustrate the large ground-motion variability in a single earthquake, I plot horizontal peak ground accelerations and peak ground velocities recorded at 47 sites in the Sept 28, 2004, M 6.0 Parkfield earthquake (Fig. 1), contrasted with ground-motion intensities (*PGA*, *PGV*, *PGD*) for the Sept 20, 1999, M 7.6 Taiwan (Chi-Chi) earthquake, shown at 441 locations (Figs. 2–4) (waveforms from the COSMOS database, <http://db.cosmos-eq.org>). First-order observations from these data are: (i) the maximum *PGA* is larger for the smaller earthquake while the highest *PGV* values are roughly identical; (ii) ground-motions tend to be larger close to the fault trace and decay in amplitude with increasing distance from the fault. However, closer inspection reveals strong variability in the near-field motions for each earthquake. The Parkfield data exhibit ground-motion differences between the two horizontal components of motion (fault-parallel and fault-normal, both for *PGA* and *PGV*) and large variability between neighboring sites. This site dependence is even more pronounced for the Chi-Chi event. *PGA*-values are highest for sites very close to the fault, on the hanging-wall (east of the fault trace) and at a few locations in the northward and southward extension of the fault-trace. This pattern changes, however, for the recorded *PGV* and *PGD*-values which are largest on the foot-wall (west side of the fault trace) and towards the northern end of the rupture.

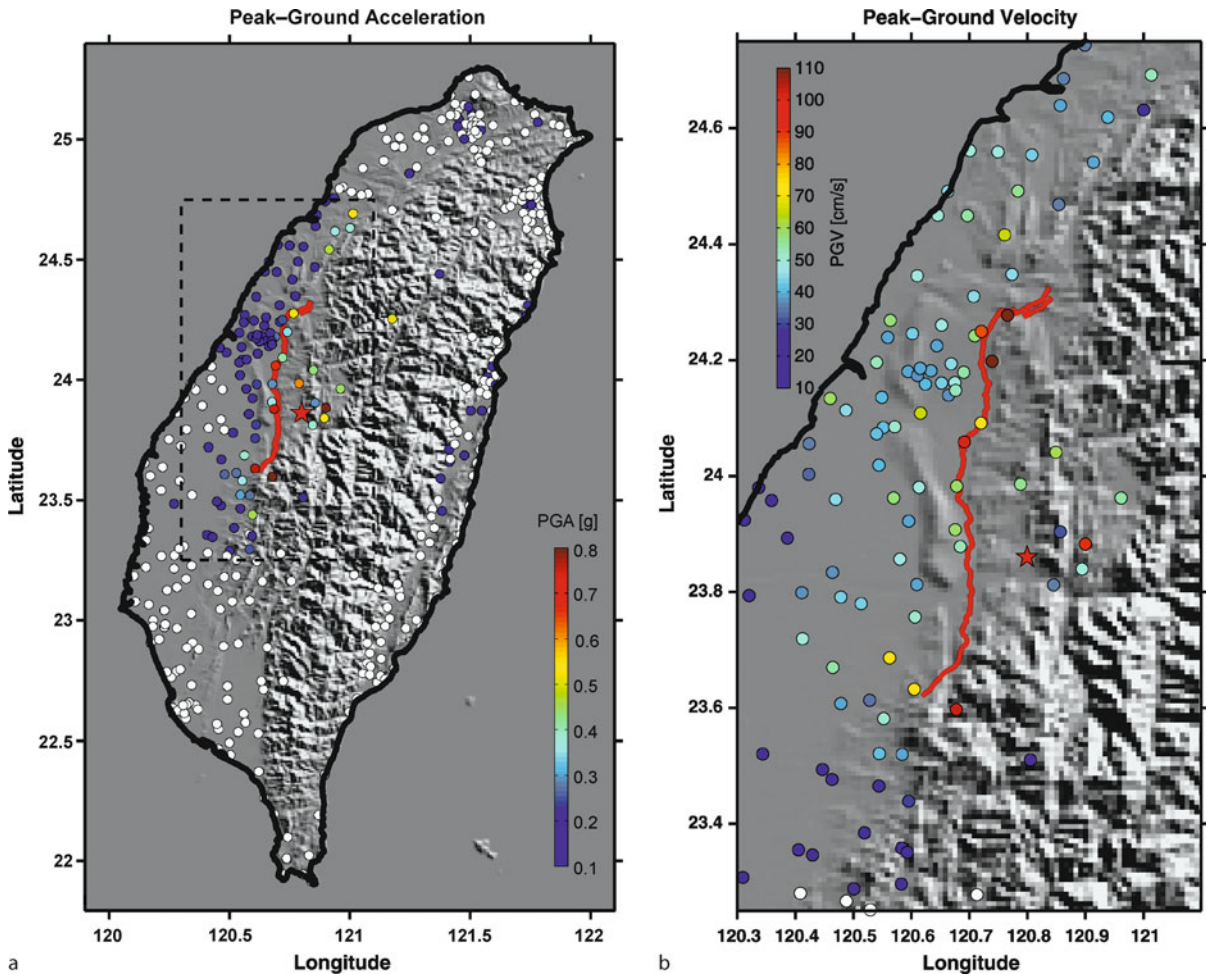
In addition to the scalar measures of shaking amplitude (intensity), Figs. 1–4 display recorded waveforms for selected sites which further illustrate the complexity and variability of near-field motions. Acceleration time series accentuate the high-frequency content of ground-motion and display an almost “shaped random noise” character, with amplitudes that increase rapidly in the beginning and then taper off gradually in the seismic coda (the incoherent wavefield after the arrival of prominent seismic phases). In contrast, velocity waveforms usually show very distinct arrivals and energetic pulses (directivity pulses) whose frequency-dependent amplitude and waveform character change with source-site geometry and rupture propagation direction. Their waveform character also varies strongly between closely spaced stations for a given event, illustrating the importance of localized source properties and site effects. Note also the large differences in ground-motion amplitudes and waveform shapes between





Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures, Figure 1

Ground-motion intensities (*colored circles*) for the Sept 28, 2004 Parkfield (M 6.0) earthquake, recorded at near-field stations, and selected waveforms (data from COSMOS database). The *black star* denotes the epicenter, (Lon =  $-120.374$ , Lat =  $35.815$ ) *red-lines* show the mapped fault trace of the San Andreas Fault in the Parkfield area, the *gray background* displays a shaded relief map. **a** Peak ground accelerations (PGA) of fault-normal component; **b** PGA of fault-parallel component; **c** Peak ground velocity (PGV) of fault-normal component; **d** PGV of fault-parallel component. Seismic traces are shown for 30s, the small number indicates the corresponding PGA or PGV value of that record



Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures, Figure 2

Ground-motion intensities (*colored circles*) for the Sept 20, 1999 Taiwan (M 7.6) earthquake, recorded at 441 locations (data from COSMOS database). The *red star* denotes the epicenter (Lon = 120.7995, Lat = 23.860), *red-lines* show the mapped fault trace of the Chelungpu fault, *heavy black lines* trace the coast line, the *gray background* displays a shaded relief map. **a** Large-scale view of PGA, white dots are sites with PGA < 0.1 g. **b** Zoomed view for PGV (white dots mark PGV < 10 cm/s). **c** Zoomed view for PGD (white dots mark PGD < 10 cm)

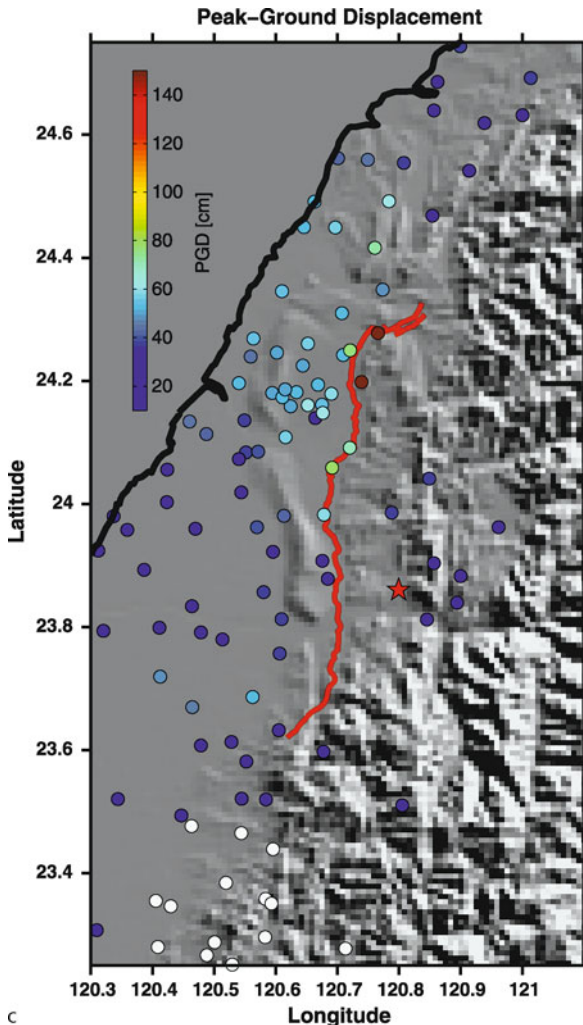
the two events for stations at similar distance to the fault, indicating the dependence on earthquake source properties of near-field ground motions. Section “**Future Directions**” compares the near-field recordings for these two earthquakes against empirical predictions for several published attenuation relations to further illustrate ground-motion complexity.

The variability in shaking intensity and the complexity in near-fault seismograms results from three physical processes: (I) the complex dynamics of earthquake rupture and the associated radiation of seismic waves; (II) the propagation of these seismic waves through the heterogeneous Earth; (III) the interaction of the seismic wave-

field with local geology/morphology, referred to as the site conditions at each observation points. Mathematically, the time-dependent ground displacement  $u_k(t)$  at a particular location  $k$  is described as

$$u_k(t) = s(t) * g_k(t) * l_k(t) \quad (1)$$

where  $*$  denotes the convolution operator,  $s(t)$  represents the **source effects** due to the earthquake rupture process,  $g_k(t)$  describes the **path effects** due to wave propagation from the source to site  $k$ , and comprises  $l_k(t)$  the **local site effects** due to the small-scale geological conditions at the  $k$ th observation point. Equation (1) quantifies ground-motion generation, omitting for simplicity



c  
Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures, Figure 2  
(continued)

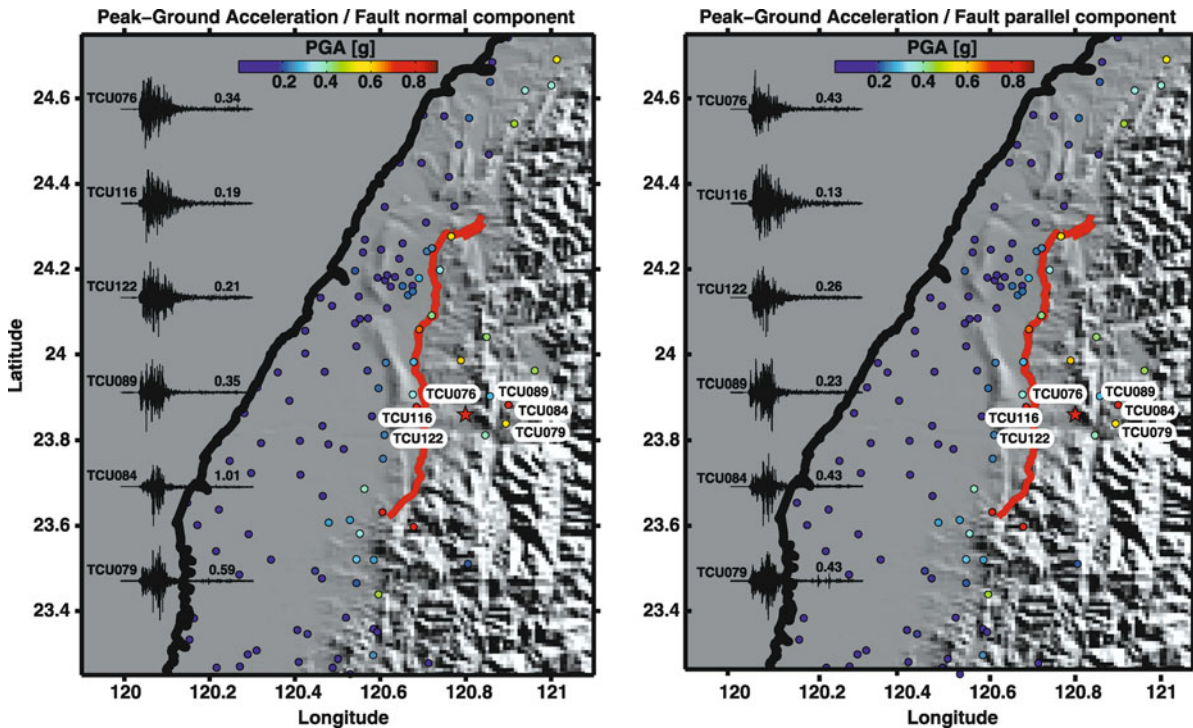
an additional instrument response  $i_k(t)$  that modulates the seismic recording. The earthquake-source contribution  $s(t)$  can be further subdivided into effects originating from the local time-dependent particle motion on the fault,  $p_{ij}(t)$ , and from rupture finiteness,  $f(t)$ . The term  $g_k(t)$  represents the Earth transfer function and contains contributions from layered Earth structure and seismic-wave attenuation, but may also comprise effects due to random heterogeneities in the Earth and/or basin and topographic structures. The factor  $l_k(t)$  describes the effects of soil structure and may also account for non-linear soil behavior.

In the following, I will use Eq. (1) as a “roadmap” for this article to illustrate the various factors of ground-

motion complexity. In Sect. “[Characterizing Earthquake Source Complexity](#)”, I present methods and relevant parameters to quantify the earthquake rupture process. Wave propagation in complex media, both deterministic and stochastic, are described in Sect. “[Wave Propagation in Complex Media: Path and Site Effects](#)”. A large body of literature exists for each of these topics, both from an observational/experimental view and from theoretical work; it is beyond the scope of this article to provide an in-depth review of all relevant material. Instead, I will focus on some of the key aspects of earthquake ruptures and waves in inhomogeneous media that are most relevant for understanding the complexity of near-field ground motions. Note also that a number of specialized articles in this encyclopedia provide more detailed information on earthquake source physics (see ▶ [Earthquake Nucleation Process](#), ▶ [Earthquake Scaling Laws](#)) and wave-propagation phenomena in complex media (see ▶ [Seismic Wave Propagation in Media with Complex Geometries, Simulation of](#), ▶ [Seismic Waves in Heterogeneous Earth, Scattering of](#)). Section “[Wave Propagation in Complex Media: Path and Site Effects](#)” also includes various aspects of local site conditions that lead to pronounced site effects, in particular non-linear site phenomena. Section “[Ground-Motion Scaling Relations](#)” focuses on empirical ground-motion prediction equations (GMPE’s). Since many studies have been published on non-linear site effects and ground-motion prediction equations, both from the seismology and the earthquake-engineering communities, an extensive review of the developments in these fields is not attempted. Instead, I concentrate on non-linear soil behavior directly beneath a site of interest that substantially affects the ground-shaking levels, and summarize some of the latest findings related to ground-motion attenuation relations. The article concludes with an outlook onto future tasks and challenges for characterizing, quantifying and predicting ground-motion complexity.

### Characterizing Earthquake Source Complexity

Geologic faults are generally geometrically complex multi-scale structures, characterized by one or more main fault strands, with associated subsidiary branches (fault segments) that form three-dimensional (3D) fault networks. However, in many cases and for almost all practical purposes, faults are approximated as planar surfaces. Using this plane-fault approximation, Fig. 5 illustrates the three main factors affecting near-source ground-motion complexity – source, path and site effects. This Section focuses on the earthquake rupture process, i.e. the source, and its properties important for ground-motion genera-



Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures, Figure 3

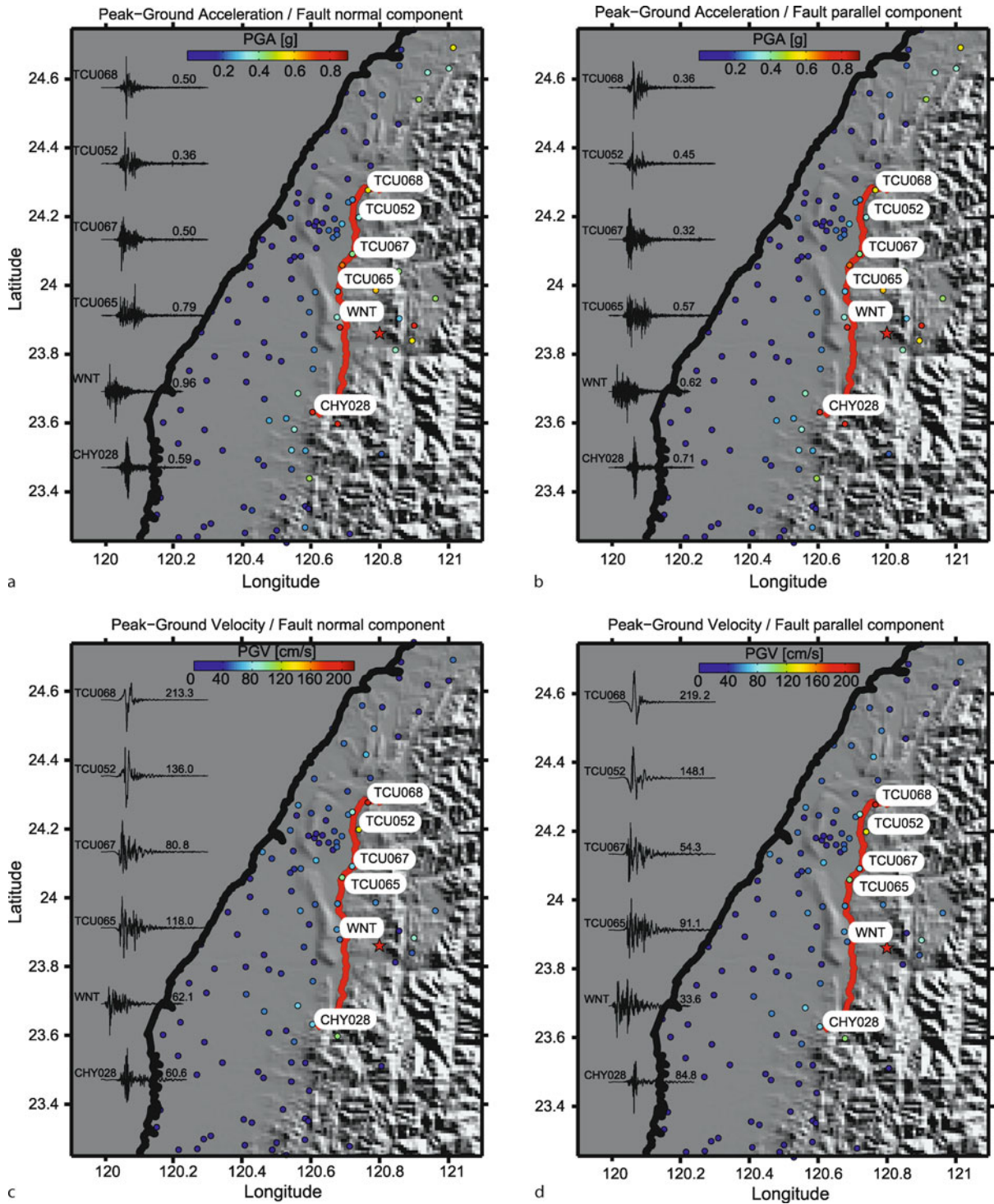
Near-field ground accelerations for the fault-normal (*left*) and fault-parallel (*right*) motion during the 1999 Chi-Chi earthquake; *waveforms* are plotted for selected stations in the vicinity of the epicenter (*red star*). Sites located on the footwall, i. e. west of the fault-trace, generally show lower ground-motions than sites located on the hanging-wall (east of the fault trace) despite similar epicentral distance

tion. A fault (shown as a planar surface with color-coded slip distribution) is embedded in a rock volume, and has a specific orientation in space described by the strike-angle  $\Phi$  (representing the azimuth of the fault's projection onto the surface, measured clockwise from North) and the dip-angle  $\delta$  (measured downward from the surface to the fault in the vertical plane perpendicular to the strike). The strike direction is defined such that, using the right-hand rule, the dip-angle is smaller than  $90^\circ$  (for a vertically dipping fault,  $\delta = 90^\circ$ , the strike direction is arbitrarily either direction).

Given the overall source geometry, the slip-vector on the fault plane defines the relative motion between the two blocks. The angle of slip, or rake angle  $\lambda$ , measured in the fault plane from the strike direction, shows the movement of the hanging wall relative to the foot wall (see inset in Fig. 5). The following definitions apply:  $\lambda = 0^\circ$  – *left-lateral strike-slip*, i. e. the hanging wall (or near-side of a vertical fault) moves horizontally to the right, so the opposite side moves to the left;  $\lambda = 180^\circ$  – *right-lateral strike-slip*, i. e. the hanging wall (or near-side of a vertical fault) moves horizontally to the left, so the opposite side moves to the

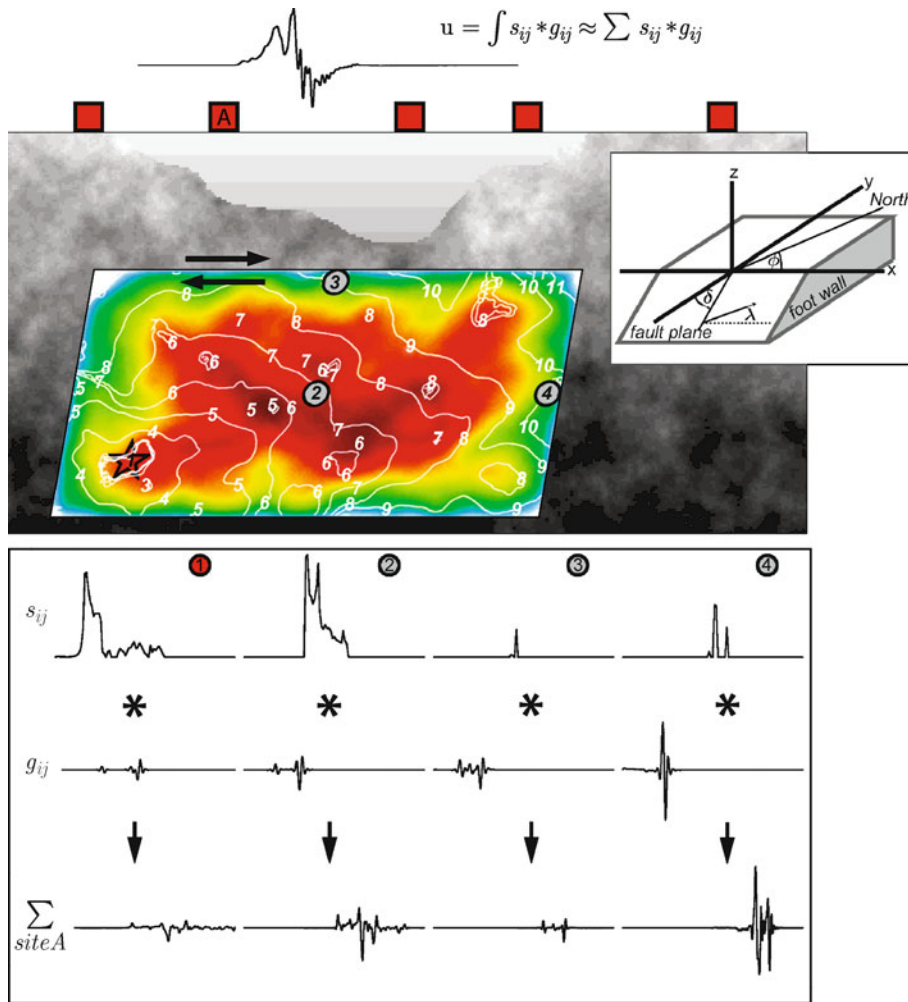
right. For  $\lambda = 90^\circ$ , the hanging wall moves upward (*thrust faulting*), for  $\lambda = 270^\circ$  the hanging wall moves downward (*normal faulting*). Figure 5 also displays the rupture model for a right-lateral strike-slip earthquake with  $\lambda = 180^\circ$  (indicated by the black arrows) on a fault-plane that dips  $80^\circ$ . The strike in this hypothetical case is undefined. The amount of displacement (slip) on each point of the fault is color-coded, white contours (at  $\Delta t = 1$  s spacing) show the expanding rupture front. Examples of past earthquakes show that the rake angle may also vary over the fault plane (Fig. 6c, d). Geometrical complexity, manifested in several fault segments, is often present (Fig. 9); in these cases the faulting-style is characterized by the predominant fault direction and slip angle.

The strength of seismic radiation is typically quantified by an earthquake magnitude, whereby a variety of magnitude scales exist. The seismic moment, defined as  $M_0 = \mu \cdot A \cdot D$ , ( $\mu$ : shear modulus in the source region;  $A = L \cdot W$ : fault area, given by fault-length  $L$  and fault-width  $W$ ;  $D$ : average displacement on the fault) is considered the best scalar quantity characterizing earthquake size. Ben-Zion [20,23] proposes to use the more basic concept of



Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures, Figure 4

Near-field ground accelerations (top) for the fault-normal **a** and fault-parallel **b** motions during the 1999 Chi-Chi earthquake; waveforms are plotted for selected stations along the fault trace (red line). **c** and **d** show the corresponding ground-velocities. Sites located at either end of the fault trace recorded waveforms with shorter duration of the dominant wave-energy. Velocity records in the North show strong rupture-directivity effects as high-amplitude short-duration velocity pulses

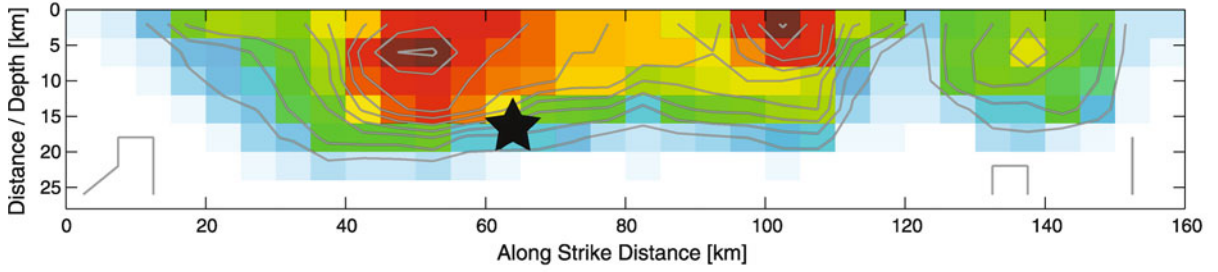


Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures, Figure 5

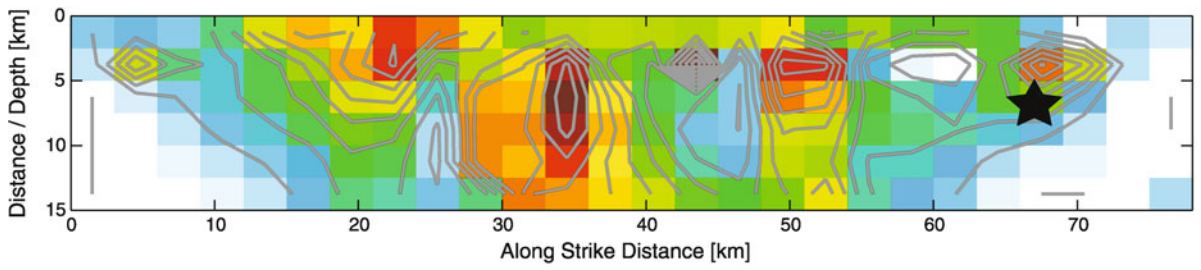
Sketch of the physical components of near-source ground motions (*top-right inset*: illustrates the source-geometry definition). A simulated earthquake rupture occurs on an assumed planar fault, with color-coded slip amplitudes (0–2.5 m) and white contours of the propagating rupture front (emanating from the hypocenter shown by the *red star*). The rupture is embedded in a gray-scale model of the Earth's crust. Darker gray-tones denote higher  $v_p$ ,  $v_s$ ,  $\rho$  than lighter gray, reflecting the overall trend of higher velocities at greater depth, locally disturbed by random heterogeneities. A basin-structure with less compliant layered sedimentary rocks exhibits complex sub-surface topography. A near-field seismogram (for an arbitrary horizontal component) is shown at a representative site A. Ground-motions are composed as the summation of the slip-functions  $s_{ij}$  at grid-points  $i, j$  on the fault, convolved with the corresponding Green's functions  $g_{ij}$  for this observer and grid points (the site term  $l_k$  in Eq. (1) is neglected for simplicity)

seismic potency,  $P_0 = A \cdot D$ , to avoid a source quantification that includes ambiguously defined or poorly known material properties; however, seismic potency is rarely used in the earthquake engineering. The moment magnitude follows from  $M_w = 2/3 \cdot \log M_0 - 6.07$  (with  $M_0$  given in Nm). The static stress drop  $\Delta\sigma$  is related to the ratio of average slip to a characteristic fault length, therefore seismic moment is proportional to static stress drop:  $M_0 \propto \Delta\sigma \cdot A^{3/2}$ . Observational evidences [106,186,187] confirm

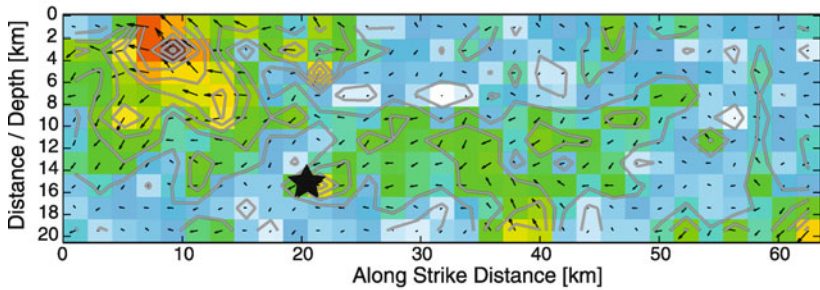
this scaling for moderate to large earthquakes, suggesting – constant (i. e. scale-invariant) static stress drop and self-similar earthquake source scaling: large earthquakes are only scaled-up versions of small earthquakes. Other scaling relations between observable fault parameters (fault area, length, and width; average slip) and magnitude (seismic moment) [79,122,169,187] partially confirm the self-similarity of earthquake rupture, but also provide evidence for the break-down of self-similarity for very small earth-



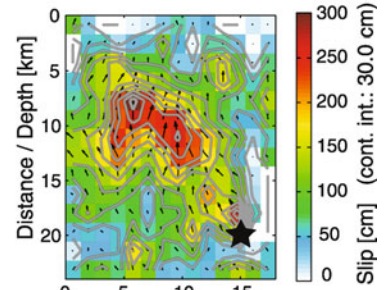
a



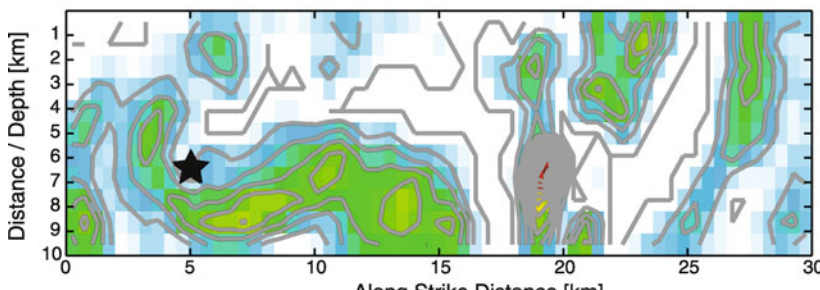
b



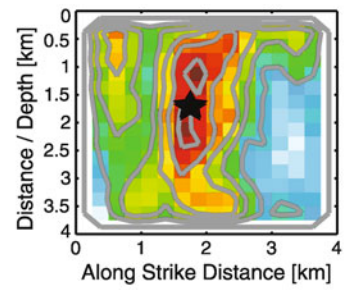
c



d



e



f

◀ Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures, Figure 6

Selection of finite-source rupture models, obtained from inversion of seismic recordings and geodetic data, for six earthquakes in the magnitude range  $4.5 \leq M_w \leq 7.5$ , illustrating the heterogeneous distribution of earthquake slip (color-coded using different color scale for each model). The *black star* denotes the hypocenter, generally located in the vicinity of a high-slip patch (also called “asperity”). a 1999 Izmit earthquake ( $M7.5$ ) [51]; b 1992 Landers earthquake ( $M7.3$ ) [184]; c 1995 Kobe event ( $M7.0$ ) [165]; d 1994 Northridge earthquake ( $M6.7$ ) [185]; e 1984 Morgan Hill event ( $M6.1$ ) [29]; f Hida Mountains event #5 ( $M4.5$ ) [96]. The rupture length of b is about half that of a; models b, c and d are drawn to scale with respect to each other. Model e has fault length  $L = 30$  km, a factor of two shorter than c, while f has source dimensions  $4 \times 4$  km<sup>2</sup>

quakes [23] and very long strike-slip earthquakes [154]. The exact physics leading to these source-scaling properties is a current topic of active research [90,91].

The overall faulting-style, defined by the dip and rake angle, affects the radiation pattern of P- and S-waves (see Chap. 4 in [6] for more details), and thus the resulting ground-motion characteristics. Together with the earthquake magnitude, the style-of-faulting is used as a predictor variable in empirical ground-motion prediction equations (see Sect. “Future Directions”). However, ground-motion complexity arises largely from the details of the rupture process (Fig. 5). The earthquake nucleates at the hypocenter (indicated by the red star), and propagates over the fault plane (white contour lines) with a rupture velocity that may exhibit local variations due to initial stress conditions and frictional properties on the fault. Stress state and friction also determine the slip-time history  $s_{ij}$  for all points  $ij$  on the fault (which start sliding once reached by the propagating rupture front). To first order, the shape of the local slip-velocity function depends on the overall fault dimensions (i. e. fault aspect ratio), the relative position of each point with respect to the rupture nucleation point and its distance to the closest fault edge [52,58,59,78] (variations in  $s_{ij}$  at four different locations on the fault plane of lower panel). The duration of slip (rise time) is related to the length of these local slip-functions. Once sliding has stopped at all points on the fault, the rupture has attained its final slip distribution (color-coded in Fig. 5) within a characteristic time (called rupture duration) which depends on fault dimensions and rupture velocity (and marginally on rise time). As depicted in Fig. 5, the slip distribution, the rupture velocity and the rise times are highly heterogeneous over the fault plane, illustrating earthquake source complexity in terms of kinematic source parameters.

The governing equation that relates ground displacement to the motion on the fault is given by a representation theorem (Chap. 3 in [6]), which we use in the following notation:

$$\mathbf{u}(\mathbf{x}, t) = \int d\tau \int_S \Delta u(\xi, \tau) c_{ijkl} v_j G_{nk,l}(\xi, \tau; \mathbf{x}, t) dS. \quad (2)$$

Equation (2) states that the time-dependent ground displacement at observer location  $\mathbf{x}$  and time  $t$  depends on the space-time integral over the space-time-dependent slip function  $\Delta u(\xi, \tau)$  on the fault plane ( $\xi$  defines the position on the fault,  $\tau$  is time), the elasticity tensor  $c_{ijkl}$ , the fault-normal vector  $v_j$ , and the Green’s tensor  $G_{nk,l}$  (subscripted comma indicate the derivative with respect to the subsequent variable). Equation (2) is related to Eq. (1), but contains the functional dependency on the source explicitly; in this context the Green’s function typically neglects the site effects  $l_k(t)$ . The term  $\int_S \Delta u(\xi, \tau) dS$  contains the time-dependent local particle motion on the fault,  $p_{ij}(t)$ , and the effects of fault finiteness,  $f(t)$ .

Equation (2) can be applied to retrieve the spatio-temporal slip distribution on the fault plane from seismic recordings (and other data) by either forward-modeling or a formal inversion procedure. Earthquake-source inversions have been carried out since the early 1980ies, manifesting the complexity of the rupture process as depicted in Figs. 6 and 9 (see also database of finite-source rupture models [121]). These earthquake source images represent kinematic rupture models that quantify  $\Delta u(\xi, \tau)$  based on observations, but do not explicitly derive  $\Delta u(\xi, \tau)$  from physical principles as attempted in dynamic rupture models e. g. [14,15,28,78].

For examining earthquake source complexity in more detail, it is mandatory we need to distinguish kinematic from dynamic rupture models, and we need to quantify slip heterogeneity (and its associated stress-change distribution) on the fault plane. It is also important to note that the position of the rupture nucleation point is not arbitrary on the fault plane for a given slip (stress) distribution but adheres to fundamental concepts of energy balance during the rupture process. This Section characterizes source complexity in space and time before introducing to *isochrone theory*, a powerful tool to visualize how the details of the rupture process determines near-fault ground-motions.

### Kinematic Rupture Models

Kinematic rupture models characterize the space-time evolution of earthquake rupture in terms of a time-dependent displacement field (distribution of slip vectors)



on a predefined fault plane without considering the forces and stresses that cause these motions on the fault. The local slip-rate (or slip-velocity) function is specified along with the rupture propagation properties. Ground motions can then be computed using Eq. (2) with any kinematic rupture model  $\Delta u(\xi, \tau)$ . Current ground-motion simulation approaches are largely based on kinematic source models since they can be efficiently generated; recent advancements try to capture at least the basic principles of source dynamics by proposing *pseudo-dynamic* source models [77,78,120]. Earthquake source inversions parameterized in terms of Eq. (2) retrieve only a kinematic rupture model which, in principle, does not need to obey any physical laws. The database of finite-source rupture models [121] provides a compilation of kinematic source models for past earthquakes, obtained by applying Eq. (2) to a variety of data and using different methods to solve the inverse problem.

### Dynamic Rupture Models

Dynamic rupture models build a physical understanding of the earthquake rupture process based on the material properties around the source volume, and the initial and boundary conditions for the forces/stresses acting on the fault plane [80]. The distribution of slip vectors  $\Delta u(\xi, \tau)$  on the fault plane is obtained by solving the elasto-dynamic equations of motion under the assumption of a constitutive law (i. e. a friction model), considering essentially the energy balance at the crack tip during rupture growth (for details see Chap. 11 in [6]). Dynamic rupture models have been developed for (i) canonical model to study general feature of dynamic rupture e. g. [10,58,59] (ii) for existing kinematic source models to infer their specific dynamic rupture process e. g. [14,15,126,181] (iii) for heterogeneous initial conditions in stress and/or friction and material distribution to investigate rupture details for classes of events e. g. [11,13,63,68,78,83,133,151]. Due to the high computational demands, dynamic rupture models are not (yet) developed routinely for ground-motion simulations, but rather to investigate source physics for general cases, earthquakes of special interest, and to study rupture behavior for certain classes of initial conditions.

### Quantifying Slip Heterogeneity

Recent approaches to characterize and quantify slip heterogeneity demonstrate that slip distributions exhibit statistical properties and empirical laws grounded in physical principles. First-order observations for these slip distributions (obtained in finite-source inversions based on Eq. (2)), can be made in Fig. 6, a compilation of slip models

for six earthquakes in the magnitude range  $4.5 \leq M_w \leq 7.5$ . The source dimensions increase from  $4 \times 4 \text{ km}^2$  to  $25 \times 160 \text{ km}^2$  while the corresponding maximum fault displacements grow by two orders of magnitude (from  $\sim 5 \text{ cm}$  to over  $500 \text{ cm}$ ). Mai and Beroza [122] have shown that the overall source-scaling relations of such rupture models is roughly consistent with global earthquake scaling laws (i. e.  $M_o \propto A^{3/2}$ ), but that there is evidence that the commonly assumed self-similar constant stress-drop scaling may not hold, because slip on the fault does not saturate but keeps increasing for growing fault dimensions (albeit at a progressively lower rate). The topic of general earthquake scaling laws is still hotly debated e. g. (see ► [Earthquake Scaling Laws](#) [90,91], and directly affects ground-motion prediction [188,189] (see Sect. “[Future Directions](#)”).

More fundamentally, earthquake slip is heterogeneous on the rupture plane (Fig. 6), i. e. regions of little displacement are separated from areas of high slip (often called “asperities”). The distribution and properties of these high-slip patches strongly influences seismic radiation and hence near-fault ground motions. Characterizing and quantifying slip heterogeneity is thus important for accurately predicting and simulating ground-motions for future earthquakes, but also to better understand the physics of earthquake rupture. Two basic avenues have been pursued in the recent past to quantify slip complexity: (i) a deterministic approach that counts the number of high-slip patches and extracts their properties in terms of size, displacement and other quantities [127,169]; (ii) a stochastic approach that characterizes slip heterogeneity in terms of a random-field model [113,114,123]. Results of both methods can be used for simulating stochastic slip distributions for ground-motion calculation to investigate how source complexity affects near-field ground shaking.

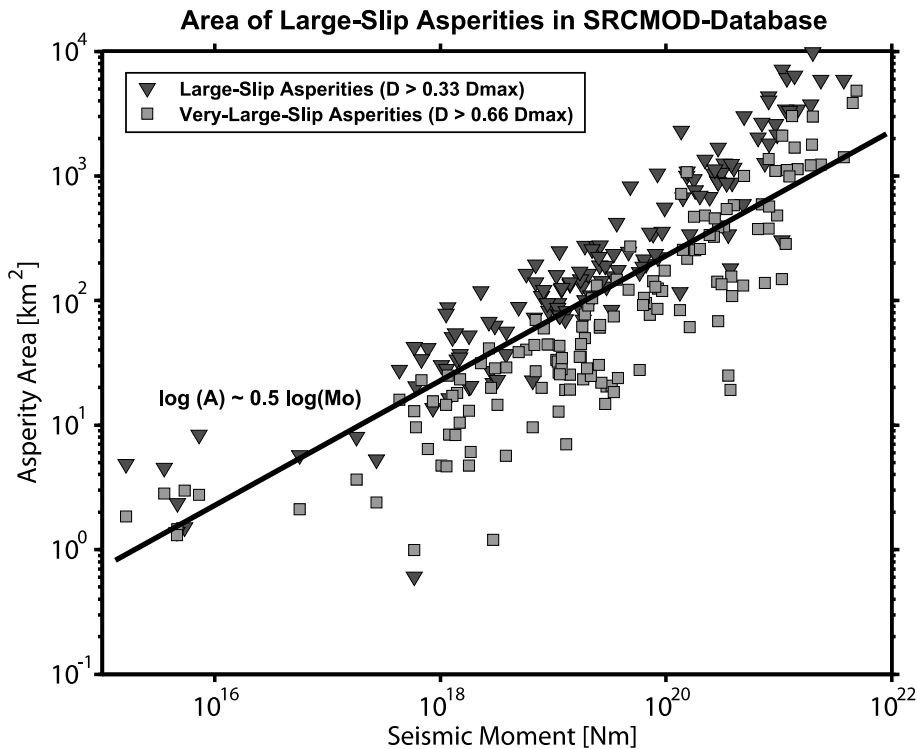
Before interpreting and utilizing inferred source-rupture models for future science or engineering applications, the reliability and resolution of these source-inversion images needs to be addressed. Examining the intra-event variability (source models for the same earthquake obtained by different research groups) often reveals large differences in slip distributions and temporal rupture parameters for the set of source models (online-database [121] for examples and [25] for a critical review). The degree of (dis-)similarity between these slip histories is controlled by differences in the Green’s functions (due to computational methods and the choice of the Earth model), variations in the fault parameterization, different inversion schemes and their control parameters (tuning parameters in non-linear inversion; damping/smoothing constraints in linearized multi-time-window inversion). Source-rupture

ture models are also affected by the type of data (seismic, geodetic, or both; additional geologic data), their selection and processing, and their weighting in the inversion.

Consequently, variations in inverted source models are expected, but the actual uncertainties are rarely quantified. Only recently, more efforts are devoted to perform rigorous uncertainty estimation by testing different inversion algorithms, model parameterizations, and data selection criteria for the same earthquake [56,86,116,117]. The work by Monelli and Mai [128] even estimates posteriori probability density functions for the model parameters of interest, using a non-linear inversion strategy coupled to a Bayesian inference technique. Despite the variability in imaged slip distributions for a given earthquake, several source quantities are stably estimated: rupture dimension, seismic moment and average displacement (both within a factor of 2 generally), and also the slip near the hypocenter. Estimates for the average rupture velocity may vary strongly (up to  $\sim 30\%$ ) between models for the same event because rupture speed estimates trade off with the slip-rate function. The locations of high-slip patches on the fault (“asperities”) are relatively well located if sufficient data

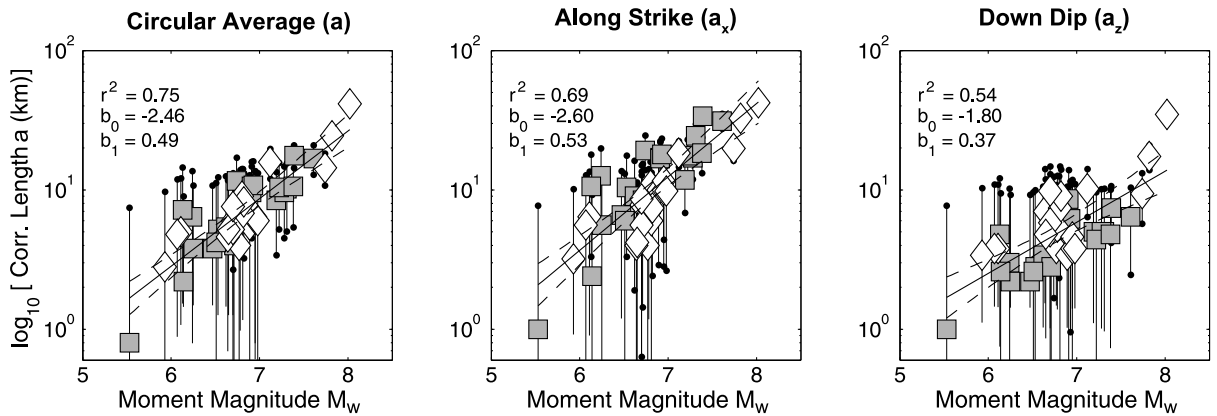
are used, but exceptions exist (i. e. the 1999 Izmit earthquake). The intra-event variability of maximum fault slip is generally quite high, and depends on the particular inversion strategy. However, estimates of correlation lengths or dominant wave-lengths of heterogeneous slip maps are robust for the different models of a given earthquake [123], indicating that the statistical properties of earthquake rupture are well imaged. Also the hypocenter location with respect to the regions of high-slip is estimated reliably in finite-source inversions [127].

Measuring slip heterogeneity deterministically by defining large-slip ( $\frac{1}{3} \cdot D_{\max} \leq D < \frac{2}{3} \cdot D_{\max}$ ) and very-large-slip ( $D \geq \frac{2}{3} \cdot D_{\max}$ ) asperities ( $D = D(x, z)$  is the local slip on the rupture plane,  $D_{\max}$  the corresponding maximum slip), [127] examine the scaling of total asperity area ( $A_{\text{ta}}$ ) with respect to seismic moment for 90 finite-source rupture models (Fig. 7). Without a formal regression, the measurements suggest the scaling  $\log_{10}(A_{\text{ta}}) \propto 0.5 \cdot \log_{10}(M_o)$ , meaning that the area occupied by high-slip patches grows slower with increasing magnitude than the total fault size (which scales as  $\log_{10}(A) \propto \frac{2}{3} \cdot \log_{10}(M_o)$ ). This scaling requires relatively larger maximum displace-



Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures, Figure 7

Asperity area, measured for  $\sim 90$  rupture models (<http://www.seismo.ethz.ch/srcmod>), plotted against seismic moment. A scaling of  $\log_{10}(A) \propto 0.5 \cdot \log_{10}(M_o)$  is shown for reference. Asperity size is measured based on slip values being a certain fraction of the maximum slip on the fault [127]:  $1/3 \cdot D_{\max} \leq D < 2/3 \cdot D_{\max}$  for “large-slip asperities”,  $D \geq 2/3 \cdot D_{\max}$  for “very-large-slip” asperities



Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures, Figure 8

Correlation lengths,  $a$ ,  $a_x$ ,  $a_z$ , versus moment magnitude  $M_w$  for 44 slip models (modified after [123]). Filled squares denote strike-slip earthquakes, open diamonds represent dip-slip events, vertical lines and black dots mark the  $1\sigma$ -error-estimates. The slope of  $b_1 \approx 0.5$  of the regression curves (solid lines) indicates that the correlation lengths scale self-similarly with moment magnitude

ments on the high-slip patches for large and very large earthquakes, which in turn affects seismic radiation and the scaling properties of near-fault ground motions. Large fault slip, concentrated on relatively small portions of the rupture plane, lead to large stress heterogeneity (with high static stress drop locally), thus implying large ground-motion variability for moderate to large earthquakes.

Alternatively, slip distributions (Fig. 6) can be treated as spatial random fields [113,123] to estimate the fractal dimension or the correlation lengths for an auto-correlation function (ACF). Such measurements are typically based on the two-dimensional power spectrum  $P(\mathbf{k})$  of the slip map. Assuming a simple fractal model for a random field, its power spectral density is given as

$$P(\mathbf{k}) \propto \mathbf{k}^{-\nu-1} \quad (3)$$

with wavenumber vector  $\mathbf{k}$  and scaling exponent  $\nu$ . [113] estimate one-dimensional scaling exponents for a small number of rupture models, using only horizontal slices of slip distributions, and find  $0.8 \leq \nu \leq 1.5$ , while [123] find for the two-dimensional exponents ( $\nu + 1$ ) values of about 1.7, implying a fractal dimension  $D = 2.3$ . The differences in these estimated scaling exponents can be reconciled when accounting for slightly different initial processing of the slip distributions and fitting strategies [114]. Testing a variety of auto-correlation functions to fit the spectral properties of a large number of slip distributions, [123] observe that a von Karman auto-correlation function, with its power-spectral density given by

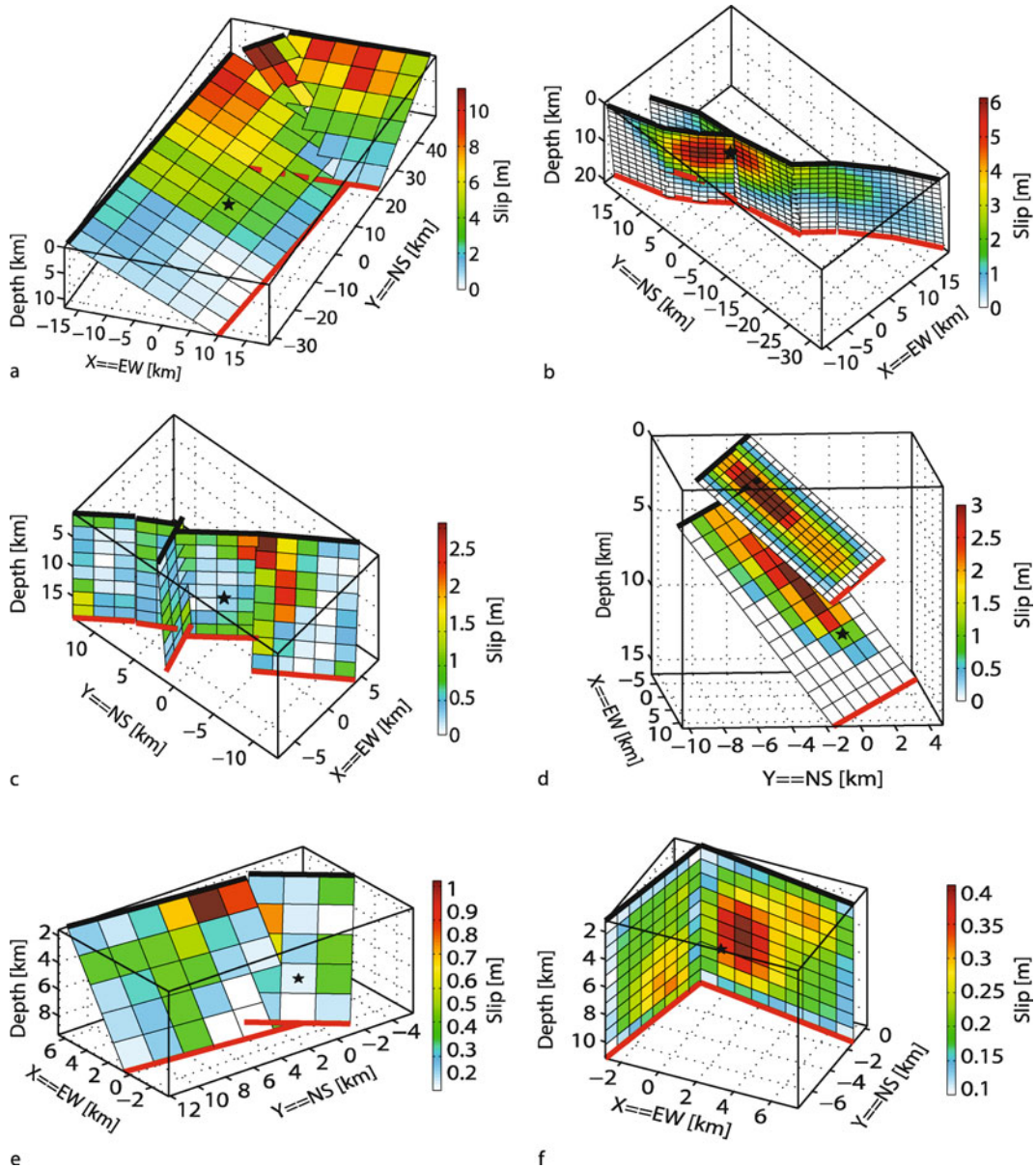
$$P(\mathbf{k}) = \frac{4\pi H}{K_H(0)} \cdot \frac{a_x \cdot a_z}{(1 + \mathbf{k}^2)^{H+1}}, \quad (4)$$

with magnitude-dependent correlation lengths  $a_x$ ,  $a_z$ , and a scale-invariant Hurst number  $H \approx 0.7$  best matches the slip heterogeneity spectra ( $K_H$ : modified Bessel function order  $H$ ;  $\mathbf{k}$  wavenumber). Figure 8 displays their estimated correlation lengths for 44 slip models [123], along with a least-squares regression that exhibits scaling relations of the form

$$\begin{aligned} \log_{10}(a_x) &\propto \frac{1}{2} M_w \\ \log_{10}(a_z) &\propto \frac{1}{3} M_w. \end{aligned} \quad (5)$$

Interestingly, the scaling law in Eq. (5) for correlation lengths of heterogeneous slip maps is similar to the relationship inferred for the deterministic measure of asperity size. This corroborates the previous argument that correlation lengths (asperity size) increases with increasing magnitude, though at a lower rate than the overall fault dimensions. Thus, to accommodate the corresponding seismic moment, the displacements (and associated stress drops) on these high-slip patches need to grow faster than a self-similar scaling would predict. This conclusion agrees with the conjecture of Heaton [88] that large earthquakes require a “very strong kick” (due to an area of large stress drop) in order to be able to grow into a very large rupture.

Treating earthquake slip as being distributed on a planar fault surface strongly simplifies geologic observations of geometrical fault complexity, where the degree of complexity depends on fault-zone maturity [20,90]. Earthquakes that break several fault segments can still be imaged with Eq. (2) using an appropriate parameterization, but planar-fault ruptures exhibit very different rupture dy-



Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures, Figure 9

Three-dimensional views of heterogeneous slip on geometrically complex faults, imaged using seismic and/or geodetic data. The *black star* shows the hypocenter, *thick black lines* are the top of the fault segments, *thick red lines* the bottom. a The 1999 Chi-Chi earthquake ( $M_w$  7.6) [101]; b the 1999 Hector Mine earthquake ( $M_w$  7.2) [102]; c the 2000 Tottori earthquake ( $M_w$  6.6) [100]; d the 1971 San Fernando event ( $M_w$  6.6) [87]; e the 2003 Miyagi-hokubo event ( $M_w$  6.1) [89]; f the 1997 Kagoshima earthquake ( $M_w$  6.0) [94]

namics than earthquakes that have to overcome geometrical obstacles [82,84,131,134]. Figure 9 shows a collection of such geometrically complex fault models, involving two or more segments that form either a system of sub-parallel fault planes (Fig. 9d), or a branching fault (Fig. 9b), or fault planes oriented at arbitrary angles to

each other (Fig. 9a, c, e, f). The seismic radiation from earthquakes on geometrically complex faults is more complicated than from a single-plane rupture, a topic that is investigated in the earthquake source dynamics community e.g. [14,15,43,81,82,84,131,134]. Unfortunately, little attention has been devoted in the past to rigorously exam-

ine and quantify the degree of ground-motion complexity generated by such fault systems. Thus, future research, using advanced numerical techniques on high-performance computing architecture, needs to properly represent fault systems at sufficient spatial resolution to capture the intricacies of rupture dynamics and associated seismic radiation on geometrically complex faults.

### Rupture Nucleation and Directivity

Besides the slip heterogeneity and the temporal rupture evolution (discussed later), the location of the hypocenter (point of rupture nucleation) is a critical factor affecting near-source ground motions. Some metric of hypocentral distance enters empirical attenuation relations (Sect. “Ground-Motion Scaling Relations”), while the on-fault hypocenter location determines the directivity effect [170,171]. This global directivity effect due to fault-hypocenter-site geometry results in large velocity-pulses, in particular at sites close to the fault, because of the constructive interference of S-wave energy which is continuously radiated from the propagating crack front and arrives within a short time window in the forward-direction of rupture propagation (see velocity records in Fig. 1 at station SlackC, FZ1, SC1E for the Parkfield event; Fig. 4 at station TCU052, TCU068 for the Chi-Chi event). In contrast, for sites in the backward direction of rupture propagation the arriving seismic energy is spread over a longer time interval, generating lower-frequency motions with smaller amplitudes. The directivity effect is therefore not a purely source-related phenomenon, but also depends on the observer location.

Additionally, there is an “on-fault” directivity effect which operates on smaller scales and is most prominent for large earthquakes and very near-fault sites. The on-fault directivity effect can be efficiently quantified using isochrone theory (see Sect. “Isochrone Theory”). For an explanatory description, consider the slip distribution and hypocenter location in Fig. 6a, for which the large-scale “global” directivity effect was observed towards the right (i. e. East in this case of the 1999 Izmit earthquake). However, any site located at about  $30 \leq X \leq 50$  km in along-strike distance (above the left-most high-slip region) experiences strong “local” directivity effects as the rupture propagates from the hypocenter towards the site and across the high-slip patch. In this case, the integrated slip along this rupture trajectory generates a high-amplitude short-duration pulse. Considering a more westerly hypocenter (e. g. at  $X = 30$  km instead of  $X = 65$  km) the large-scale directivity effect remains essentially unchanged while the integrated slip along the rupture trajectory for

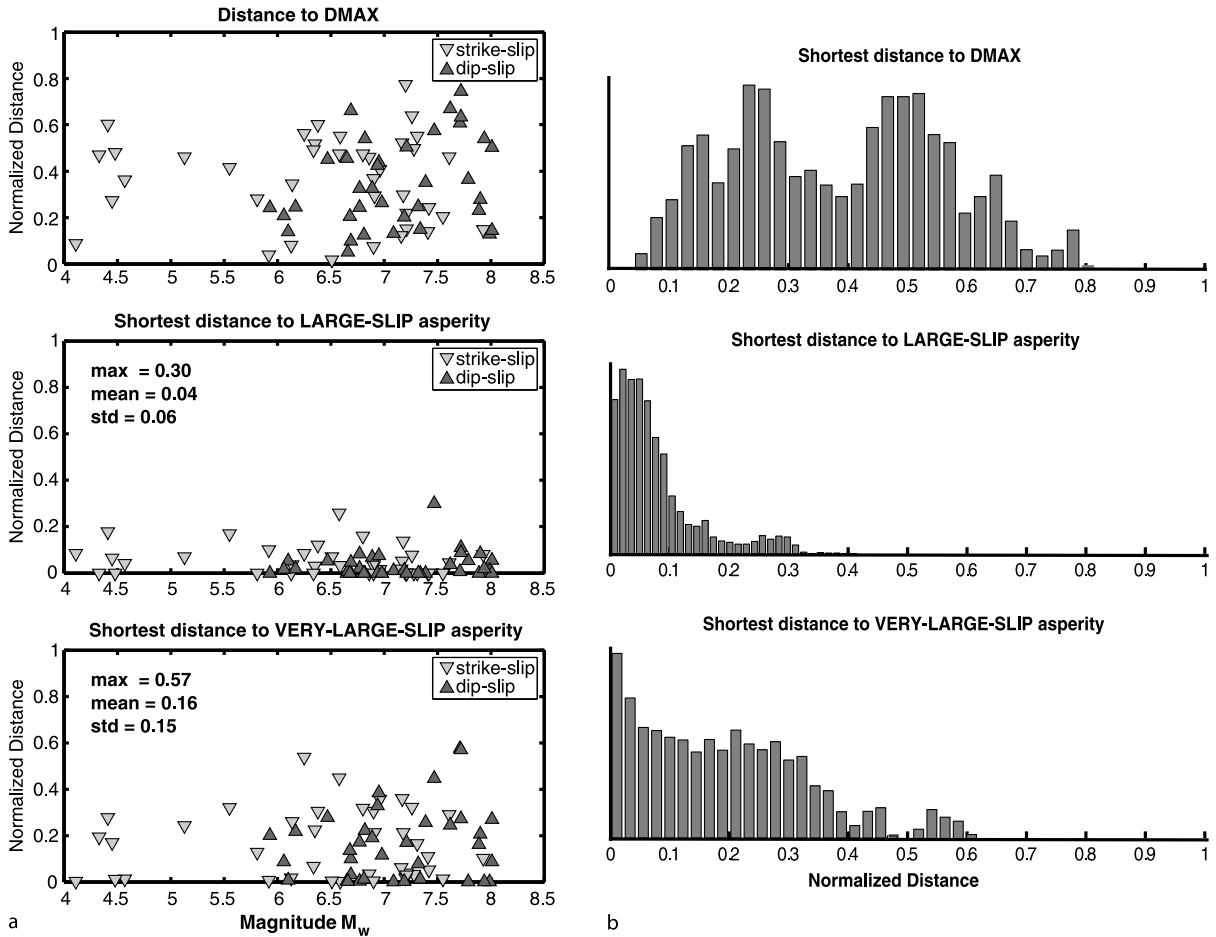
stations at  $30 \leq X \leq 50$  km is much smaller, thus diminishing on-fault directivity. The spatial relation between the hypocenter and large-slip regions plays an important role for ground-motion complexity, but is also fundamentally related to the dynamics of the rupture process.

Examining the hypocenter positions in Fig. 6 and Fig. 9, two observations are evident: (a) ruptures generally do not nucleate at the fault boundaries, but in the interior of the fault plane, but rarely exactly in the center; (b) the nucleation point is generally not located in regions of low slip (light colors,  $D \leq 0.1 \cdot D_{\max}$ ) but in areas where  $0.2 \cdot D_{\max} \leq D \leq 0.4 \cdot D_{\max}$  and the distance to a nearby large-slip zone (asperity) is small compared to the overall source dimensions [127]. This study investigated these first-order observations statistically using a database of  $\sim 80$  finite-source rupture models, and concluded that hypocenters are not randomly located on a fault but are located either within or close to regions of large slip (Fig. 10). More specifically, ruptures nucleate within or very close to large-slip asperities ( $\frac{1}{3} \cdot D_{\max} \leq D < \frac{2}{3} \cdot D_{\max}$ ), but rarely start on very-large-slip asperities ( $D \geq \frac{2}{3} \cdot D_{\max}$ ; often located far away from the hypocenter).

These observational constraints of slip heterogeneity and rupture nucleation are rooted in the energy balance of earthquake source physics, and have been confirmed by dynamic rupture simulations [78,133,151]. The essence of the physical mechanism is that a sustained large earthquake can only be generated if a sufficiently large amount of energy is furnished to the propagating crack tip to facilitate rupture growth. In a simplistic view, an earthquake can only grow in size if the energy absorbed to create new crack surface (fracture energy) balances the available elasto-static energy and the energy radiated by seismic waves (Chap. 11 in [6]). If the hypocenter is located in regions of low slip (low stress drop) and far away from any point of significant stress drop, the required fracture energy will become too large to allow further crack propagation, and the rupture will stop prematurely. The spatial correlation between hypocenter and asperity is thus consistent with this simplified energy budget of dynamic rupture. Moreover this hypocenter-asperity relation manifests the “on-fault” directivity and strongly effects seismic radiation and thus ground-motion complexity.

### Temporal Rupture Evolution

The characteristics of the slip distribution and the hypocenter position alone are insufficient to parameterize  $\Delta u(\xi, \tau)$  in Eq. (2) because the temporal rupture evolution is not yet specified. Rupture velocity and the local slip-velocity function with its associated rise time (slip duration)



Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures, Figure 10

Hypocenter distance to “large-slip” and “very-large-slip” regions in finite-source rupture models. **a** Raw measurements of shortest distances between hypocenter and point of maximum displacement (*top*), hypocenter and closest large-slip asperity (*middle*), and hypocenter and closest very-large-slip asperity (*bottom*), separated into strike-slip and dip-slip earthquakes. **b** Distributions of shortest distances between the hypocenter and the point of peak-slip, the closest large-slip, and the closest very-large-slip asperity, generated using Monte Carlo simulations that include hypocenter uncertainties. The distributions indicate that 16% of all hypocenters occur on a very-large-slip asperity, 35% within a large-slip asperity, and 48% right outside an asperity (compiled from [127])

need to be given, either pre-defined by means of a kinematic rupture modeling approach or solved for from physical principles in a dynamic rupture model.

**Rupture Velocity** After earthquake nucleation the rupture propagates over the fault plane with a rupture velocity  $v_r$  which generally is close to the local shear-wave velocity (typically in the range  $0.6 \cdot v_s \leq v_r \leq 0.9 \cdot v_s$ ). Seismic radiation and thus near-field motions strongly depend on rupture velocity: slow earthquakes radiate little seismic energy while fast ruptures generate higher ground-motion amplitudes. Earthquake source inversions mostly assume a constant rupture velocity over the fault plane, al-

though dynamic rupture modeling and evidence from particularly well recorded earthquakes indicates that rupture speed may vary significantly on the fault plane [16,44]. Dynamic modeling shows that the initial rupture speed (during and right after the nucleation phase) may be small (e.g.  $v_r^{\text{init}} \leq 0.5 \cdot v_s$ ), but then rapidly increases during the rupture’s dynamic breakout and propagation phase (Chap. 11 in [6]). Depending on the stress conditions and the frictional parameters on the fault, and the geometrical properties of the rupturing fault, the rupture speed may actually exhibit very strong small-scale variations (see Fig. 5 for example). Variations in rupture velocity are a source of high-frequency seismic radiation [119,173],

and thus strongly contribute to near-field ground motion complexity.

While the rupture speed for most earthquakes is lower than the local S-wave velocity (i. e. sub-shear rupture propagation), there is evidence from laboratory measurements [155] and from seismic data and source modeling e. g. [16,44,134] that the crack front may travel at super-shear speeds at least over parts of the fault plane. Such super-shear ruptures generate rather distinct ground-motion characteristics at the few sites where they were recorded, but it is not yet clear what the “generic” ground-motion signature of super-sonic rupture speed would be. [1] address some aspects of this question by kinematic rupture modeling with various rupture speeds and computing the resulting near-field motions, indicating very peculiar “mode-switching” of the seismic-energy distribution on the two horizontal components as the rupture speed becomes very large. Furthermore dynamic rupture simulations [10,58,59,63,68,134] provide physical models for the occurrence of super-shear propagation, which depends on local stress and frictional conditions on the fault plane. The potential occurrence of super-shear rupture burst [62,63] and rupture speed variations in general are critical for understanding ground-motion complexity. To efficiently include such effects into ground-motion simulations, [78] developed an initial physics-based representation of rupture-velocity heterogeneity for kinematic source modeling.

**Slip-Velocity Function and Rise Time** As the propagating rupture sweeps over the fault plane points on the fault plane are “activated” and start slipping. Each point traces an individual slip-time history (slip-velocity function) whose shape and duration depends on the stress conditions and frictional properties on the fault, but also on overall fault size and the position of each point with respect to rupture nucleation [59]. Figure 5 illustrates the variability of local slip-velocity functions  $s_{ij}$  on the fault. The slip duration (rise time  $\tau_r$ ) is usually defined by integrating the slip-velocity function and measuring the time it takes to complete 5–95% of the total displacement at each point.

Several approximations of these complicated slip-velocity functions are in use: a boxcar or isosceles-triangle function with a rise time equal to the width of the function, or symmetric and asymmetric cosine-functions. The classical approximation uses the solution of a quasi-dynamic crack model [110], showing an  $1/\sqrt{t}$ -decay after a rapid onset. Recent modifications to this Kostrov-type slip function [129,180] are compatible with earthquake dynamics, while ground-motion simulations and source in-

version often assume simple parameterization of overlapping triangles [78,92].

As Fig. 5 indicates, rise time varies over the fault plane. However many kinematic source inversions and simple source-model simulations assume constant rise time. A self-similarly expanding crack model predicts longer rise times in the center of the fault e. g. [58,110] where the rupture nucleates, but short rise times at the crack periphery due to the earlier arrival of the healing front from the crack rim. Analysis of source-inversion results [88] and dynamic rupture modeling [28] however suggest that rise times are in general short, i. e. the rupture propagates as a self-healing pulse over the fault plane with rise times determined by local healing of the rupture front.

For fixed rupture velocity and slip distribution, ground-motions are very sensitive to rise time variations. Shorter rise times lead to larger ground-motions as the seismic energy is released in a shorter time interval. The detailed shape of the slip-velocity function is less important for the seismic waveforms, because they are determined by the summation of slip-functions, convolved with the appropriate Green’s function, over the entire fault plane (effectively filtering out small-scale features of the slip-velocity parameterizations). However, the peak slip-velocity value that the slip-rate function may attain is crucial for the final ground-motion amplitudes.

### Isochrone Theory

The effects of various kinematic source parameters (slip and slip-velocity distribution, rupture propagation) on near-field ground-motions can be efficiently visualized in the framework of the isochrone theory, a high-frequency (ray-theory) approximation to calculate seismic radiation from earthquake ruptures [27,173]. By approximating the elastic wave Green’s functions with far-field body waves, surface-waves and low-frequency near-field terms are not considered in the isochrone method. It is correct, however, at higher frequencies, where far-field terms dominate, and in the distance range to the fault where surface waves have not yet developed but near-field terms are unimportant. The main aspect of this approach is that the space-time integration in Eq. (2) can be replaced by a series of line integrals over the fault for simple slip-function parameterizations. For each time  $t_i$  in the observer seismogram a line integral is computed for which the integration path comprises only those points on the fault which radiate seismic waves that arrive at the observer location at exactly the time  $t_i$ . Defining an arrival-time function, composed of the rupture-front arrival time at each point on the fault and the corresponding seismic-wave travel time to the ob-

server, the integration path represents an isochrone of this arrival-time function. The corresponding isochrone velocity,  $\mathbf{c}$  (the spatial derivative of the arrival-time function) is related to rupture velocity  $v_r$ , and thus resembles the directivity function. Since ground velocity is proportional to isochrone velocity, the characteristics of high-frequency body-waves in near-field seismograms can easily be examined with respect to arbitrarily heterogeneous slip and rupture velocity distributions.

The omission of near-field terms and surface-waves restricts the isochrone method to cases in which seismic radiation from compact slip zones or sudden changes in rupture velocity dominate the ground-motions. Such source behavior has been reported for many earthquakes, consistent with [119] who showed that the high-frequency far-field radiation is emitted at the propagating crack tip, with additional radiation coming from stopping phases as the rupture heals. The isochrone theory has also been successfully applied to earthquake source inversions [29].

The detailed theoretical development of the isochrone theory is beyond the scope of this article, but a brief description will help to illustrate the concept. Spudich and Frazer [173] link the representation theorem, Eq. (2), to geometrical ray theory (without Fraunhofer approximation) for the far-field displacements of P- and S-waves. According to [176] a simple, yet versatile, parameterization of the slip function  $s(\mathbf{y}, t)$ , is given by

$$s(\mathbf{y}, t) = \mathbf{s}_r(\mathbf{y})f_r[t - t_r(\mathbf{y})], \quad (6)$$

with rupture time  $t_r$  at positions  $\mathbf{y}$  on the fault plane, position-independent shape-function  $f_r$  for the slip-velocity function, and position-dependent amplitude  $\mathbf{s}$ . Any heterogeneous rupture model can be approximated by Eq. (6). Inserting this expression into Eq. (2), shown here for S-waves only, one obtains

$$\mathbf{u}^S(\mathbf{x}, t) = f_r(t) \cdot \int_S \mathbf{s}_r \cdot G_a^S \delta(t - t_a^S) dS \quad (7)$$

where  $t_a^S(\mathbf{y}, \mathbf{x}) = t_r(\mathbf{y}) + t^S(\mathbf{y}, \mathbf{x})$  is the arrival time function for an observer at location  $\mathbf{x}$  due to an S-wave radiated at point  $\mathbf{y}$  on the fault  $G_a^S$  is the corresponding Green's function. The surface integral in Eq. (7) is non-zero only if the argument of the  $\delta$ -function is zero. Curves  $\mathbf{y}(t_a^S, \mathbf{x})$  define the contours of equal arrival time (isochrones) at observer  $\mathbf{x}$ ; the surface-integral in Eq. (7) thus reduces to a line integral

$$\mathbf{u}^S(\mathbf{x}, t) = f_r(t) \cdot \int_{\mathbf{y}(t, \mathbf{x})} (\mathbf{s}_r \cdot G_a^S) \cdot c(\mathbf{y}, \mathbf{x}) dl \quad (8)$$

in which  $c(\mathbf{y}, \mathbf{x}) = |\nabla t^S(\mathbf{y}, \mathbf{x})|^{-1}$  represents the ‘‘isochrone velocity’’ of these curves along the fault surface ( $\nabla$  denotes

the gradient operator). Note that the isochrone velocity is related to the directivity function; isochrone theory therefore helps to visualize on-fault directivity effects [172].

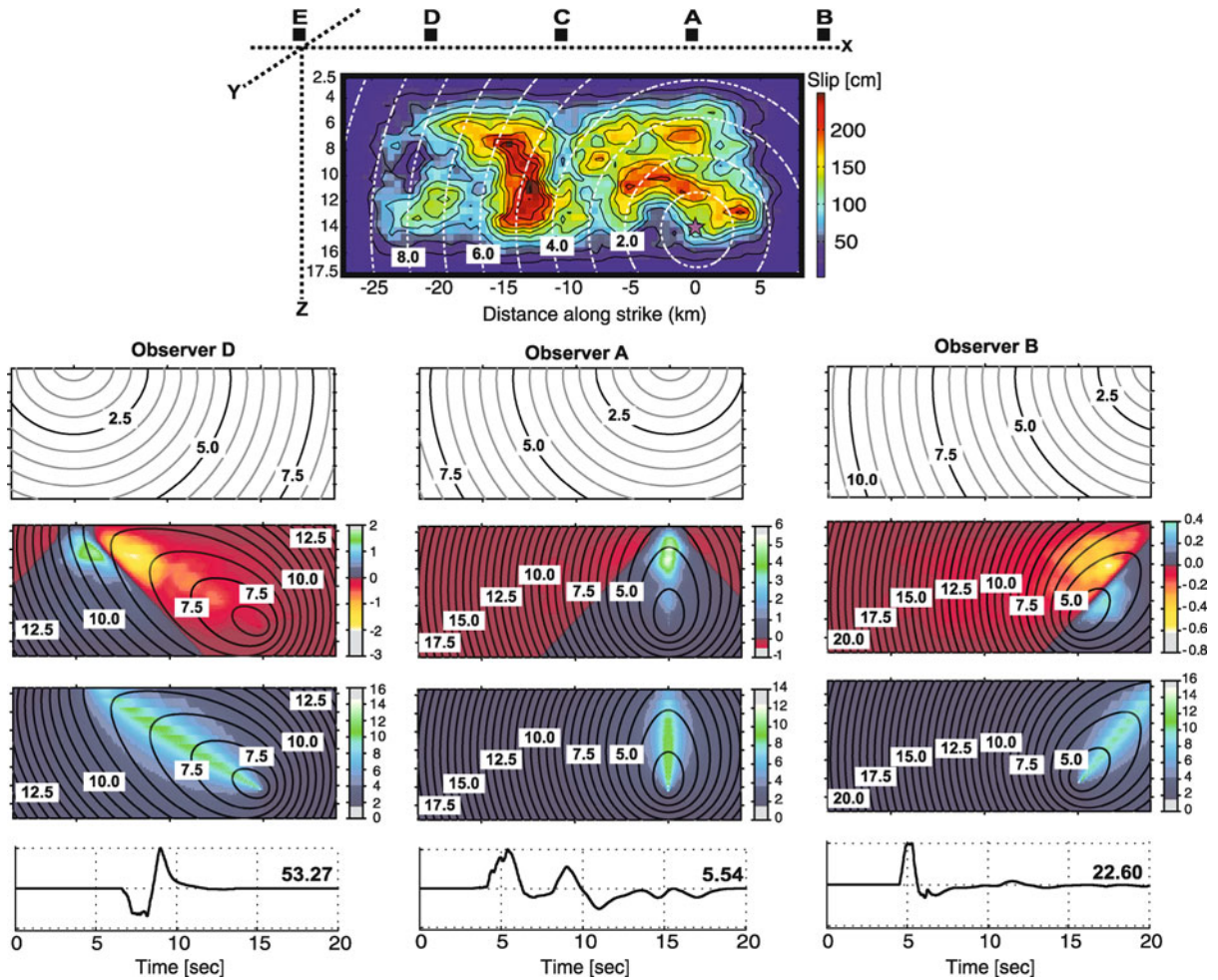
As an application of the isochrone method, we compute near-field seismograms for a hypothetical rupture ( $M_w$  6.7) with constant rupture velocity (radially spreading rupture front) at three sites, each at fault-perpendicular position  $y = 0.2$  km, but various positions along-strike (Fig. 11). A simple layered medium is assumed, and anelastic attenuation is excluded. For each observer, the S-wave travel time function (travel times from the site to each point on the fault) is shown (top panels), the color-coded integrand (Eq. (8)) with isochrone contours (spacing  $dt = 0.5$  sec) (2nd row of panels) and the isochrone velocity with isochrone contours (3rd row of panels). S-wave pulses in the fault-normal velocity seismograms (bottom) can then be attributed to isochrone properties. Consider for example observer D. The first isochrone appears at  $\sim 7$  sec, consistent with the S-wave arrival time in the seismogram. Isochrones up to 8 sec are entirely in areas where the integrand is negative (yellow to red colors), hence the down-ward initial motion. The peak velocity (53.2 cm/sec) occurs at  $t \approx 9$  sec when the corresponding isochrones hit the large-amplitude regions of the integrand (green areas). Observer D exhibits stronger directivity than observers A or B due to its position with respect to the hypocenter and the dominant asperity, generating large amplitudes due to high isochrone velocity over regions of large slip. At site D, the seismic energy radiated from the high-slip patch at  $x \approx -15$  km,  $y \approx 10$  km arrives within a short time window, whereas at site A, the contributions from this slip patch are spread over a longer time window. For site A, the isochrone velocity is smaller than for D, hence directivity is less pronounced. As a results, site A exhibits lower peak ground motions but a longer shaking duration.

The current implementation of the isochrone method [176] could be extended to include near-field terms [104] and to be applicable in three-dimensional velocity structures by adding 3D ray-tracing capabilities. Spatially variable slip-velocity functions and anelastic attenuation are further possible extensions. Finally, effects of wave-scattering in random media could be implemented using scattering operators [125,192,193,194] convolved with the local slip-rate function  $s(\mathbf{y}, t)$ .

### Wave Propagation in Complex Media: Path and Site Effects

Returning to Eq. (1), this Section discusses path effects,  $g_k(t)$ , due to source-to-site wave propagation, and local





Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures, Figure 11

Ground-motion simulation using isochrone theory for a hypothetical  $M_w$  6.7 earthquake, buried at 2.5 km depth and embedded in a layered medium. *Top graph:* Rupture spreads from the hypocenter (red star) at constant rupture speed ( $v_r = 2.7$  km/s, white contours) over the fault with heterogeneous slip (color-coded). *Bottom graphs:* Isochrone quantities at three sites. *Top-most panels:* S-wave arrival time function; *2nd row of panels:* isochrones overlain over the real part of the integrand in Eq. (8) (in units of  $s_r \cdot G_0^S \cdot c(y, x)$ ); *3rd row of panels:* isochrones overlain over isochrone velocity (in km/s). *Lower-most panels:* Resulting fault-normal velocity seismograms dominated by large S-wave pulses (in cm/s; peak velocity indicated at the end of each trace)

site effects,  $l_k(t)$ , due to the small-scale geological conditions at each observation point. The wave propagation from the earthquake to the observation point depends on the complexity of Earth structure, which can be separated into two parts: (i) deterministic wave-propagation within a layered medium, three-dimensional basin effects, topographic features and large-scale geological structures; (ii) wave-propagation in a stochastic random medium with small-scale heterogeneity leading to incoherently scattered wave energy. Local site effects in the shallow near-surface structure (usually the top-most 100m) beneath the observation site further complicate near-fault

ground-motions, leading to either increased or decreased motions (compared to bed-rock level), increased shaking duration, shifts in the dominant frequency of ground-motion, and perhaps nonlinearity effect. These wave-propagation effects have to be incorporated into the characterization of ground-motion complexity, and should be applied in strong-motion simulations for predicting the intensity and variability of seismic shaking. However, the success in incorporating these effects hinges on the available knowledge of seismic properties of the Earth, which in turn depends on the dominant wavelengths (frequencies) at which the Earth is sampled by observational data.

Similarly, inversion for Earth structure and forward simulation of seismic waves require an appropriate spatial discretization of the modeling region for the numerical calculations, which essentially provides an upper limit on either the maximum resolvable frequency or the size of the computational domain for ground-motion simulations.

In this Section, I first review some aspects of “deterministic” wave propagation in a flat-layered attenuating Earth, and then qualitatively describe effects of basin structures, Earth topography and local features (e. g. fault zones, narrow belts of low shear-wave velocity). I then examine aspects of wave propagation in random media and corresponding simulation methods. Site effects and nonlinearity conclude this section. Due to the large body of literature on these topics, a detailed quantitative review is not attempted in this Section. Instead I merely select a few key elements of particular interest for characterizing ground-motion complexity.

### Wave-Propagation in a Flat-Layered, Attenuating Earth

The first-order approximation of Earth’s internal structure is a radially layered medium, which for short source-site distances ( $< 600$  km) can be treated as a flat-layered structure. Many research branches in seismology successfully apply this approximation of a one-dimensional depth-dependent velocity-density model: earthquake location, arrival-time measurements of various seismic wave types, focal-mechanism determination, moment-tensor inversion, finite-source inversion. A number of well tested methods exist to compute Greens functions for a flat-layered Earth model, although the exact frequency range in which this model is accurate is difficult to determine and depends on the particular application and region. At long periods ( $T > 20$  sec and corresponding long wavelengths ( $\lambda \approx 120$  km for P-waves traveling at speed of  $v_p \approx 6$  km/sec), used for instance in moment-tensor inversions), basin structures, mountain ranges and other geologic features with minor changes in physical properties (density, wave-speed) can be ignored at spatial scales on the order of tens of kilometers, thus allowing accurate modeling of the complete low-frequency seismogram. For small areas with relatively simple Earth structure, body-wave Greens functions can be accurately synthesized up to frequencies of  $f \approx 1$  Hz, a property that is often used in finite-source inversions. In these cases, ground-motion complexity is assumed to be largely due to the heterogeneity in the rupture process, since the local, observer-specific Greens functions (containing the impulse response in a layered medium

due to point-source excitation) are simple (examples are shown in Fig. 5).

In many cases, however, sites of interest may be located on top of basin structures, in zones that are geologically different on a small-scale, or even in a narrow fault-zone belt which may have distinctly different seismic properties. By examining observed ground-motions at many sites, site-specific one-dimensional Earth models can be used to compute corresponding Greens functions, a cumbersome and error prone approach since source and receiver are located at points with different seismic properties. In such cases, 2D or 3D-models are preferable for the Greens function calculation, as they allow to appropriately include basin effects, local geology and potentially also topography, usually at the expense of reduced frequency resolution and/or a smaller computational domain.

### Effects of Sedimentary Basins, Fault-Zones, Topography

Geological basins, containing layers of compliant sedimentary rock units and covered by potentially poorly consolidated, low shear-wave-velocity sediments, have a variety of effects on incoming seismic waves. Characterizing and quantifying the basin-related ground-motion complexity is particularly important for seismic hazard studies since many major urban areas are built in geological basins that are located within or close to a seismically active region (e. g. Mexico City, Tokyo, Los Angeles).

For instance, the 1985  $M_w$  8.0 Michoacan earthquake generated major damage in Mexico City, several hundreds of kilometers away from the epicenter. While the near-source strong-motion stations did not show unusual ground motions, the shaking level in Mexico City was unexpectedly large in those parts of the town constructed on lake bed sediments [9]. Ground-motions and associated damage exhibited high spatial variability; long-period surface waves, generated at the edge of the basin, had particularly severe consequences for tall high-rise building with eigenperiods of several seconds. Many of them suffered complete collapse. Strong basin effects have been simulated for earthquakes in the greater Los Angeles area, indicating localized basin-response effects caused by surface waves generated at the edge of the basin [74,75,135,136,138] and the focusing of seismic energy due to basin geometry [76,108,138]. A particularly interesting effect has been found for the 1994  $M_w$  6.7 Northridge earthquake, attributed to an isolated high-damage area in Santa Monica to focusing effects caused by a small-scale localized 3D lens-like high-velocity structures [57].

Other prominent examples of strong basin response are given for the Wellington Basin (New Zealand) [24], for the Tokyo plain (Japan) [109,159,160,162], and the Kobe area (Japan), which was strongly damaged during the 1995 Kobe earthquake ( $M_w$  6.9). In the latter case, the largest damage did not occur in direct vicinity of the rupturing fault, but was concentrated within an elongated band offset to the southeast of the fault. Modeling studies showed that this “damage belt” was caused by constructive interference of seismic waves taking different paths: (i) directly through the low-velocity basin structure and (ii) through high-velocity rocks outside the basin and then reflected back into the basin [107,143,144]. Additionally, topographic effects may have contributed to the complexity of near-fault motions for the Kobe earthquake [145]. Another example for strong topographic effects on near-source motion is the Tarzana (California) site, located atop a small hill, which recorded a value of  $PGA = 1.78$  g on the EW-component of motion during the 1994  $M_w$  6.7 Northridge earthquake [174]. However, it is difficult to separate topographic effects from subsurface-structure properties beneath the topographic features, and firm conclusions on the contribution of topographic effects are not well defined [108].

These basin-induced or topographic site effects strongly depend on the direction from which the incoming wave field arrives, as shown by strong-motion simulations for the Basel (Switzerland) area [139] or the greater Los Angeles region [118,138]. Strong variations in the basin response, and hence ground-motions at individual sites, depend on whether the incoming waves are predominantly polarized in the direction parallel or perpendicular to the main geologic structures (i. e. the large-scale basin shape and major faults inside and bounding the basin).

Low-frequency ground-motions, modulated by basin effects, topography, or narrow fault-zone-related regions of low shear-wave velocity, exhibit also longer shaking duration (due to surface-waves arriving after the dominant S-arrivals), and potentially larger amplitudes in case of constructive interference of wave packets (often occurring as trapped waves) [141]. Due to the dependency of the basin response on the direction of the incoming wavefield and the detailed small-scale structure of the basin, individual sites within a particular basin will experience very different ground-shaking. The ground-motion complexity is thus greatly increased by these effects, and general scaling relations for ground-motion (de-)amplification and prolongation due to basin and topographic structures are difficult to derive. Only numerical simulations for particular earthquakes or specific scenario events can help to understand the corresponding ground-motion variability which

can then be related to standard ground-motion attenuation relationships (see Sect. “Ground-Motion Scaling Relations”).

While there has been considerable progress in interpreting and modeling strong-motion waveforms for frequencies  $f < 1$  Hz, one of the major challenges is to calculate reliable broadband near-source seismograms for the frequency range of engineering interest which extends to  $f \sim 10$  Hz. The works described above examine and model the low-frequency wavefield contribution, using 1D-, 2D- or 3D-finite-element, finite-difference or spectral-element techniques (see ► [Seismic Wave Propagation in Media with Complex Geometries, Simulation of](#)). Seismic source properties, Earth structure, and site effects strongly affect high-frequency motions, but computational limitations still prohibit purely deterministic ground-motion simulations for frequencies above  $f \approx 1$  Hz. Instead, at least some part of the simulation procedure needs to involve a stochastic component since Earth structure, and to some extent the earthquake source, are essentially unknown at short spatial scales required for accurate high-frequency simulations. The difficulty is thus to capture and correctly quantify the scattering properties of the Earth at the scale-lengths and frequencies that are of interest for seismic hazard purposes.

### Scattering in Inhomogeneous Media

The seismic coda, the energy in the seismogram after the prominent direct P- and S-wave arrivals, consist of P-, S- and surface-wave energy scattered in the inhomogeneous rock volume between the source and the recording site. Figures 1, 3, 4, display near-field waveforms for which the coda-characteristics show large variability. In some cases, the coda waves decay fast, in other cases they persist for a long time; the frequency content of coda waves also appears to be site dependent. These observations manifest that, scattering and attenuation of seismic waves in inhomogeneous media strongly contribute to ground-motion complexity.

The basic method of estimating attenuation properties of seismic waves uses the Fourier amplitude spectrum of observed ground motion  $\mathbf{u}(r; f)$ , for a spherical S-wave of frequency  $f$  given as

$$\mathbf{u}^S(r; f) \propto \frac{1}{r} \cdot \exp \left[ -\frac{\pi \cdot f \cdot r}{Q_S \cdot \beta} \right] \quad (9)$$

where  $r$  is distance,  $\beta$  the S-wave velocity, and  $Q_S$  is the attenuation coefficient for S-waves, containing both intrinsic anelastic attenuation and scattering attenuation. For S-waves,  $Q$ -values in the lithosphere range from  $20 \leq$

$Q_S \leq 10\,000$ , for P-waves they are about a factor of two larger [158]. However, the frequency dependence of  $Q_{S,P}$  above about  $f \approx 1$  Hz indicates that attenuation increases for higher frequencies as  $Q_{S,P}^{-1} \propto f^{-n}$  (with  $0.5 \leq n \leq 1.0$ ) [161]. While the effects of seismic wave attenuation due to geometrical spreading are well understood and can be readily measured and incorporated into ground-motion simulation methods. The scattering losses of seismic waves are more difficult to estimate.

Based on seismic array measurements, recorded coda envelopes or sonic logs from borehole measurements, several studies estimated characteristic scale lengths of seismic scattering in the Earth [71,93,152,190]. Alternatively, numerical simulations for assumed random-media realizations have been used to assess scattering properties of the Earth e.g. [65,66]. These works show that the heterogeneity spectrum of velocity fluctuations in the Earth (i. e. the random variations of P- and S-wave velocities around a depth-dependent velocity gradient) can be adequately modeled as a fractal medium (e. g. Eq. (3)) or with a correlation function that does not decay too rapidly in wavenumber domain. Smoothly varying Gaussian-type correlation functions are thus inappropriate whereas an exponential or a more general van Karman correlation function Eq. (4) well models seismic scattering in the Earth crust [93,159].

Due the tectonic history of the Earth and the multi-scale nature of geologic structures, scattering parameters depend on region and depth. Correlation lengths of  $a \approx 10$  km are inferred for wave-scattering in the lower crust e. g. [66], whereas [86] assumed  $a \approx 5$  km for upper-crustal wave-field simulations. Upper-mantle heterogeneities may be modeled with correlation length  $a \approx 20$  km. The actual velocity fluctuations are on the order of 2–10%. A concise review of seismic scattering and estimated scattering parameters is given in (see ► [Seismic Waves in Heterogeneous Earth, Scattering of](#)) [161], a thorough introduction is published by Sato and Fehler [158].

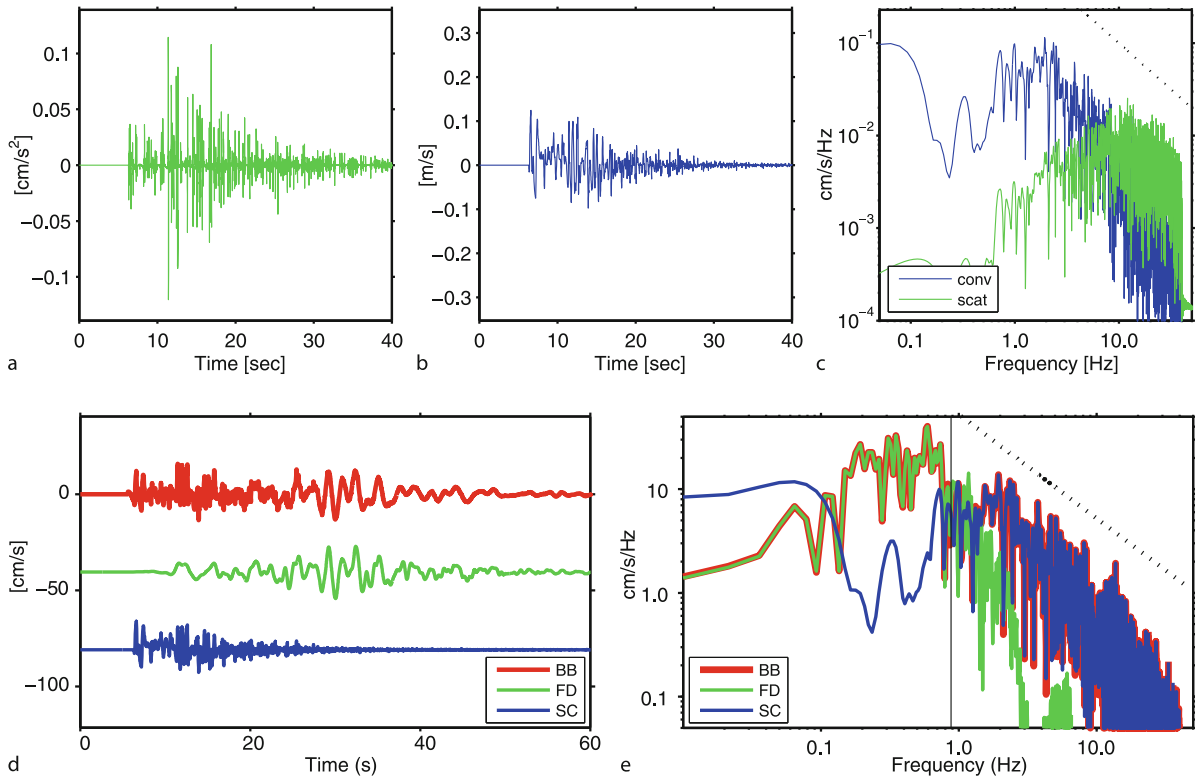
Ground-motion simulations for seismic hazard assessment or studies on the nature of ground-motion complexity should incorporate seismic scattering, i. e. or at least to some degree the stochastic nature of high-frequency seismograms. In earthquake engineering random-vibration theory has been used for this purpose e. g. [35,103], which oversimplifies the earthquake rupture process but appropriately accounts for the apparent randomness of high-frequency ground-motions. More advanced techniques combine deterministic low-frequency motions with stochastic high-frequency signals [99,146,147], in which the stochastic part reflects the short-scale variability in source prop-

erties and seismic-wave scattering. The stochastic signal used in these methods is generated as random white noise and does not contain any physical scattering mechanism.

More physical approaches calculate the 3D-seismic wavefield in a 3D-heterogeneous medium, a computationally expensive task which cannot be carried out for large-scale simulations or many scenario events for which high-frequency ground-motions are needed. To devise a simplified method, [86] have combined realistic small-scale heterogeneity in the source properties with scattering operators for a von-Karman random medium to compute broadband time histories. Their study indicates that the effects of scattering are masked by the heterogeneity in the kinematic source characterization. Note also that, near-field seismic wave scattering in a heterogeneous medium may also replicate apparent nonlinear site effects [130] (see Section “[Site Effects and Nonlinearity](#)”).

A computationally less demanding approach to include seismic scattering into ground-motion simulation uses radiative transfer theory to model the space-time distribution of the seismic-energy envelope due to scattered waves [192,194]. In this formulation, a time-dependent multiple S-to-S scattering process occurs due to a shear-dislocation point source embedded in a 3D-medium with background velocity  $v_0$  in which point-like isotropic scatterers of cross section  $\sigma_0$  are randomly distributed with density  $N$ . The total scattering coefficient is thus  $g_0 = N \cdot \sigma_0$ . The detailed theoretical developments are beyond the scope of this article, but it is worth noting that the calculation of the energy-density envelopes is computationally efficient and can be used for high-frequency ground-motion simulations [193].

The multiple S-to-S scattering theory has recently been applied to compute hybrid broadband near-field seismograms [125]. Their technique joins low-frequency 3D-finite-difference (FD) synthetics (which may contain the effects of a sedimentary basin) with site-dependent high-frequency scattering operators to synthesize broadband ground motions. The high-frequency scattering operators are convolved with an appropriate source-time function, and then these site-dependent “scatterograms” are combined in the Fourier domain with the corresponding low-frequency synthetics using a phase-matching optimization technique [124] (conceptually depicted in Fig. 12). Comparing data and simulations for a site that recorded the 1994 Northridge earthquake indicates a good agreement in terms of waveforms, amplitude spectra and spectral acceleration (Fig. 13). The model bias (logarithm of the ratio between the observed and simulated quantity) for spectral acceleration at 30 sites that recorded the 1994 Northridge event (Fig. 14a) shows only small deviations from zero, ex-



Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures, Figure 12

Conceptual diagram for computing hybrid broadband seismograms. **a** site-specific scattering Greens function for a point-source at the hypocenter. **b** time series representing the local high-frequency scattering contribution (obtained for instance by convolving the scattering Greens function in **a** with an appropriate slip-rate function). **c** Fourier amplitude spectra for the time series in **a** and **b**; the velocity-scatterogram decays as  $1/\omega$  (dotted line) beyond the corner frequency. **d** broadband seismogram (top) computed by combining the LF-seismogram (center) with the site-specific HF-scatterogram (bottom). **e** amplitude spectra for the time series in **d**; the spectra of the broadband synthetics exactly represent the LF-motions at low-frequencies and the HF-scattering contribution at high frequencies

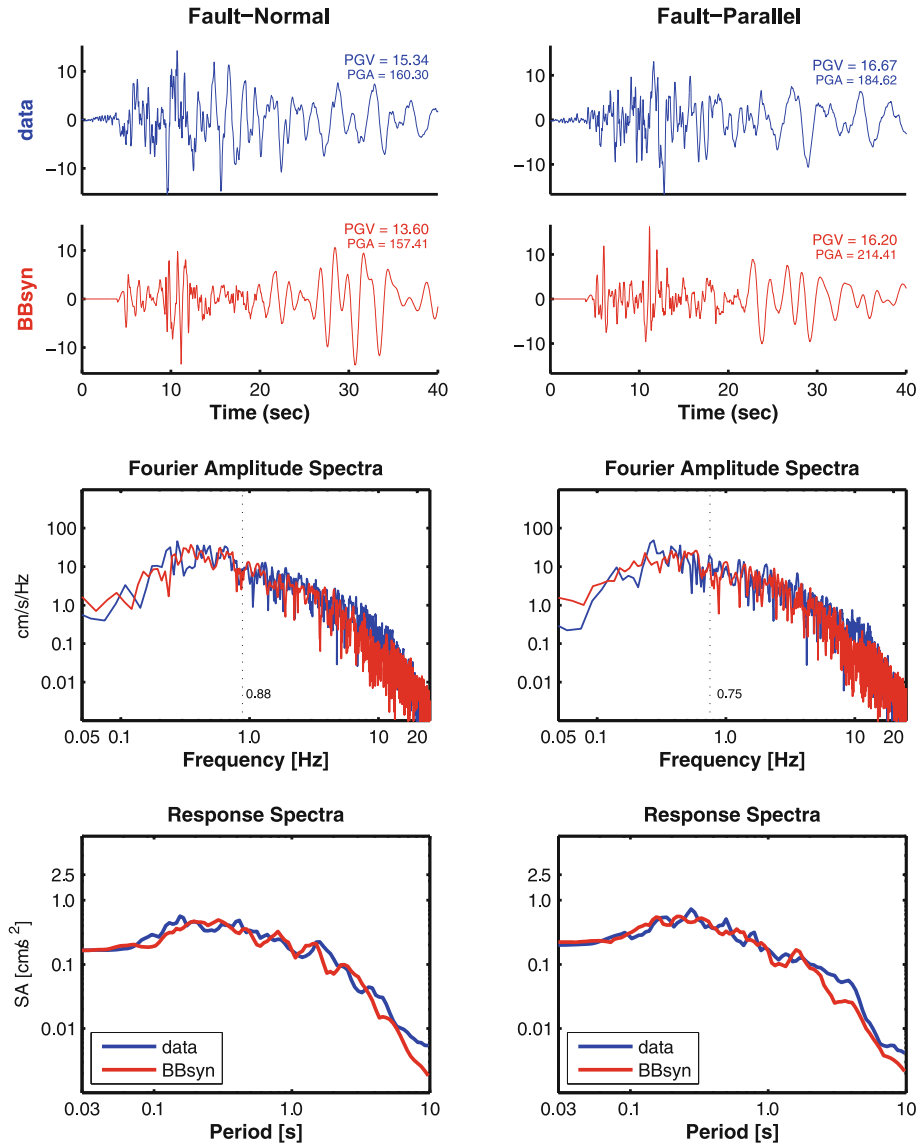
hibiting also narrow 90%-confidence limits. Considering also the general agreement between observed and simulated PGV and PGA values (Fig. 14b) indicates that such hybrid broadband wavefield simulations can reproduce ground-motion complexity, where differences at individual site remain due to unmodeled site and path effects.

Adding seismic scattering into ground-motion characterization and simulation helps to capture the large degree of complexity in the near-source seismic wavefield, and thus presents an important future research topic. Here I have only briefly discussed a few key elements of seismic scattering and its application to a specific case study. However, to fully explore the range of realistic scattering parameters, combined with the inherent complexity of the source-rupture process, requires extensive numerical simulations that need to be calibrated and validated against observational data.

### Site Effects and Nonlinearity

Site effects play a major role in characterizing and quantifying strong ground motion as they may amplify or deamplify the incoming “bedrock” motions in the uppermost velocity layers beneath the observer. Since site-amplification factors may reach two orders of magnitude [108], these effects cannot be neglected in earthquake engineering practice. A comprehensive review of various approaches to estimate and model site effects is given by Kawase [108], so I restrict the following discussion to a few key aspects.

There is no strict boundary between site effects and path effects, but site-effects usually refer to wave-field modifications in the immediate vicinity of the observer location (top 30 m of the local sedimentary cover). Usually, the time-averaged shear-wave velocity in the top 30–100 m



Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures, Figure 13

Comparison of data (blue) and broadband simulations (red) at a site that recorded the 1994  $M_w$  6.7 Northridge earthquake (*left*: fault-normal; *right*: fault parallel motions). Note the consistency between the Fourier amplitude spectra (vertical line denotes the matching frequency) and between the response spectra ( $\zeta = 5\%$  damping). PGA- and PGV-values are also similar

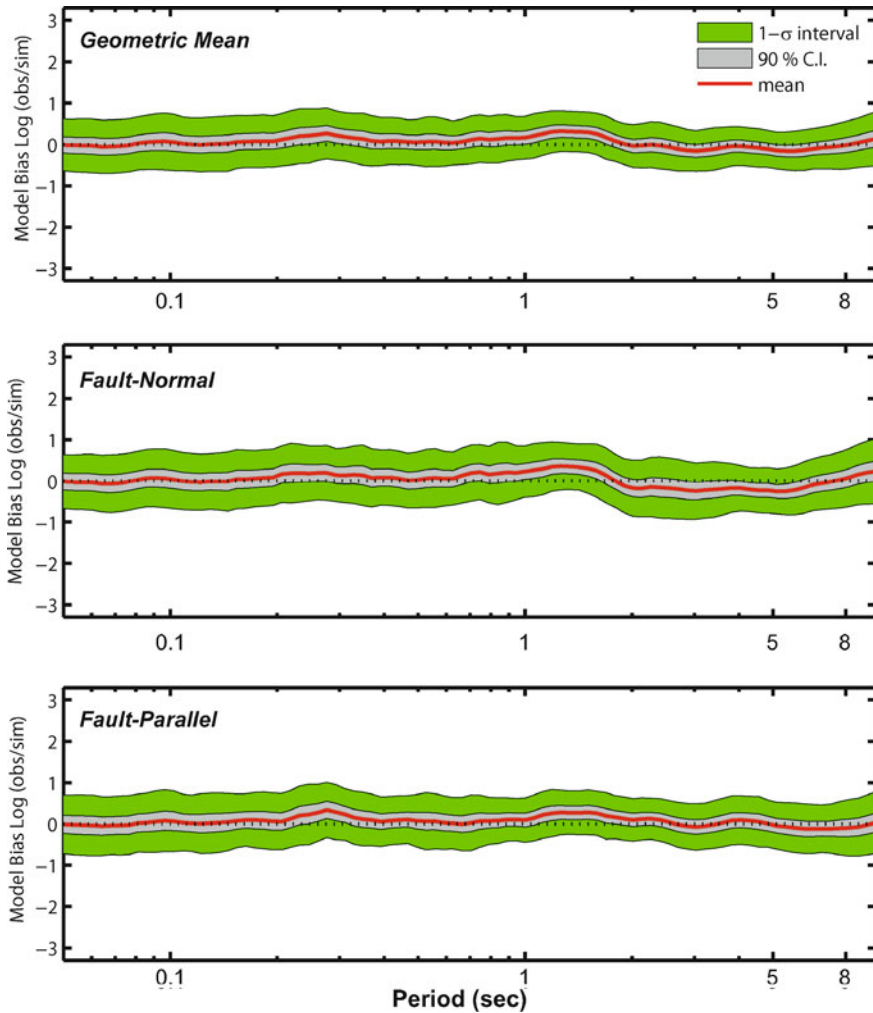
( $V_{S30}$ ) is used to define the station-specific soil classification. Site effects may be affected by the water table, and even soil-building interaction can be considered a site effect.

The classical method to estimate site effects from seismic observations is based on Eq. (1), and attempts to separate source and path from site effects. Expressing Eq. (1) in the Fourier domain and considering the ground motions due to earthquake  $j$  observed at site  $k$ , we obtain source-

site specific Fourier amplitude spectra as

$$A_{jk}(f) = S_j(f) \cdot G_{jk}(f) \cdot L_k(f). \quad (10)$$

The Green's function term,  $G_{jk}(f)$  can be expressed in terms of geometrical spreading factors with respect to the source-site distances,  $r_{jk}$ ; combining intrinsic and scattering attenuation into a common  $Q(f)$  one obtains (for



Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures, Figure 14a

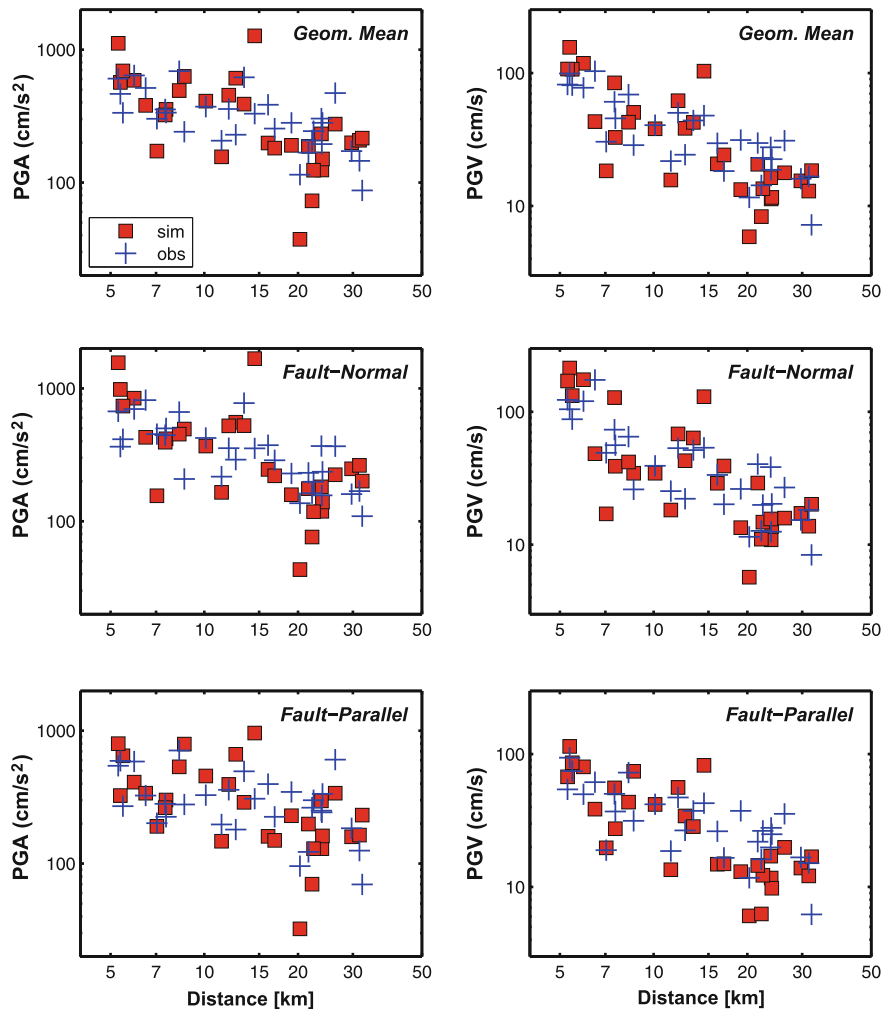
Hybrid broadband ground-motion calculations using low-frequency finite-difference synthetics and high-frequency scattering operators. The *three panel* show, for the geometric mean (*top*), the fault-normal (*center*), and fault-parallel (*bottom*) component, the model bias, computed as  $\log\left(\frac{\text{obs}}{\text{sim}}\right)$ , for spectral acceleration at 30 sites that recorded the 1994 Northridge earthquake

S-waves)

$$G_{jk}(f) = \frac{1}{r_{jk}} \cdot \exp\left[-\frac{\pi \cdot r_{jk}}{\beta \cdot Q(f)}\right] \quad (11)$$

where  $\beta$  is the representative shear-wave velocity over the entire path. Eq. (10) can then be solved by adding at least one independent constraint on either source, site or path. A common approach uses the average site factor as a reference, meaning that the logarithm of the sum of all site factors is equal to zero. Other methods select the site with the smallest site factor as reference, or make an a priori reference-site selection based on geologic or other information.

A recent method for site-effect estimation uses microtremor, ambient noise measurements, or aftershock recordings. Taking the spectral ratio of the horizontal component of motion with respect to the vertical component (so called H/V-ratios), these microtremor ratios often show similar site characteristic as inferred by independent methods [108]. However, the exact physical meaning of this H/V-ratio is not well understood, and since H/V-measurements do not directly represent the true soil amplification, those need to be supplemented and calibrated by numerical modeling [69,178]. Nonetheless, H/V-measurements are useful for microzonation studies which the detailed properties of the uppermost soil layers for a small



Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures, Figure 14b

Comparison of observed (*blue crosses*) and simulated (*red squares*) peak-ground acceleration (*left*) and peak-ground velocity (*right*) for 30 strong-motion sites that recorded the 1994  $M_w$  6.7 Northridge earthquake. Note the overall agreement between simulated and recorded motions. Discrepancies at individual stations are due to unmodeled ground-motion complexity

area, thus characterizing the susceptibility of particular locations to strong site effects [70].

Site-correction factors have been derived also from the seismic coda. Because coda waves essentially consist of scattered S-waves, local site amplification of the coda itself should be an average of the S-wave amplification factor for S-waves that arrive from different azimuths under different incidence angles. However, the assumption that the seismic coda consists only of scattered S-waves is too simplistic, because surface waves and strong P-coda waves may contaminate the S-wave coda. Site-effect estimation based only on S-coda measurements should be therefore interpreted with care [108].

A common assumption in seismic wave propagation is that strong and weak motions are affected in an identical manner. Analysis of near-field ground-motions provides evidence for nonlinear effects (i.e. a non-linear stress-strain relationship) [26,64,156], due to lower shear-wave velocity and increased damping within the sedimentary cover [54]. Local (or temporal) modification of the seismic properties of the rock material underneath a site or along the wave path may shift resonant modes to lower frequencies and generate reduced amplitudes. Nonlinearity occurs once a certain ground-motion intensity threshold is exceeded. Previous studies suggest  $PGA > 0.3$  g, or  $PGV > 20$  cm/sec, or peak-strains in excess of 0.06% [9],



roughly consistent with values inferred for strong-motion recordings of the 1989  $M_w$  6.9 Loma Prieta earthquake for which the threshold for large S-wave travel-time delays in repeating earthquakes occurs at  $PGA > 0.4$  g or  $PGV > 40$  cm/sec [156].

Effects of nonlinearity greatly increases the complexity of ground-motions, but only on a very localized scale. Recently, detailed spectrogram analyses of seismic recordings of the 2003  $M_w$  8:3 Tokachi-Oki (Japan) earthquake demonstrate liquefaction, quantified by a dramatically reduced high-frequency contents of the waveforms [55]. However, it is very difficult to establish general scaling laws that account for potential nonlinearity effects. In case of water-saturated loose sands, strong shaking may result in nonlinear soil liquefaction due to a rapid, temporary increase of pore water pressure, resulting in a dramatic loss of soil stiffness. Liquefaction effects are extremely important for earthquake engineering, but have received little attention from seismology in the past.

### Ground-Motion Scaling Relations

Despite the fact that observed near-source ground-motions show complicated time histories and large variability in intensity measures, there is the need in earthquake engineering and seismic-hazard analysis to devise simple approaches for estimating expected ground-motions in future earthquakes. For this reason, ground-motion prediction equations (GMPE's) are developed, providing mathematical expressions that relate shaking intensities to seismological quantities, source-site geometry, and potentially site-specific parameters. Thus, the complexity of the earthquake rupture process, the wave-propagation effects from the source to the site, and the detailed site conditions are condensed into relative simple functional forms, comprising a few parameters that constitute empirical ground-motion scaling relations.

### Development of Empirical Scaling Relations Based on Recordings of Past Events

Using strong-motion observations, ground-motion prediction equations are developed using a number of different parameterizations. The most fundamental form is given as [52]

$$Y = a_1 \cdot e^{a_2 M} \cdot R^{a_3} \cdot e^{a_4 r} \cdot e^{a_5 F} \cdot e^{a_6 S} \cdot e^\sigma \quad (12)$$

where  $Y$  is the ground-motion intensity measure of interest ( $PGA$ ,  $PGV$ ,  $S_A$ , ...).  $M$  is the earthquake's magnitude,  $r$  the source-to-site distance;  $F$  is a parameter that characterizes the type of faulting, and  $S$  captures the local site

conditions.  $R$  is an additional magnitude-dependent distance function which can take on alternative forms, for example [52]

$$R = \begin{cases} r + c_7 \cdot e^{c_8 M} \\ \sqrt{r^2 + [c_7 + e^{c_8 M}]^2} \end{cases} \quad (13)$$

This model contains some of the basic physics of earthquakes (terms with  $M$  and  $F$ ), the attenuation due to geometrical spreading of seismic waves (terms related to distance  $r$  and  $R$ ) and the site conditions ( $S$ ). Due to the approximately log-normal distribution of ground-motion intensities, attenuation relations are parameterized using the natural logarithm of  $Y$ . Correspondingly, the standard deviation of the zero-mean random error term  $\sigma$  is estimated as the standard error of  $\ln Y (\sigma_{\ln Y})$ .

Over the years, different models have been proposed to capture various aspects of the inherent complexity of near-fault shaking. One class of models assumes a functional form whose shape is magnitude-independent for all distances. A simple parametric form is given by [38]

$$\begin{aligned} \ln(Y)_{M,R,F} \\ = a_1 + a_2 \cdot M + a_3 \cdot M^2 + a_4 \cdot \ln(r + a_5) + a_6 \cdot F. \end{aligned} \quad (14)$$

However, many observations show that ground-motion intensities saturate at close distances to the fault, i. e. moderate-magnitude earthquakes ( $5.0 \leq M \leq 6.5$ ) may generate about the same level of high-frequency shaking as large magnitude events ( $M > 6.5$ ). The second class of ground-motion scaling relations captures these observations. For instance, starting from Eq. (14) Abrahamson and Silva [5] propose the relation

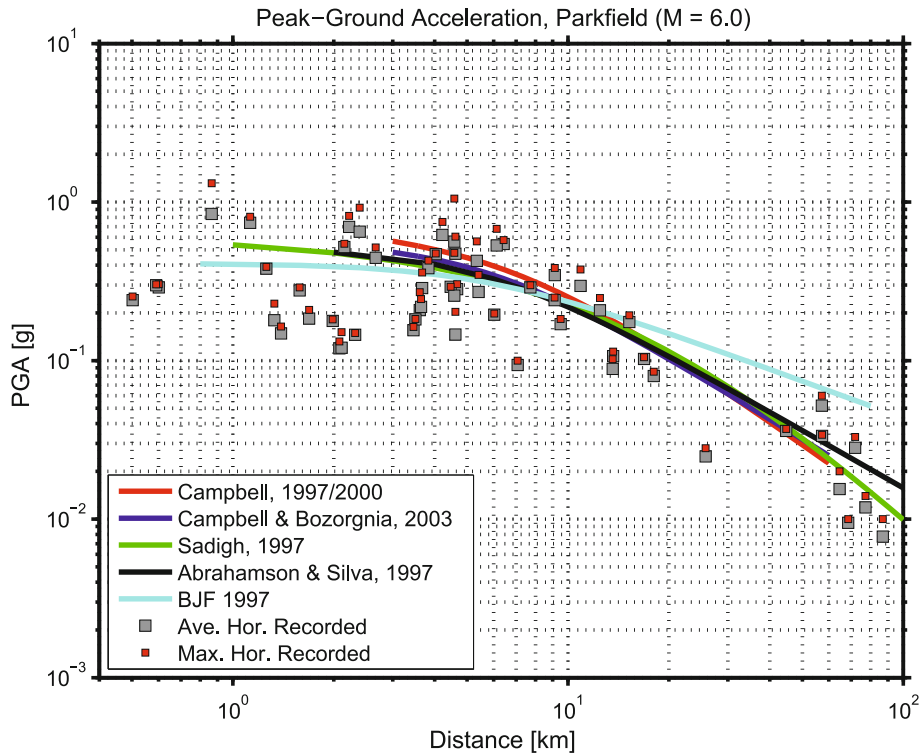
$$\begin{aligned} \ln(Y)_{M,R,F} = a_1 + a_2 \cdot M + a_3 \cdot M^2 \\ + (a_4 + f_1(M)) \cdot \ln(r + a_5) + a_6 \cdot F \end{aligned} \quad (15)$$

with  $f_1(M) = a_7 M$ , while Sadigh et al. [157] developed an attenuation model given by

$$\begin{aligned} \ln(Y)_{M,R,F} \\ = a_1 + a_2 \cdot M + a_3 \cdot M^2 + a_4 \cdot \ln(r + f_2(M)) + a_6 \cdot F \end{aligned} \quad (16)$$

where  $f_2(M) = c_7 \cdot e^{c_8 M}$ . The difference between these two scaling relations becomes evident at larger distances ( $r > 50$  km), while for short distances they lead to about the same ground-motion intensities (Fig. 15).

Empirical ground-motion prediction like those presented in Eqs. (14), (15), (16) have been widely applied to



Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures, Figure 15

Observed peak ground accelerations for the Sept 28, 2004  $M_w$  6.0 Parkfield earthquake, compared against five different empirical attenuation relationships. Data are plotted using the geometrical mean of the two horizontal components (fault-normal and fault-parallel) ( $PGA = \sqrt{PGA_{FN} \cdot PGA_{FP}}$ ), and for the maximum of the two components. The median of the observations is generally in good agreement with the empirical predictions, but the data exhibit large variability with much higher (and lower) motions than empirically predicted, independent of the closest distance to the fault. Note that the recording at station “Fault Zone 16” is not shown as it was not available at the COSMOS database. The PGA-value for site “Fault Zone 16”, located at  $\sim 0.6$  km distance from the fault, is estimated to be in the range 2.0–2.5 g [167]

global and regional ground-motion data sets, but a number of observations call for modification to these “generic” models. First of all, attenuation models for different tectonic provinces have been published accounting for variations in the regional geology [7,8,17,18,31,67,175]. While there is the need to distinguish between stable-cratonic areas (like Eastern North America), extensional regimes (like the Basin and Range province in the western US), zones of complex active tectonics (like California), or subduction zones (like in Mexico, Chile, Japan), recent work [7] suggests that data from different tectonically active areas with crustal seismicity can still be jointly analyzed. In certain regions the geologic conditions cause the presence of several of such tectonic regimes over a rather small area (e.g. Italy, Switzerland, Japan) which further complicates the development of adequate ground-motion prediction equations.

Besides the local geology, the source-site geometry greatly influences seismic shaking. In the near-fault

regime, ground-motions are strongly affected by geometrical source-site effects, leading to large long-period pulses of motion termed “directivity pulses” (see Sect. “**Isochrone Theory**”). Recent work includes such directivity pulses in the attenuation equation [170,171]. Large permanent offset of the ground due to the rupture process will also generate long-period large motions; this so-called “fling-step” is distinguished from the directivity pulse [32]. For thrust-faulting earthquakes, the “hanging-wall” effect has been shown to generate very different ground motions [3] for sites located on the hanging-wall than on the footwall (Fig. 3). This effect is explicitly included in several recent attenuation models [3,53].

An intriguing observation has been made following the large surface-breaking earthquakes in 1999 in Turkey (the Aug. 17  $M_w$  7.5 Izmit and the Nov. 12  $M_w$  7.1 Duzce events) and Taiwan (the Sept. 20  $M_w$  7.6 Chi-Chi event): near-field ground-motions were significantly lower than empirically predicted [105]. This counter-intuitive result

can partly be explained by the general differences in stress-release patterns between buried-faulting and surface-rupturing earthquakes. The primary cause for the ground-motion variation between these two classes of earthquakes lies in general differences in the dynamic rupture process and the associated energy balance of the system (see ► [Earthquake Scaling Laws](#)). Ultimately, these effects could be related to depth-dependent fault-zone properties which exhibits highly damaged rocks with intense microcracking and lower seismic velocity close to the surface, but more compliant and less damaged material at depths [22]. Fault-zone structure affect the dynamics of the rupture process [13,21,22,83] and the resulting seismic radiation [141], and therefore strongly influence the resulting ground-motion intensities.

Besides the functional form and the given geological, geometrical, and physical parameters a number of technical issues affect ground-motion scaling equations (a comprehensive review is given in [60]): (i) the choice of the ground-motion intensity measure  $Y$  determines the values for the coefficients  $a_i$  (Eqs. (12)–(16)) derived by regression analysis; (ii) the specific approach for the regression (data fitting) matters (common techniques are weighted non-linear least-squares regression, two-step regression, or random-effects regression); (iii) the data selection criteria for assembling the individual strong-motion database; (iv) data corrections may be applied to the recorded strong-motions (e. g. uniform instrument response for all records, baseline correction, frequency filtering) and also to the meta-data (magnitudes, distances, site information etc.) which represent the independent parameters.

In developing ground-motion prediction equations, a uniform and well-calibrated magnitude definition is particularly important, but in practice often difficult to achieve when merging datasets from various institutions whose magnitude values may not be compatible. The style-of-faulting factor  $F$  in Eqs. (14), (15), (16) is generally well defined, but the site-effect factor  $S$  is often poorly known. One of the most critical parameters in ground-motion scaling is the distance  $r$  between the site and the source. In particular for extended rupture planes, the applied distance metric for the source-to-site geometry becomes crucial; different source-to-site distance definitions are in use [4] but the particular choice for  $r$  in turn will affect the seismic-hazard calculations [163].

In an effort to harmonize and calibrate ground-motion data, related meta-data, and the development of empirical ground-motion prediction equations, the Pacific Earthquake Engineering Research Center (PEER) carried out the New Generation Attenuation of Ground Motions

(NGA) Project (completed in Jan. 2008). In this context, [37] for instance refine Eq. (14) to accommodate the effects of anelastic attenuation when modeling far-distance recordings ( $R > 80$  km) and to include an “effective” magnitude-dependent geometrical spreading (allowing to predict ground-motion amplitudes out to distances  $R = 400$  km). Their data-driven equation includes only terms that are truly needed to adequately fit the data, involving (i) a complicated magnitude-scaling function,  $f_M(M)$ ; (ii) a versatile distance function,  $f_D(R_{JB}, M)$ ; (iii) a site-effect function that includes potential nonlinearity effects,  $f_S(V_{S30}, R_{JB}, M)$ :

$$\ln(Y)_{M,R,F} = f_M(M) + f_D(R_{JB}, M) + f_S(V_{S30}, R_{JB}, M) + \epsilon \cdot \sigma_T. \quad (17)$$

All terms in Eq. (17) are period-dependent;  $M$  is moment magnitude,  $R_{JB}$  is the Joyner–Boore distance (the closest distance to the surface projection of the fault)  $\epsilon$  is the fractional number of standard deviations of a single predicted value of  $\ln(Y)$ , and  $\sigma_T = \sqrt{\sigma_{\text{intra}}^2 + \sigma_{\text{inter}}^2}$  describes the uncertainty by combining the intra-event and inter-event aleatory uncertainty. In this context, the exact definition for the ground-motion parameter  $Y$ , usually taken as the geometric mean of the two horizontal components, has received increased attention. Because of misaligned recording instruments or complicated (multi-)pathing of radiated waves, simple measures of “mean ground-motion” appeared to be incorrect when taking the sensor orientation as installed in the field. Works by [30,39,150] discuss the usefulness of orientation-independent ground-motion measures based on data and numerical simulations.

### Relating Earthquake Source-Scaling to Ground-Motion Prediction

When using empirical ground-motion attenuation relations for predicting the expected shaking level at a given site the analyst needs quantitative information regarding the type of faults and their properties, the fault locations with respect to the site, and the site conditions. Probabilistic seismic hazard analysis (PSHA) requires additional information about the recurrence rate of earthquakes on the chosen faults [111]. Therefore, earthquake statistics (see ► [Geo-complexity and Earthquake Prediction](#)) [111] and earthquake physics (see ► [Earthquake Scaling Laws](#)) [6] are major components for accurate seismic hazard calculations. However, ground-motion prediction Eqs. (14), (15), (16) only contain a faulting-style factor ( $F$ ) and the magnitude dependence as earthquake-source related parameters. A correct estimation of the potential magni-

tudes for earthquakes occurring on the selected faults is the most crucial step in any seismic hazard study. Earthquake physics (see ► [Earthquake Scaling Laws](#)) [6] provides the theoretical foundation for defining these magnitudes and to explain observational data and empirical relations e. g. [106,122,187].

In modern seismic hazard analysis, the magnitude  $M$  relates to moment magnitude,  $M_w$ , derived from the seismic moment as  $M_w = 2/3 \cdot \log M_0 - 6.07$ . Since  $M_0 = \mu \cdot L \cdot W \cdot D$ , source-scaling relations between observable fault dimensions (length  $L$ , width  $W$ , area  $A = L \cdot W$ , average displacement  $D$ ) and magnitude are used in empirical ground-motion prediction for obtaining self-consistent input parameters.

Based on geologic observations it is often possible to estimate the length of a fault, considering also cases with multiple segments that may rupture in individual events or jointly (these cases are then treated in PSHA using logic trees) [188]. Estimating fault width is more difficult; it may be inferred from the location of background seismicity in a crustal tectonic setting, or from modeling interseismic deformation which constrains the fault's locking depth, but often a generic fault width of  $W = [10 - 20]$  km is assumed. Several studies have published earthquake scaling relations between various fault parameters [79,120,122,169,186,187]; these source-scaling relations are not only useful for seismic hazard analysis but also provide important insight into earthquake mechanics (see ► [Earthquake Scaling Laws](#)) [164]. A generic relation between magnitude  $M_w$  and fault area  $A$  (in  $\text{km}^2$ ) is given by:

$$M_w = p + q \cdot \log_{10} A \quad (18)$$

For  $q \equiv 1$ , Eq. (18) is consistent with self-similar constant average-stress-drop scaling (see ► [Earthquake Scaling Laws](#)). Rule-of-thumb values for the coefficients in Eq. (18) are  $p = 4$  and  $q = 1$  (adapted from [187] who find  $p = 3.98$ ,  $q = 1.02$  for strike-slip earthquakes, and  $p = 4.07$ ,  $q = 0.98$  for all faulting styles), leading to a magnitude  $M_w 7$  earthquake for a fault area  $A = 1000 \text{ km}^2$ . However, a number of studies have found considerable deviations from self-similar earthquake scaling [79,122,169], and reported values in the range  $3.97 \leq p \leq 4.39$  and  $0.97 \leq q \leq 1.33$  [120]; these differences strongly affect empirically predicted ground-motion intensities and seismic hazard calculations [188,189]. For instance, [188] proposes  $p = 4.2$ ,  $q = 1$ , resulting in an  $M_w 7.2$  earthquake for  $A = 1000 \text{ km}^2$ , with twice the seismic moment and twice the displacement as an  $M_w 7.0$  rupture with identical source area. Correspondingly, Eqs. (14)–(17) lead to higher ground-motion estimates for the larger event,

consequently also different seismic hazard, due to small changes in the source-scaling relation.

### Quantifying Uncertainty in Ground-Motion Prediction Equations

Analyzing, quantifying and modeling strong-motion uncertainties is a mandatory part of any reliable seismic-hazard study [168]. The variability in source-scaling relations reflects in part the complexity of the earthquake rupture process, treated as an epistemic uncertainty (i. e. as more data become available and seismologists better understand earthquake dynamics, this variability will eventually decrease). The standard deviation of the residuals, obtained when deriving ground-motion prediction equations by regression analysis measures the aleatory variability (randomness) of ground-motion parameters. This standard deviation,  $\sigma_{InY}$ , is then partitioned into two error terms for the intra-event and inter-event variability,

$$\sigma_{InY} = \sqrt{\sigma_{inter}^2 + \sigma_{intra}^2}.$$

If the regression uses the geometric mean of the two horizontal components and the component-to-component variability is needed, a third term  $\sigma_{comp}$  has to be included. The intra-event variability can be further separated into a site-to-site component  $\sigma_{comp}$ , and the remaining variability  $\sigma_0$  (after accounting for source and site effects):

$$\sigma_{intra} = \sqrt{\sigma_s^2 + \sigma_0^2} \quad [52].$$

The standard deviation  $\sigma_{InY}$  has previously been found to be a function of magnitude (e. g. [5,157]) with decreasing variability for increasing magnitudes. However, this counter-intuitive result could be affected by a sampling bias due to fewer near-source strong-motion recordings for moderate-to-large earthquakes. More recent work suggests that  $\sigma_{InY}$  is independent of magnitude [2]. The exact value of  $\sigma_{InY}$  varies between different studies and ground-motion parameters, but is generally in the range  $0.4 \leq \sigma_{InY} \leq 0.8$ . Including this standard deviation in the ground-motion estimation is indispensable to assess the variability in shaking intensity – common practice is to use at least one or two standard deviations. However, for critical infrastructures (e. g. nuclear power plants, nuclear waste repositories), it is not clear where to truncate the ground-motion distribution (i. e. how many  $\sigma$ 's to include, [177]) since very long return periods need to be considered and conclusive physical arguments for upper limits on near-field ground motions have not yet been made. Bommer et al. [33] demonstrate that the choice of the truncation level will significantly affect seismic hazard estimates at very low probabilities (annual frequencies of

exceedance of  $10^{-6}$  and lower), while the effect on hazard at probabilities traditionally used in PSHA (annual frequencies of exceedance of  $10^{-4}$ ) is small. The reason is that low-probability seismic hazard may contain contributions from rare but extreme ground-motion values.

The preceding discussion raises the issue of defining upper bounds of near-source ground motion [33]. For instance, the maximum values for *PGA* in the NGA-database is  $PGA = 1.56\text{ g}$  (for the Tarzana record of the 1994  $M_w$  6.7 Northridge earthquake), but values of over 2 g have been reported for the 2004 Parkfield earthquake [167,168] or the 2003 Miyagi (Japan) earthquake. Note that sites and events showing the largest *PGA*-values generally do not coincide with those showing the highest *PGV*-values. For instance, the largest *PGV*-value reported in the NGA-database is 205 cm/s (observed at station TCU068 for the  $M_w$  7.6 Chi-Chi earthquake, Fig. 4), but the *PGA* at this site was only 0.53 g. *PGV*-values in the NGA-database frequently exceed 75 cm/s, and values of 100 cm/s are not uncommon. Closer inspection of the rapidly growing online databank of near-source recordings, obtained with modern dense observational networks (e.g. the K-Net and KiK-net stations in Japan) may return even higher maximum *PGA*- and *PGV*-values than listed above.

Current research addresses the physical limits to maximum ground motions, imposed by the complexity of the source-rupture process, the wave propagation, the site conditions and the strength of the rock. Investigating the combination of these complex physical processes with numerical methods and large-scale simulations helps to better understand the overall distribution of ground-motion parameters and their complexity.

An innovative observational method to examine maximum ground-motion has been developed by Brune and colleagues [19,46,47,48,49,50], based on precariously balanced rocks. These are free-standing (large) boulders, created by erosional processes, which appear as if small ground-accelerations could overturn (topple) them; this toppling-acceleration can be measured. By also dating the age of these rocks it is possible to determine which ground-motion levels have not been exceeded in the corresponding time interval in this region. These “natural seismoscopes” therefore provide important constraints for maximum shaking levels and probabilistic seismic hazard analysis.

### Comparing Observations with Empirical Predictions

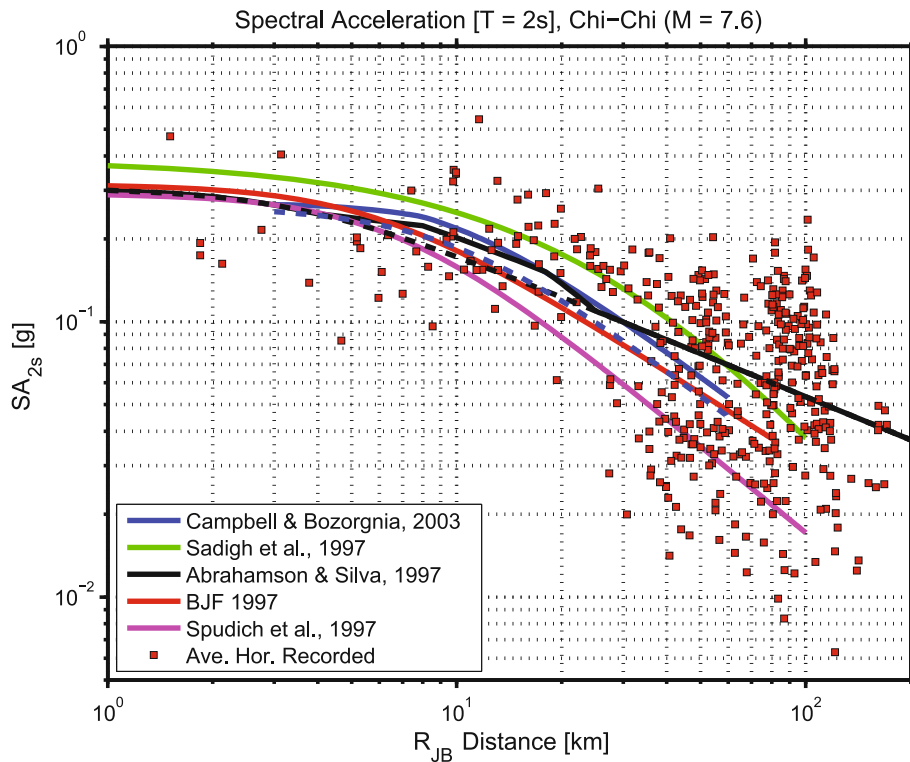
The introduction of this article presents near-field ground-motion observations for two recent well recorded earth-

quakes, the  $M_w$  6.0 strike-slip earthquake in Parkfield (09/28/2004) and the  $M_w$  7.6 thrust-faulting earthquake in Taiwan (09/20/1999) (Figs. 1–4). I now compare these records against empirical predictions for a number of widely used ground-motion attenuation relations. Figures 15 and 16 display recorded *PGA* for the Parkfield event and spectral acceleration ( $S_A^{T=2\text{sec}}$ ) for the Chi-Chi earthquake, respectively. In both cases, the empirical predictions are generally consistent with the median of the observations, but the large scatter in the data also leads to observations occur far outside the standard  $1\sigma$ -bounds of the predictions.

For the Parkfield data, observed *PGA*-values are both higher and lower than empirically predicted (for  $r < 100\text{ km}$ ), and no obvious pattern is visible that may explain the large ground-motion variability (Fig. 15). Peak ground acceleration close to and above 1 g were observed at a number of stations. At one site (station “Fault Zone 16”), the *PGA*-value is estimated to be on the order of 2.0–2.5 g [167]. The plotted *PGA*-values are not corrected or grouped according to the soil classification for each site, i.e. parts of the observed variability can be attributed to site effects. However, much of the variability originates from the particular position of each station with respect to the spatial complexity of the earthquake rupture process, and, to a lesser extent, from localized wave-propagation phenomena.

Examining recorded motions and empirical predictions for the Chi-Chi earthquake reveals an overall consistency between the observed median spectral acceleration ( $S_A^{T=2\text{sec}}$ ) and ground-motion attenuation relations, despite significant scatter in the data (Fig. 16). At large distances ( $R_{JB} \geq 50\text{ km}$ ) many of the recorded motions are significantly larger than any of the empirical prediction, while at distances close to the fault ( $R_{JB} \leq 10\text{ km}$ ) most of the  $S_A$ -values are significantly lower than predicted. This unexpected behavior of lower than-predicted ground-motions for large surface-breaking earthquakes has been recently observed for a number of earthquakes [105], and is attributed to effects of earthquake source dynamics occurring in the uppermost shallow and more compliant crustal layers [126].

These two examples illustrate the large variability of near-source motions recorded at a number of stations for the same earthquake, thus characterizing the intra-event variability described above. Site, path, and source effects are responsible for this large degree of ground-motion complexity, but separating the contributions of each of these effects is difficult. Deciphering the detailed physical processes and corresponding parameters that lead to a specific near-field ground-motion is an active area of research.



Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures, Figure 16

Spectral accelerations (SA) at  $T = 2$  s for ground-motion recordings of the Sept 20, 2004  $M_w$  7.6 Chi-Chi (Taiwan) earthquake, compared against five different empirical attenuation relationships. Data are plotted using the geometrical mean of the two horizontal components (fault-normal and fault-parallel) ( $SA = \sqrt{SA_{FN} \cdot SA_{FP}}$ ). The Campbell & Bozorgnia [53] and Abrahamson & Silva [5] relations contain a hanging-wall factor, indicated by the *solid blue* and *black lines*, respectively; the corresponding regular relations are plotted with *dashed lines*. The median of the observations is consistent with the empirical predictions, but the data exhibit large variability, with much higher (and lower) motions than empirically predicted. Note that at large distances ( $R_{JB} \geq 50$  km) many sites reveal higher-than-predicted spectral accelerations while at very short distances ( $R_{JB} \leq 10$  km) many sites exhibit significantly lower ground-motions.

Ultimately, this will allow us to better model ground-motions and their variability for future earthquakes, a key ingredient for improved seismic hazard analysis.

### Future Directions

This article reviews the complexity of near-field ground motions, generated by the space-time-dependent heterogeneous earthquake rupture process, transformed by wave-propagation through complex geologic structure and inhomogeneous media, and finally subjected to localized site conditions. As a consequence, the waveform complexity recorded at dense strong-motion arrays provides both a challenge and an opportunity for future research.

A largely overlooked aspect of ground-motion complexity pertains to the rotational motions (also called vorticity) of the displacement field. Translational ground

displacements (velocity, acceleration) are recorded as seismograms used for monitoring seismic activity and ground motions; strain measurements capture the deformation of the Earth, but the theoretically predicted [6] ground rotation (a vectorial quantity) has rarely been reliably measured in the past. Observations of rotational motions are challenging, due to their small amplitudes [42] and inadequate instrument sensitivity, but recent work provides evidence that rotational motions are significant [95,97,98,179]. Moreover, they are particularly important for engineering applications [182,183] (see ► [Earthquake Source: Asymmetry and Rotation Effects](#) for a comprehensive review on rotational motions). As improved sensors are developed to measure rotational motions, they may become a new observable for (engineering) seismology, potentially adding complementary information on earthquake source processes, Earth structure,

and ground shaking [98]. For deciphering ground-motion complexity, rotational seismology constitutes a new emerging research field; consequently, the International Working Group on Rotational Seismology (IWGoRS) was formed in 2006 to foster the exchange of ideas, data, and software on rotational seismology.

As strong-motion arrays become more abundant and better equipped, the recorded near-source seismograms provide a wealth of information on the rupture process, the wave-propagation phenomena and the site structure. Harvesting these data to learn more about these complex physical processes will be a research focus for years to come. For example, current earthquake source inversions generally use low-pass filtered seismograms ( $f \leq 1$  Hz) because of incomplete knowledge of Earth structure at shorter wavelengths and the increasing ill-conditioning of the inverse problem as higher frequencies are included. This limitation does not allow proper imaging of high-frequency radiation on the fault, excited at small-scale geometrical complexities or by sudden changes in the dynamic rupture process. Those higher frequencies, however, are particularly damaging to much of the built environment and critical for reliable seismic hazard studies.

Currently, seismologists either estimate earthquake source properties and assume an Earth model, or simplify the source to a known mechanism to understand Earth structure. Future work will increasingly consider earthquake source modeling and imaging of Earth structure as a coupled (joint) inverse problem, potentially including shorter wavelength and/or deriving appropriate stochastic media characterizations. With the advent of high-performance computing facilities and innovative numerical methods (see ► [Seismic Wave Propagation in Media with Complex Geometries, Simulation of](#)) [61,112,149], multi-level optimization strategies will be developed to solve the highly non-linear inverse problem of inferring earthquake rupture dynamics from radiated seismic waves that propagate through heterogeneous media.

Empirical ground-motion equations remain important in engineering practice, but accurately quantifying ground-shaking variability for GMPE's continues to be difficult, despite an increasing number of near-source recordings. It is likely that each future large earthquake, recorded by a dense network, may generate unusual ground-motions outside the commonly assumed standard-deviation of current empirical predictions. Precise and reliable ground-motion estimation is the key challenge for seismologists and earthquake engineers for proper seismic hazard assessment and earthquake loss mitigation in future events. This task requires the ability to model not only median values of ground-mo-

tion intensity, but also to accurately capture their variability. Moreover, it will become increasingly important to not only model scalar ground-motion intensities, but to compute large suites of realistic synthetic near-field seismograms that reproduce the observed ground-motion complexity. Thus, innovative source-modeling approaches are required to capture the complexity of the dynamic rupture process (e.g. [77,120]). Multi-scale dynamic simulations with initial conditions based on stochastic parameter distributions [133,151], need to be coupled to large-scale broadband wave-propagation computations (e.g. [138,139]) for many realizations of heterogeneous Earth models (perhaps with stochastic properties at short wave-lengths). Such simulation-based ground-motion prediction is needed to advance the current practice in seismic hazard assessment.

## Acknowledgments

I am indebted to J. Bühler for generating several figures for this article. J. Ripperger provided simulation data for Fig. 5. Thanks to P. Spudich for computing isochrone quantities and synthetics shown in Fig. 11. Strong-motion data were taken from the COSMOS strong-motion database (<http://db.cosmos-eq.org>). Critical comments and helpful suggestions by J. Clinton, G. Cua, and S. Jonsson greatly improved the article. I am also grateful for constructive reviews by R. Harris and Y. Ben-Zion. Parts of this work was supported by the Southern California Earthquake Center; SCEC is funded by NSF Cooperative Agreement EAR-0106924 and USGS Cooperative Agreement 02HQAG0008. This is SCEC contribution number 1154.

## Bibliography

### Primary Literature

1. Aagaard B, Heaton TH (2004) Near-source ground motions from simulations of sustained intersonic and supersonic fault ruptures. *Bull Seis Soc Am* 94(6):2064–2078
2. Abrahamson NA, Silva WJ (2005) Preliminary results of the A&S 2005 attenuation relation, presented at the USGS workshop on Attenuation Relations to be used in the National USGS Seismic Hazard Maps, Menlo Park CA, October 24
3. Abrahamson NA, Somerville PG (1996) Effects of the hanging-wall and foot-wall on ground motions recorded during the Northridge earthquake. *Bull Seis Soc Am* 86:93–99
4. Abrahamson NA, Shedlock KM (1997) Overview. *Seis Res Lett* 68(1):9–23
5. Abrahamson NA, Silva W (1997) Empirical response spectral attenuation relations for shallow crustal earthquakes. *Seis Res Lett* 68(1):94–127
6. Aki K, Richards PG (2002) *Quantitative Seismology*. University Science Books, Sausalito

7. Akkar S, Bommer JJ (2007) Empirical prediction equations for peak ground velocity derived from strong-motion records from Europe and the Middle East. *Bull Seis Soc Am* 97(2):511–530
8. Ambraseys NN (1995) The prediction of earthquake peak ground acceleration in Europe. *Earthq Eng Struct Dyn* 24(4):467–490
9. Anderson JG (2002) Strong-motion seismology. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part B*. Academic Press, San Diego, 937–965
10. Andrews DJ (1976) Rupture propagation with finite stress in antiplane strain. *J Geophys Res* 81(20):3575–3582
11. Andrews DJ (1980) A stochastic fault model; 1. Static case. *J Geophys Res* 85(B7):3867–3877
12. Andrews DJ (1981) A stochastic fault model; 2. Time-dependent case. *J Geophys Res* 86(11):10821–10834
13. Andrews DJ, Ben-Zion Y (1997) Wrinkle-like slip pulse on a fault between different materials. *J Geophys Res* 102(1):552–571
14. Aochi H, Fukuyama E (2002) Three-dimensional non-planar simulation of the 1992 Landers earthquake. *J Geophys Res* 107(B2):art 2035. doi:10.1029/2000JB000061
15. Aochi H, Madariaga R (2003) The 1999 Izmit, Turkey, earthquake: Non-planar fault structure, dynamic rupture process, and strong ground motion. *Bull Seis Soc Am* 93(3):1249–1266
16. Archuleta RJ (1984) A faulting model for the 1979 Imperial Valley earthquake. *J Geophys Res* 89(6):4559–4585
17. Atkinson G, Boore DM (1995) New ground motion relations for eastern North America. *Bull Seis Soc Am* 85:17–30
18. Bay F, Fäh D, Malagnini L, Giardini D (2003) Spectral shear-wave ground-motion scaling in Switzerland. *Bull Seis Soc Am* 93(1):414–429
19. Bell JW, Brune JN, Zeng Y (2004) Methodology for obtaining constraints of ground motion from precariously balanced rocks. *Bull Seis Soc Am* 94:285–303
20. Ben-Zion Y (2003) Appendix 2, Key Formulas in Earthquake Seismology. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part B*. Academic Press, San Diego, pp 1857–1875
21. Ben-Zion Y, Andrews DJ (1998) Properties and implications of dynamic rupture along a material interface. *Bull Seis Soc Am* 88(4):1085–1094
22. Ben-Zion Y, Sammis CG (2003) Characterization of Fault Zones. *Pure Appl Geophys* 160:677–715
23. Ben-Zion Y, Zhu L (2002) Potency-magnitude scaling relations for southern California earthquakes with  $1.0 < M_L < 7.0$ . *J Geophys Res* 107:F1–F5
24. Benites R, Olsen KB (2005) Modeling strong ground motion in the Wellington Metropolitan Area, New Zealand. *Bull Seis Soc Am* 95(6):2180–2196
25. Beresnev IA (2003) Uncertainties in finite-fault slip inversions: To what extent to believe? (A critical review). *Bull Seis Soc Am* 93(6):2445–2458
26. Beresnev IA, Wen KL (1996) Nonlinear soil response – A reality. *Bull Seis Soc Am* 86:1964–1978
27. Bernard P, Madariaga R (1984) A new asymptotic method for the modeling of near field accelerograms. *Bull Seis Soc Am* 74:539–558
28. Beroza GC, Mikumo T (1996) Short slip duration in dynamic rupture in the presence of heterogeneous fault properties. *J Geophys Res* 101(10):22,449–22,460
29. Beroza GC, Spudich P (1988) Linearized Inversion for Fault Rupture Behavior. Application to the 1984 Morgan-Hill, California, Earthquake. *J Geophys Res* 93(B6):6275–6296
30. Beyer K, Bommer JJ (2006) Relationships between median values and between aleatory variabilities for different definitions of the horizontal component of motion. *Bull Seis Soc Am* 96(4):1512–1522
31. Bindi D, Parolai S, Grosser H, Milkereit C, Durukal E (2007) Empirical ground-motion prediction equations for northwestern Turkey using the aftershocks of the 1999 Kocaeli earthquake. *Geophys Res Lett* 34:L08305. doi:10.1029/2007GL029222
32. Bolt BA, Abrahamson NA (2002) Estimation of strong seismic ground motions. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part B*. Academic Press, San Diego, 983–1001
33. Bommer JJ, Abrahamson NA, Strasser FO, Pecker A, Bard P-Y, Bungum H, Cotton F, Fäh D, Sabetta F, Scherbaum F, Studer J (2004) The challenge of defining upper bounds on earthquake ground motions. *Seis Res Lett* 75(1):82–95
34. Bommer JJ, Martinez-Pereira A (1998) The effective duration of earthquake strong motion. *J Earthq Eng* 3(2):127–172
35. Boore DM (1983) Stochastic simulation of high frequency ground motions based on seismological models of the radiation spectra. *Bull Seismol Soc Am* 73:1865–1894
36. Boore DM (2001) Comparisons of ground motions from the 1999 Chi-Chi earthquake with empirical predictions largely based on data from California. *Bull Seis Soc Am* 91(5):1212–1217
37. Boore DM, Atkinson GM (2008) Updated Reports of Next Generation Attenuation (NGA) Models, PEER Lifelines Program, report and auxiliary material available at <http://peer.berkeley.edu/products/Boore-Atkinson-NGA.html>
38. Boore DM, Joyner WB et al (1997) Equations for estimating horizontal response spectra and peak acceleration from western North American earthquakes; a summary of recent work. *Seis Res Lett* 68 (1):128–153
39. Boore DM, Lamprey JW, Abrahamson NA (2006) Orientation-independent measures of ground motion. *Bull Seism Soc Am* 96(4A):1502–1511. doi:10.1785/0120050209
40. Borcherdt RD (1994) Estimates of site-dependent response spectra for design (methodology and justification). *Earthquake Spectra* 10(4):617–653
41. Borcherdt RD (2002) Empirical evidence for acceleration-dependent amplification factors. *Bull Seis Soc Am* 92(2):761–782
42. Bouchon M, Aki K (1982) Strain, tilt, and rotation associated with strong ground motion in the vicinity of earthquake faults. *Bull Seismol Soc Am* 72:1717–1738
43. Bouchon M, Streiff D (1997) Propagation of a shear crack on a nonplanar fault: A method of calculation. *Bull Seis Soc Am* 87(1):61–66
44. Bouchon M, Vallee M (2003) Observation of long supershear rupture during the magnitude 8.1 Kunlunshan earthquake. *Science* 301(5634):824–826
45. Brune JN (1970) Tectonic Stress and Spectra of Seismic Shear Waves from Earthquakes. *J Geophys Res* 75(26):4997–5009
46. Brune JN (1996) Precariously balanced rocks and ground-motion maps for southern California. *Bull Seis Soc Am* 86(1):43–54



47. Brune JN (1999) Precarious rocks along the Mojave section of the San Andreas Fault, California: Constraints on ground motion from great earthquakes. *Seis Res Lett* 70:29–33
48. Brune JN (2003) Precarious rock evidence for low near-source accelerations for trans-tensional strike-slip earthquakes. *Physics of the Earth and Planetary Interiors* 137(1–4):229–239
49. Brune JN, Anooshehpour A, Purvance MD (2006) Band of precariously balanced blocks between the Elsinore and San Jacinto, California, fault zones: Constraints on ground motion for large earthquakes. *Geology* 34:137–140
50. Brune JN, Whitney JW (1992) Precariously balanced rocks with rock varnish—Paleoindicators of maximum ground acceleration? *Seis Res Lett* 63:21
51. Cakir Z, de Chabaliér JB, Armijo R, Meyer B, Barka A, Peltzer G (2003) Coseismic and early post-tensional slip associated with the 1999 Izmit earthquake (Turkey), from SAR interferometry and tectonic field observations. *Geophys J Int* 155 (1):93–110
52. Campbell KW (2002) Strong-motion attenuation relations. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part B*. Academic Press, San Diego, pp 1003–1012
53. Campbell KW, Bozorgnia Y (2003) Updated near-source ground-motion (attenuation) relations for the horizontal and vertical components of peak ground acceleration and acceleration response spectra. *Bull Seis Soc Am* 93(3):1413–1413
54. Choi Y, Stewart JP (2005) Nonlinear site amplification as function of 30 m shear-wave velocity. *Earthq Spect* 21(1):1–30
55. Clinton JF (2004) Modern digital seismology-instrumentation, and small amplitude studies in the engineering world. <http://resolver.caltech.edu/CaltechETD:etd-05202004-225044>
56. Custodio S, Lui P-C, Archuleta RJ (2005) The 2004  $M_w$  6.0 Parkfield, California, earthquake: inversion of near-source ground motion using multiple data sets. *Geophys Res Lett* 32:L23312
57. Davis P, Rubinstein JL, Liu K, Gao S, Knopoff L (2000) Northridge earthquake damage caused by geologic focusing of seismic waves. *Science* 289(5485):1746–1750
58. Day SM (1982a) Three-dimensional finite difference simulation of fault dynamics; rectangular faults with fixed rupture velocity. *Bull Seis Soc Am* 72(3):705–727
59. Day SM (1982b) 3-Dimensional Simulation of Spontaneous Rupture – the Effect of Non-uniform Prestress. *Bull Seis Soc Am* 72(6):1881–1902
60. Douglas J (2003) Earthquake ground motion estimation using strong-motion records: a review of equations for the estimation of peak ground acceleration and response spectral ordinates. *Earth-Science Reviews* 61:43–104
61. Dumbser M, Käser M (2006) An arbitrary high-order discontinuous Galerkin method for elastic waves on unstructured meshes II— The three dimensional case. *Geophys J Int* 167:319–336
62. Dunham E, Archuleta RJ (2004) Evidence for a super-shear transient during the 2002 Denali earthquake. *Bull Seis Soc Am* 94:5256–5268
63. Dunham E, Favreau P, Carlson J (2003) A supershear transition mechanism for cracks. *Science* 299:1557–1559
64. Field EH, Zeng YH, Johnson PA, Beresnev I (1998) Nonlinear sediment response during the 1994 Northridge earthquake: Observations and finite source simulations. *J Geophys Res* 103(B11):26869–26883
65. Frankel A, Clayton RW (1984) A finite-difference simulation of wave propagation in two-dimensional random media. *Bull Seis Soc Am* 74(6):2167–2186
66. Frankel A, Clayton RW (1986) Finite-Difference Simulations of Seismic Scattering Implications— for the Propagation of Short-Period Seismic-Waves in the Crust and Models of Crustal Heterogeneity. *J Geophys Res* 91(B6):6465–6489
67. Fukushima Y, Tanaka T (1990) A new attenuation relation for peak horizontal acceleration of strong earthquake ground motion in Japan. *Bull Seis Soc Am* 80(4):757–783
68. Fukuyama E, Olsen KB (2002) A condition for super-shear rupture propagation in a heterogeneous stress field. *Pure Appl Geophys* 159(9):2047–2056
69. Fäh D, Kind F, Giardini D (2001) A theoretical investigation of average H/V ratios. *Geophys J Int* 145(2):535–549
70. Fäh D, Rüttener E, Noack T, Kruspan P (1997) Microzonation of the city of Basel. *J of Seism* 1:87–102
71. Goff JA, Holliger K (1999) Nature and origin of upper crustal seismic velocity fluctuations and associated scaling properties: Combined stochastic analyses of KTB velocity and lithology logs. *J Geophys Res* 104(B6):13169–13182
72. Graves RW (1996) Simulating seismic wave propagation in 3D elastic media using staggered-grid finite differences. *Bull Seis Soc Am* 86(4):1091–1106
73. Graves RW (1998) Three-dimensional finite-difference modeling of the San Andreas fault: source parameterization and ground-motion levels. *Bull Seis Soc Am* 88(4):881–897
74. Graves RW, Clayton RW (1992) Modeling path effects in 3-dimensional basin structures. *Bull Seis Soc Am* 82(1):81–103
75. Graves RW, Pitarka A, Somerville P (1998) Ground-motion amplification in the Santa Monica area: Effects of shallow basin-edge structure. *Bull Seis Soc Am* 88(5):1224–1242
76. Graves RW, Wald DJ (2004) Observed and simulated ground motions in the San Bernardino basin region for the Hector Mine, California, earthquake. *Bull Seis Soc Am* 94(1):131–146
77. Guatteri M, Mai PM, Beroza GC (2004) A pseudo-dynamic approximation to dynamic rupture models for strong ground motion prediction. *Bull Seis Soc Am* 94(6):2051–2063
78. Guatteri M, Mai PM, Beroza GC, Boatwright J (2003) Strong-ground motion prediction from stochastic-dynamic source models. *Bull Seis Soc Am* 93(1):301–313
79. Hanks TC, Bakun WH (2002) A bilinear source-scaling model for  $M$ -log  $A$  observations of continental earthquakes. *Bull Seis Soc Am* 92(5):1841–1846
80. Harris RA (2004) Numerical simulations of large earthquakes: dynamic rupture propagation on heterogeneous faults. *Pure Appl Geophys* 161:2171–2181
81. Harris RA, Archuleta RJ et al (1991) Fault steps and the dynamic rupture process; 2-D numerical simulations of a spontaneously propagating shear fracture. *Geophys Res Lett* 18(5):893–896
82. Harris RA, Day SM (1993) Dynamics of fault interaction; parallel strike-slip faults. *J Geophys Res* 98(3):4461–4472
83. Harris RA, Day SM (1997) Effects of a low-velocity zone on a dynamic rupture. *Bull Seis Soc Am* 87(5):1267–1280
84. Harris RA, Day SM (1999) Dynamic 3D simulations of earthquakes on en echelon faults. *Geophys Res Lett* 26(14):2089–2092
85. Hartzell S, Guatteri M, Mai PM, Liu P-C, Fiske M (2005) Calculation of broadband time histories of ground motion: Part

- II, Kinematic and dynamic modeling with theoretical Green's functions. *Bull Seis Soc Am* 95(2):614–645
86. Hartzell S, Liu P-C, Mendoza C, Ji C, Larson K (2007) Stability and uncertainty of finite-fault slip inversions: Application to the 2004 Parkfield, California, earthquake. *Bull Seis Soc Am* 97:1911–1934
  87. Heaton TH (1982) The 1971 San-Fernando Earthquake – a Double Event? *Bull Seis Soc Am* 72(6):2037–2062
  88. Heaton TH (1990) Evidence for and implications of self-healing pulses of slip in earthquake rupture. *Phys Earth Planet Int* 64:1–20
  89. Hikima K, Koketsu K (2004) Source processes of the foreshock, mainshock and largest aftershock in the 2003 Miyagi-ken Hokubu, Japan, earthquake sequence. *Earth Planets Space* 56:87–93
  90. Hillers G, Mai PM, Ampuero J-P, Ben-Zion Y (2007) Statistical properties of seismicity of fault zones at different evolutionary stages. *Geophys J Int* 169:515–533. doi:10.1111/j.1365-246X.2006.03275.x
  91. Hillers G, Wesnousky S (2008) Scaling relation of strike-slip earthquakes with different rate-state-dependent properties at depth. *Bull Seis Soc Am* 98(3):1085–1101
  92. Hisada Y (2001) A theoretical omega-square model considering spatial variation in slip and rupture velocity. Part 2: Case for a two-dimensional source model. *Bull Seis Soc Am* 91(4):651–666
  93. Holliger K (1997) Seismic scattering in the upper crystalline crust based on evidence from sonic logs. *Geophys J Int* 128(1):65–72
  94. Horikawa H (2001) Earthquake doublet in Kagoshima, Japan: Rupture of asperities in a stress shadow. *Bull Seis Soc Am* 91(1):112–127
  95. Huang BS (2003) Ground rotational motions of the 1991 Chi-Chi, Taiwan earthquake as inferred from dense array observations. *Geophys Res Lett* 30(6):1307–1310
  96. Ide S (2001) Complex source processes and the interaction of moderate earthquakes during the earthquake swarm in the Hida-Mountains, Japan, 1998. *Tectonophysics* 334(1):35–54
  97. Igel H, Cochard A, Wassermann J, Schreiber U, Velikoseltsev A, Dinh NP (2007) Broadband observations of rotational ground motions. *Geophys J Int* 168(1):182–197
  98. Igel H, Schreiber U, Flaws A, Schuberth B, Velikoseltsev A, Cochard A (2005) Rotational motions induced by the M8.1 Tokachi-oki earthquake, September 25, 2003. *Geophys Res Lett* 32:L08309. doi:10.1029/2004GL022336
  99. Irikura K, Kamae K (1994) Estimation of strong ground motion in broad-frequency band based on a seismic source scaling model and an empirical Green's function technique. *Annali Di Geofisica* 37:1721–1743
  100. Iwata T, Sekiguchi H (2002) Source process of the 2000 western Tottori Prefecture earthquake and near-source strong ground motion. *Proceedings of the Japan. Earthquake Eng. Symposium*, vol 11. pp 125–128
  101. Johnson KM, Hsu YJ, Segall P, Yu SB (2001) Fault geometry and slip distribution of the 1999 Chi-Chi, Taiwan earthquake imaged from inversion of GPS data. *Geophys Res Lett* 28(11):2285–2288
  102. Jonsson S, Zebker H, Segall P, Amelung F (2002) Fault slip distribution of the 1999 M-w 7.1 Hector Mine, California, earthquake, estimated from satellite radar and GPS measurements. *Bull Seis Soc Am* 92(4):1377–1389
  103. Joyner WB, Boore DM (1988) Measurement, characterization, and prediction of strong ground motion. *Conference on Earthquake Engineering and Soil Dynamics II: Recent Advances in Ground Motion Evaluation*, Park ASCE City UT, USA, p 43–102
  104. Joyner WB, Spudich P (1994) Including near-field terms in the isochrones integration method for application to finite-fault or Kirchhoff boundary integral problems. *Bull Seis Soc Am* 84:1260–1265
  105. Kagawa T, Irikura K, Somerville PG (2004) Differences in ground motion and fault rupture process between surface and buried rupture earthquakes. *Earth Planets Space* 56(1):3–14
  106. Kanamori H, Anderson DL (1975) Theoretical basis of some empirical relations in seismology. *Bull Seis Soc Am* 65(5):1073–1095
  107. Kawase H (1996) The cause of the damage belt in Kobe: The basin edge effect – constructive interference of the direct S-wave with the basin-induced diffracted Rayleigh waves. *Seis Res Lett* 67(5):25–34
  108. Kawase H (2003) Site effects on strong ground motions In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology*, Part B. Academic Press, San Diego, pp 1013–1030
  109. Koketsu K, Kikuchi M (2000) Propagation of seismic ground motion in the Kanto basin, Japan. *Science* 288:1237–1239
  110. Kostrov BV (1964) Self-similar problems of propagating shear cracks. *App J Math Mech* 28:1077–1078
  111. Kramer SL (1996) *Geotechnical Earthquake Engineering*. Prentice Hall, Upper Saddle River, New Jersey
  112. Käser M, Mai PM, Dumbser M (2007) Geometrically complex finite source rupture models and their accurate treatment using tetrahedral meshes. *Bull Seis Soc Am* 97(5):1570–1586
  113. Lavallee D, Archuleta R (2003) Stochastic modeling of slip spatial complexities for the 1979 Imperial Valley, California, earthquake. *Geophys Res Lett* 30:1245. doi:10.1029/2002GL015839
  114. Lavallee D, Liu PC, Archuleta R (2006) Stochastic model of heterogeneity in earthquake slip spatial distributions. *Geophys J Int* 165(2):622–640
  115. Lee WHK, Shin T-C, Kuo KW, Chen KC, Wu CF (2001) CWB free-field strong-motion data from the 21 September Chi-Chi, Taiwan, earthquake. *Bull Seis Soc Am* 91(5):1370–1376
  116. Liu P, Archuleta RJ (2004) A new nonlinear finite fault inversion with three-dimensional Green's functions: Application to the 1989 Loma Prieta, California, earthquake. *J Geophys Res* 109:B02318
  117. Liu P-C, Custodio S, Archuleta RJ (2006) Kinematic inversion of the 2004 M 6.0 Parkfield earthquake including an approximation to site effects. *Bull Seis Soc Am* 96:143–158
  118. Ma S, Archuleta RJ, Page MT (2007) Effects of large-scale surface topography on ground motions, as demonstrated by a study of the San Gabriel Mountains, Los Angeles, California. *Bull Seis Soc Am* 97(6):2066–2079
  119. Madariaga R (1983) High frequency radiation from dynamic earthquake fault models. *Ann Geophys* 1(1):17–23
  120. Mai PM (2001) Characterizing earthquake source complexity for improved strong motion prediction. Ph D thesis, Department of Geophysics, Stanford University, California
  121. Mai PM (2004) SRCMOD: an online database of finite-

- source rupture models. <http://www.seismo.ethz.ch/srcmod>, updated July 2007; last accessed Aug 2007
122. Mai PM, Beroza GC (2000) Source-scaling properties from finite-fault rupture models. *Bull Seis Soc Am* 90(3):604–615
  123. Mai PM, Beroza GC (2002) A spatial random-field model to characterize complexity in earthquake slip. *J Geophys Res* 107(B11):2308. doi:10.1029/2001JB000588
  124. Mai PM, Beroza GC (2003) A hybrid method for calculating near-source broadband seismograms: application to strong motion prediction. *Phys Earth Planet Int* 137:183–199
  125. Mai PM, Olsen KB (2005) Broadband ground motion simulations using finite-difference synthetics with local scattering operators. 2005 Annual SCEC Meeting, Palm Springs CA, Southern California Earthquake Center
  126. Mai PM, Somerville P, Pitarka A, Dalguer L, Miyake H, Beroza G, Song S-G, Irikura K (2006) Fracture-energy scaling in dynamic rupture models of past earthquakes. *Earthquakes: Radiated Energy and the Physics of Faulting Geophysical Monograph Series*, vol. 170. American Geophysical Union, 10.1029/170GM28:283–294
  127. Mai PM, Spudich P, Boatwright J (2005) Hypocenter locations in finite-source rupture models. *Bull Seis Soc Am* 95(3):965–980
  128. Monelli D, Mai PM (2008) Bayesian inference of kinematic earthquake rupture parameters through fitting of strong motion data. *Geophys J Int* 173:220–232. doi:10.1111/j.1365-246X.2008.03733.x
  129. Nakamura H, Miyatake T (2000) An approximate expression of slip velocity time functions for simulation of near-field strong ground motion. *Zishin (J Seis Soc Jpn)* 53:1–9
  130. O'Connell DRH (1999) Replication of apparent nonlinear seismic response with linear wave propagation models. *Science* 283(5410):2045–2050
  131. Oglesby DD, Archuleta RJ (2003) The three-dimensional dynamics of a nonplanar thrustfault. *Bull Seis Soc Am* 93(5):2222–2235
  132. Oglesby DD, Day SM (2001) Fault geometry and the dynamics of the 1999 Chi-Chi (Taiwan) earthquake. *Bull Seis Soc Am* 91(5):1099–1111
  133. Oglesby DD, Day SM (2002) Stochastic fault stress: Implications for fault dynamics and ground motion. *Bull Seis Soc Am* 92(8):3006–3021
  134. Oglesby DD, Dreger DS, Harris RA, Ratchkovski N, Hansen R (2004) Inverse kinematic and forward dynamic models of the 2002 Denali fault earthquake, Alaska. *Bull Seis Soc Am* 94(6):S214–S233
  135. Olsen KB (2000) Site amplification in the Los Angeles basin from three-dimensional modeling of ground motion. *Bull Seis Soc Am* 90(6):77–94
  136. Olsen KB, Archuleta RJ (1996) Three-dimensional simulation of earthquakes on the Los Angeles fault system. *Bull Seis Soc Am* 86(3):575–596
  137. Olsen KB, Archuleta RJ, Matarese JR (1995) 3-Dimensional simulation of a magnitude-7.75 Earthquake on the San-Andreas fault. *Science* 270(5242):1628–1632
  138. Olsen KB, Day SM, Minster JB, Cui Y, Chourasia A, Faerman M, Moore R, Maechling P, Jordan T (2006) Strong shaking in Los Angeles expected from southern San Andreas earthquake. *Geophys Res Lett* 33:L07305. doi:10.1029/2005GL025472
  139. Oprsal I, Fäh D, Mai PM, Giardini D (2005) Deterministic earthquake scenario for the Basel area Simulating – strong motion and site effects for Basel (Switzerland). *J Geophys Res* 110:B04305. doi:10.1029/2004JB003188
  140. PEER (2008) Next Generation Attenuation of Ground Motions (NGA) Project. [http://peer.berkeley.edu/products/nga\\_project.html](http://peer.berkeley.edu/products/nga_project.html)
  141. Peng ZG, Ben-Zion Y (2006) Temporal changes of shallow seismic velocity around the Karadere–Duzce branch of the North Anatolian Fault and strong ground motion. *Pure Appl Geophys* 163:567–600
  142. Piatanesi A, Cirella A, Spudich P, Cocco M (2007) A global search inversion for earthquake kinematic rupture history: application to the 2000 western Tottori, Japan earthquake. *J Geophys Res* B07314. doi:10.1029/2006JB004821
  143. Pitarka A, Irikura K (1996) Modeling 3D surface topography by finite-difference method; Kobe-JMA station site, Japan, case study. *Geophys Res Lett* 23(20):2729–2732
  144. Pitarka A, Irikura K, Iwata T, Kagawa T (1996) Basin structure effects in the Kobe area inferred from the modeling of ground motions from two aftershocks of the January 17, 1995, Hyogoken Nanbu earthquake. *J Phys Earth* 44(5):563–576
  145. Pitarka A, Irikura K, Iwata T, Sekiguchi H (1998) Three-dimensional simulation of the near-fault ground motion for the 1995 Hyogo-ken Nanbu (Kobe), Japan, Earthquake. *Bull Seis Soc Am* 88(2):428–440
  146. Pitarka A, Somerville P, Fukushima Y, Uetake T, Irikura K (2000) Simulation of near-fault strong-ground motion using hybrid Green's function. *Bull Seis Soc Am* 90(3):566–586
  147. Pulido N, Kubo T (2004) Near-fault strong motion complexity of the 2000 Tottori earthquake (Japan) from a broadband source asperity model. *Tectonophysics* 390:177–192
  148. Reid HF (1910) *The Mechanics of the Earthquake: The California Earthquake of April 18, 1906*, Report of the State Investigation Commission, Vol. 2. Carnegie Institution of Washington, Washington DC
  149. Ripperger J, Mai PM (2004) Fast computation of static stress changes on 2D faults from final slip distributions. *Geophys Res Lett* 31(18):L18610. doi:10.1029/2004GL020594
  150. Ripperger J, Mai PM, Ampuero J-P (2008) Near-Field Ground Motion from Dynamic Earthquake Rupture Simulations. *Bull Seis Soc Am* 98(3):1207–1228
  151. Ripperger J, Mai PM, Ampuero J-P, Giardini D (2007) Earthquake source characteristics from dynamic rupture with constrained stochastic fault stress. *J Geophys Res* 112:B04311. doi:10.1029/2006JB004515
  152. Ritter JRR, Mai PM, Stoll G, Fuchs K (1997) Scattering of teleseismic waves in the lower crust Observations – in Massif Central, France. *Phys Earth Planet Int* 104:127–146
  153. Robinson DP, Brough C et al (2006) The M-w 7.8:2001 Kunlunshan earthquake: Extreme rupture speed variability and effect of fault geometry. *J Geophys Res* B08303. doi:10.1029/2005JB004137
  154. Romanowicz B (1992) Strike-slip earthquakes on quasi-vertical transcurrent faults; inferences for general scaling relations. *Geophys Res Lett* 19(5):481–484
  155. Rosakis AJ, Samudrala O, Coker D (1999) Cracks faster than the shear wave speed. *Science* 284(5418):1337–1340
  156. Rubinstein JL, Beroza GC (2004) Evidence for widespread nonlinear strong ground motion in the M-W 6.9 Loma Prieta Earthquake. *Bull Seis Soc Am* 94(5):1595–1608

157. Sadigh K, Chang CY, Egan JA, Makdisi F, Youngs RR (1997) Attenuation relationships for shallow crustal earthquakes based on California strong motion data. *Seis Res Lett* 68(1):180–189
158. Sato H, Fehler M (1998) Seismic wave propagation and scattering in the heterogeneous Earth, Press AIP/Springer, New York
159. Sato H, Fehler M, Wu R-S (2003) Scattering and attenuation of seismic waves in the lithosphere. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part A*. Academic Press, San Diego, p 195–208
160. Sato T, Graves RW, Somerville PG (1999) Three-dimensional finite-difference simulations of long-period strong motions in the Tokyo metropolitan area during the 1990 Odawara earthquake (MJ 5.1) and the great 1923 Kanto earthquake (MS 8.2) in Japan. *Bull Seis Soc Am* 89:579–607
161. Sato T, Graves RW, Somerville PG, Kataoka S (1998) Estimates of regional and local strong motions during the great 1923 Kanto, Japan, earthquake (Ms 8.2). Part 2: Forward simulation of seismograms using variable-slip rupture models and estimation of near-fault long-period ground motions. *Bull Seis Soc Am* 88(1):206–227
162. Sato T, Helmberger DV, Somerville PG, Graves RW, Saikia CK (1998) Estimates of regional and local strong motions during the great 1923 Kanto, Japan, earthquake (Ms 8.2). Part 1: Source estimation of a calibration event and modeling of wave propagation paths. *Bull Seis Soc Am* 88(1):183–205
163. Scherbaum F, Schmedes J, Cotton F (2004) On the conversion of source-to-site distance measures for extended earthquake Source models. *Bull Seis Soc Am* 94(3):1053–1069
164. Scholz C (2002) *The mechanics of earthquakes and faulting*, 2nd edn. Cambridge University Press, Cambridge
165. Sekiguchi H, Irikura K, Iwata T (2002) Source inversion for estimating the continuous slip distribution on a fault-introduction of Green's functions convolved with a correction function to give moving dislocation effects in subfaults. *Geophys J Int* 150(2):377–391
166. Shakal A, Haddadi H, Graizer V, Lin K, Huang M (2006) Some key features of the strong-motion data from the M 6.0 Parkfield, California, earthquake of 28 September 2004. *Bull Seis Soc Am* 96(4B):S90–S118
167. Shakal AF, Haddadi HR, Huang MJ (2006) Note on the Very-High-Acceleration Fault Zone 16 Record from the 2004 Parkfield Earthquake. *Bull Seis Soc Am* 96(3):S119–S128
168. Sigbjörnsson R, Ambraseys NN (2003) Uncertainty analysis of strong-motion and seismic hazard. *Bull Earthq Eng* 1:321–347
169. Somerville P, Irikura K, Graves R, Sawada S, Wald DJ, Abrahamson N, Iwasaki Y, Kagawa T, Smith N, Kowada A (1999) Characterizing crustal earthquake slip models for the prediction of strong ground motion. *Seis Res Lett* 70(1):59–80
170. Somerville PG (2003) Magnitude scaling of the near fault rupture directivity pulse. *Phys Earth Planet Int* 37:201–212
171. Somerville PG, Smith NF, Graves RW, Abrahamson NA (1997) Modification of empirical strong ground motion attenuation relations to include the amplitude and duration effects of rupture directivity. *Seis Res Lett* 68(1):199–222
172. Spudich P, Chiou BSJ, Graves R, Collins N, Somerville P (2004) A formulation of directivity for earthquake sources using isochrone theory. United States Geological Survey, Open-File Report 2004–1268
173. Spudich P, Frazer LN (1984) Use of ray theory to calculate high-frequency radiation from earthquake sources having spatially variable rupture velocity and stress drop. *Bull Seis Soc Am* 74(6):2061–2082
174. Spudich P, Hellweg M, Lee WHK (1996) Directional topographic site response at Tarzana observed in aftershocks of the 1994 Northridge, California, earthquake; implications for mainshock motions. *Bull Seis Soc Am* 86(1, Part Suppl B):193–208
175. Spudich P, Joyner WB, Lindh AG, Boore DM, Margaris BM, Fletcher JB (1999) SEA99: a revised ground motion prediction relation for use in extensional tectonic regimes. *Bull Seis Soc Am* 89(1):1156–1170
176. Spudich P, Xu L (2003) Software for calculating earthquake ground motions from finite faults in vertically varying media In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part B*. Academic Press, San Diego, pp 1857–1875
177. Strasser FO, Bommer JJ, Abrahamson NA (2008) Truncation of the distribution of ground-motion residuals. *J Seis* 12:79–105
178. Sørensen MB, Oprsal I, Bonnefoy SC, Atakan K, Mai PM, Pulido N, Yalciner C (2006) Local site effects in Ataköy, Istanbul, Turkey, due to a future large earthquake in the Marmara Sea. *Geophys J Int* 167(3):1413–1424
179. Takeo M (1998) Ground rotational motions recorded in near-source region of earthquakes. *Geophys Res Lett* 25(6):789–792
180. Tinti E, Fukuyama E, Piatenesi A, Cocco M (2005) A kinematic source-time function compatible with earthquake dynamics. *Bull Seis Soc Am* 95(4):1211–1223
181. Tinti E, Spudich P, Cocco M (2005) Earthquake fracture energy inferred from kinematic rupture models on extended faults. *J Geophys Res* 110:B12303. doi:10.1029/2005JB003644
182. Trifunac MD (1982) A note on rotational components of earthquake motions on ground surface for incident body waves. *Soil Dyn Earthq Eng* 1:11–19
183. Trifunac MD (2006) Effects of torsional and rocking excitations on the response of structures. In: Teisseyre R et al (eds) *Earthquake Source Asymmetry, Structural Media and Rotation Effects*. Springer, Heidelberg, pp 569–582
184. Wald DJ, Heaton TH (1994) Spatial and temporal distribution of slip for the 1992 Landers, California, earthquake. *Bull Seis Soc Am* 84(3):668–691
185. Wald DJ, Heaton TH, Hudnut KW (1996) The slip history of the 1994 Northridge, California, earthquake determined from strong-motion, teleseismic, GPS, and leveling data. *Bull Seis Soc Am* 86(1):S49–S70
186. Wang JH, Ou SS (1998) On scaling of earthquake faults. *Bull Seis Soc Am* 88(2):758–766
187. Wells DL, Coppersmith KJ (1994) New empirical relationships among magnitude, rupture length, rupture width, rupture area, and surface displacement. *Bull Seis Soc Am* 84(4):974–1002
188. Working Group on California Earthquake Probabilities (1999) Earthquake probabilities in the San Francisco Bay region: 2000 to 2030—A summary of findings. Open USGS-File Report 99–517
189. Working Group on California Earthquake Probabilities (2003) Earthquake probabilities in the San Francisco Bay region: 2002–2031. Open USGS-File Report 03–214

190. Wu RS, Aki K (1985) The fractal nature of the inhomogeneities in the lithosphere evidenced from seismic-wave scattering. *Pure Appl Geophy* 123(6):805–818
191. Wu Y-M, Shin T-C, Chang C-H (2001) Near real-time mapping of peak ground acceleration and peak ground velocity following a strong earthquake. *Bull Seis Soc Am* 91(5):1218–1228
192. Zeng YH (1993) Theory of scattered P-wave and S-wave energy in a random isotropic scattering medium. *Bull Seis Soc Am* 83(4):1264–1276
193. Zeng YH, Anderson JG, Su F (1995) Subevent rake and random scattering effects in realistic strong ground motion simulation. *Geophys Res Lett* 22(1):17–20
194. Zeng YH, Su F, Aki K (1991) Scattering wave energy propagation in a random isotropic scattering medium. 1. Theory *J Geophys Res* 96(B1):607–619
- Geophysical Monograph Series, 170. American Geophysical Union, 10.1029/170GM28
- Lay T, Wallace TC (1995) *Modern Global Seismology*. Academic Press, San Diego
- Lee WHK, Kanamori H, Jennings PC, Kisslinger C (2002) *International Handbook of Earthquake and Engineering Seismology, Part A and B*. Academic Press, San Diego

### Web Links

- Database of finite-source rupture models <http://www.seismo.ethz.ch/srcmod>
- COSMOS strong-motion database <http://db.cosmos-eq.org>
- Kyoshin Network (Japan) <http://www.k-net.bosai.go.jp>
- PEER (Pacific Earthquake Engineering Research Center) <http://peer.berkeley.edu>
- NGA Project page at PEER [http://peer.berkeley.edu/products/nga\\_project.html](http://peer.berkeley.edu/products/nga_project.html)
- International Working Group on Rotational Seismology (IWGoRS) <http://www.rotationalseismology.org>

### Books and Reviews

- Abercrombie R, McGarr A, Kanamori H, Di Toro G (eds) (2006) *Earthquakes: Radiated Energy and the Physics of Faulting*.

## Infrasound from Earthquakes, Tsunamis and Volcanoes

MILTON GARCES<sup>1</sup>, ALEXIS LE PICHON<sup>2</sup>

<sup>1</sup> Infrasound Laboratory, HIGP, SOEST, University of Hawaii, Manoa, Kailua-Kona, USA

<sup>2</sup> CEA/DASE/LD, Bruyères-le-Châtel, France

### Article Outline

Glossary

Definition of the Subject

Introduction

Infrasound Arrays

Earthquake Infrasound

Tsunami Infrasound

Volcano Infrasound

Future Directions

Concluding Remarks

Acknowledgments

Bibliography

### Glossary

**Infrasound** atmospheric sound waves with frequencies lower than the 20 Hz hearing threshold of the human ear.

**Infrasound array** four or more horizontally separated identical microphones or microbarometers with precisely known locations that optimize the reception of a specified wavelength range.

**Trace velocity** apparent horizontal phase velocity of an acoustic arrival measured by an array.

**Celerity** effective propagation speed of a signal, measured from the ratio of the total range over the total travel time along the great circle path from a source to a receiver.

**Tremor** volcanic signal consisting of a nearly continuous oscillation of the ground and atmosphere, with durations of minutes to years. Harmonic tremor may have multiple distinct spectral peaks.

**LP** Long Period event. Transient volcanic signal with durations of tens of seconds to minutes and distinct spectral peaks.

### Definition of the Subject

Infrasound may be radiated by earthquakes, tsunamis, and volcanoes through the displacement or rupture of Earth's surface and the subsequent flow and excitation of fluids. These complex and sometimes cataclysmic phenomena share some common physics, yet have different ways of converting energy into atmospheric sound. Sig-

nals from earthquakes, tsunamis, and volcanoes captured by the present generation of infrasound arrays are introduced in this chapter through case studies. Contemporary methods used in the analysis, interpretation, and modeling of these diverse signatures are discussed and some of the associated geophysical problems that remain unsolved are considered.

### Introduction

The human ear may perceive sound in the frequency band of 20 to 20,000 cycles per second (Hz). Infrasound consists of acoustic waves in the atmosphere with frequencies lower than the 20 Hz hearing threshold of the human ear. Because of reduced acoustic attenuation and scattering in the atmosphere at long infrasonic wavelengths and the large spatial scales of the physical processes driving earthquakes, tsunamis, and volcanoes, the infrasound frequency band is well suited to the remote monitoring of these events. The ambient infrasound field at any location is rich and diverse, with sources originating from the solid Earth, the ocean, the atmosphere, space-born objects, and human activity [4,39]. These acoustic pressure waves co-exist with non-acoustic atmospheric pressure fluctuations associated with meteorological changes such as wind and frontal passages (e.g. [9]). Sound propagation paths are controlled primarily by the temperature and wind stratification in the lower, middle, and upper atmosphere [16,19,24]. The effective sound velocity of the atmosphere at a given height above the ground may be approximated as the sum of the scalar sound speed, which is proportional to the square root of temperature, and the vector wind velocity, which typically may reach magnitudes of 15–20% of the sound speed in the upper atmosphere. An acoustic waveguide may efficiently direct sound to ground-based stations, and is defined by a high sound velocity layer at the upper boundary and a lower sound velocity layer near the ground. The high temperature of the mesosphere and lower thermosphere (MLT) would always refract infrasound back to the ground, but severe attenuation above ~110 km can suppress thermospheric returns. Waveguides in the troposphere and stratosphere are expected to only transmit primarily along the downwind direction. However, observations suggest that the elevated acoustic waveguide formed by the low temperature zone in the stratosphere may routinely leak energy back to the ground through scattering and diffraction [27,28,29]. Ground cooling and low altitude winds may also produce a stable waveguide in the boundary layer near the ground surface [20,42]. A new generation of global atmospheric specifications designed

for the study of infrasound propagation has been developed by integrating multiple meteorological and upper atmosphere models [19]. Validation and further refinement of these atmospheric models is ongoing [3,43].

## Infrasound Arrays

### Background

Modern infrasound array technology emerged at the turn of the 21st century after the 1996 adoption of the Comprehensive Nuclear-Test-Ban Treaty and the subsequent growth of the International Monitoring System (IMS), which was designed for the detection of clandestine nuclear test explosions [55,61]. An infrasound array generally consists of four or more horizontally separated identical microphones or microbarometers with precisely known locations that optimize the reception of a specified wavelength range. Wind is the most pernicious source of incoherent noise, and infrasound's greatest vulnerability. An array that is sheltered from the wind by forest cover, snow cover, or topographical blocking will have a low noise floor, and provide high sensitivity and robust measurements. If a wind-sheltered site is not found, wavelength-specific wind noise reducing filters have to be designed for the boundary layer conditions at the array site [1,40]. IMS-type infrasound arrays have a maximum distance between sensors of 1–3 km and use sensors with over 100 dB of dynamic range and a flat frequency response between 0.02 and 20 Hz. Portable arrays often have apertures of  $\sim 100$  m or less, and are thus optimized for smaller wavelengths (frequencies  $>1$  Hz). Infrasonic sensors for portable array applications often operate within the 0.1–100 Hz frequency band, and may overlap into the audio band. Infrasound data are typically recorded with GPS time-stamped 24-bit digitizers and are often sent via digital communications to a central processing facility for real time analysis.

### Basic Principles

A number of calibrated microphones, precisely timed and arranged in optimal spatial configurations as arrays, present the best design for recording infrasound. A wavefront is the spatial surface containing sound emitted from a source at the same time. An array relies on the principle that sound along a (possibly curved) wavefront has recognizable features, so that as the wavefront passes through the multiple sensors in an array it is possible to determine the time of passage of a specific waveform feature (for example, the peak of an explosive pulse). From the time of arrival of a waveform feature at each known sensor location, the direction of propagation of the incident wavefront as

well as its apparent propagation speed across the array can be inferred. Once the arrival direction and speed are determined, it is possible to digitally apply time-delays or phase shifts to temporally align all the microphone waveforms along a beam of energy (beamform) to improve the ratio of the signal amplitude to the ambient noise. Based on these fundamental principles, more advanced contemporary techniques permit the extraction of waveform features with a very small signal to noise ratio [12,65].

### Array Design

If two identical microphones spaced some distance from each other record identical waveforms, the two observed signals are perfectly coherent. As the microphone spacing increases beyond the characteristic wavelength of a pulse, the coherence decreases until the waveforms no longer resemble each other. Microphone arrays are designed to detect coherent signals within a specific wavelength range, and the ability of an array to accurately determine the speed and angle of arrival of coherent sound signals depends largely on the sensor distribution relative to the wavelengths of interest. A minimum of three sensors, deployed as an L, are required to discriminate incidence angle. However, a three element array has a very poor angular resolution and leaves no margin for error, as failure of a single sensor will substantially degrade its detection capability. Four or more sensors are preferable, yielding a broader frequency response and better measurement precision.

### Detection, Location, and Signal Identification

Infrasonic signals measured a few kilometers from the source can vary from powerful explosions ( $>10^2$  Pa) that dominate the recorded time series, to a background rumble ( $10^{-3}$  Pa) buried within the ambient sound field. A single infrasonic array can discriminate between a coherent signal arriving from the direction and height of the target source and a competing signal coming from a different angle and elevation [10]. Thus, arrays can 1) separate the coherent sound field from the incoherent ambient noise field, 2) identify and extract a specific infrasonic signal within the coherent ambient sound field (clutter) and 3) separate acoustic arrivals propagating through different waveguides yet originating from a signal source which may be stationary or moving. If two arrays are available, the target signal would be recorded at each array with an arrival angle pointing to the source and an arrival time consistent with the propagation path from the source to the receiver. The two array beams would intersect at the source, thus providing a geographic loca-

tion which would be confirmed and refined with the signal travel time information [25]. Thus two properly sited infrasonic arrays can unambiguously locate a source region and discriminate between sources. By optimizing detection with arrays, performing locations with two or more arrays, and characterizing preexisting sources of clutter, it is possible to acoustically recognize a source with high confidence. If other detection (such as seismic or imaging) and identification [36]) technologies are available, uncertainties in source location and identification drop substantially. This process of signal detection, location, and identification is routinely used successfully by the international community in the monitoring of natural and man-made events [27].

### Acoustic Speed and Velocity

Infrasound studies may refer to the acoustic phase velocity, group velocity, effective sound speed, trace velocity, and the celerity of an acoustic arrival. The first three quantities depend on the sound speed and wind speed in the atmosphere along the source-receiver path. The trace velocity of a signal may be measured directly by observing the apparent horizontal speed and direction of propagation of a signal across an array. The celerity of an arrival is defined as the ratio of range over the travel time, and may be conceived as the effective, or average propagation speed over the complete propagation path. Given a known source-receiver configuration, the celerity may be computed directly. Although it is tempting to use the trace velocity for identifying a given arrival, measurement and calibration uncertainties can make this procedure rather inaccurate for distant sources. Celerity estimates from various infrasonic events with known locations suggests that 1) our knowledge of the atmosphere is presently insufficient to reliably predict all infrasonic arrivals to a station under all atmospheric conditions, and, 2) diffraction and/or scattering can feed acoustic energy to waveguides that are elevated above the ground, and these elevated waveguides may also leak sound back to the ground [10].

### Analysis Method and Scope

Array processing methods allow us to extract coherent signals from the incoherent ambient sound field. One efficient and popular technique for estimating the infrasonic wave parameters is the *Progressive Multi-Channel Correlation* method (PMCC) [12]. This method, originally designed for seismic arrays, is well adapted for analyzing low-amplitude coherent waves within incoherent noise and efficient for differentiating signals of interest from background clutter [22,26]. The PMCC method was

used for array processing of the signals from earthquakes, tsunamis, and volcanoes presented in this chapter.

Acoustic-gravity waves form a class of propagating atmospheric pressure signals which were studied in detail during the early Megaton-yield atmospheric nuclear tests (e. g. [32]). These waves are affected by buoyancy, have periods longer than 50 s, and have unique source and propagation characteristics that are beyond the scope of this paper. The IMS and portable arrays discussed in this chapter are tuned to higher frequencies, so are generally not designed to process acoustic gravity waves with high precision. The reader may refer to [33] for an excellent introduction to gravity and acoustic-gravity waves.

This chapter will also omit the acoustics of structural collapses such as pyroclastics flows from volcanoes, avalanches, landslides, or rockfalls. Although such signals may portend hazardous events (e. g. [79]), this family of acoustic signals are poorly understood and even less well modeled.

### Earthquake Infrasound

Ground vibrations can produce sound in manifold ways. Relative to the atmosphere, earthquakes can act as distributed, supersonic acoustic sources moving at seismic or fault rupture velocities. When seismic surface waves travel through mountainous regions, the predominant source of infrasound is attributed to the re-radiation of pressure waves by topography [48,60,78]. In this case, the earthquake-induced displacement perpendicular to the ground surface can locally generate ground-coupled air waves. The local conversion from seismic waves to the sound pressure has been observed on microbarometers at regional and teleseismic distances [14,15,18,69]. Seismic to acoustic wave coupling at the ground-air interface will be enhanced when the horizontal phase velocity, or trace velocity, of the infrasonic waves and the seismic waves are matched. This type of seismoacoustic coupling can be particularly efficient in sediments and loosely consolidated volcanic environments with a low shear wave velocity [24].

The generation of infrasonic waves from the epicenter region has also been postulated [59,62]. At large infrasonic periods (50–100 s) acoustic-gravity waves from the sudden strong vertical ground displacements have been detected at distances of thousands kilometers from the origin [6,57]. This mechanism would also apply for large submarine earthquakes [30,58]. In all the aforementioned cases, an actual pressure wave is radiated from the ground into the atmosphere. This process is distinguished from microphonics, where the recorded signal is due to the sensor response to accelerations and ground level elevations



independently of any pressure changes occurring in the atmosphere [2,4,45]. In this case, the microphone acts as a seismometer.

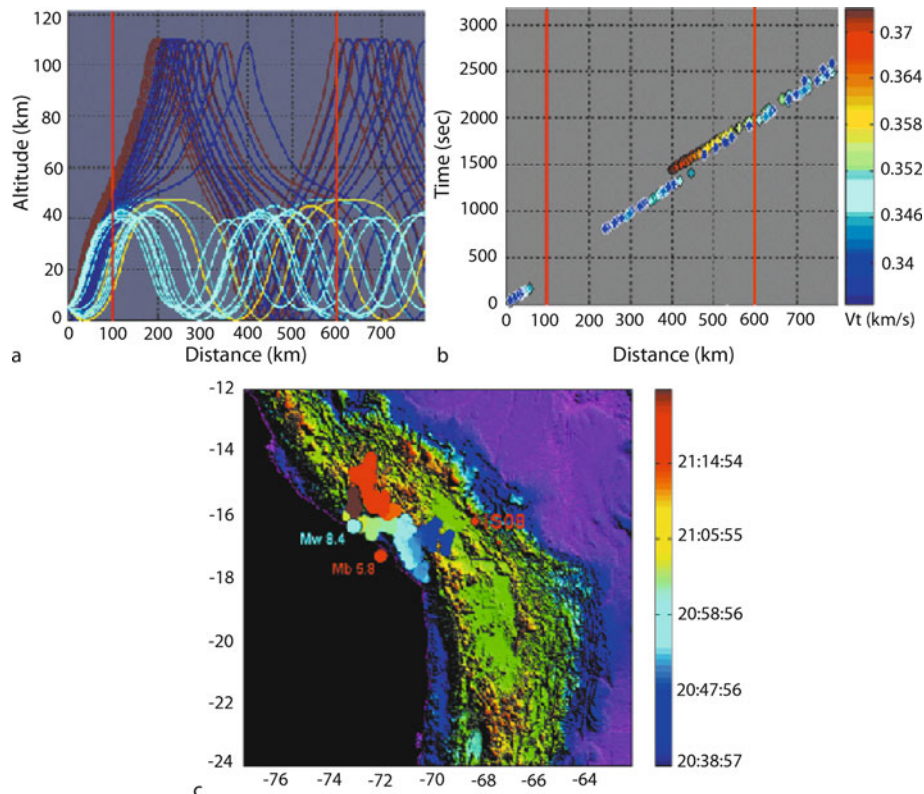
The principles introduced in the discussion of the generation, propagation, measurement, and interpretation of earthquake infrasound are applied to the progressive case studies of the Arequipa earthquake of June 23, 2001, recorded at a range of  $\sim 500$  km, the Mongolia earthquake of November 14, 2001, recorded at a range of  $\sim 1800$  km, and the Chile earthquake of June 13, 2005, detected by multiple infrasonic arrays to a maximum range of 2300 km.

### Case Study 1: $M_w$ 8.4 Arequipa Earthquake Detected by a Single Station

On June 23, 2001, at 20:33:13 UTC, a strong earthquake measuring  $M_w$  8.4 (NEIC) ripped along the coast of south-central Peru. The earthquake origin ( $16.15^\circ\text{S}$ ,  $73.40^\circ\text{W}$ , fo-

cal depth  $\sim 30$  km) was centered along the Peruvian coast about 600 km southeast of Lima and 110 km northwest of Camana (Fig. 1). The Pacific Tsunami Warning Center reported a moderate tsunami struck the Peruvian coast. Infrasonic waves associated with this event were detected for more than an hour at IMS infrasound station IS08 in La Paz, Bolivia ( $16.26^\circ\text{S}$ ,  $68.45^\circ\text{W}$ ) [48].

Due to the relative proximity of the station ( $\sim 500$  km from the epicenter), it was possible to perform relatively direct analyses of the apparent horizontal propagation speed (trace velocity) of the incident wavefield and the arrival angle of the different wave types at the array. The estimated trace velocity ranges from several kilometers per second to the sound velocity, as expected from the arrival of both seismic and infrasonic waves. Although the aperture of the IS08 array is designed for infrasonic wavelengths, array processing also yields the arrival characteristics of seismic waves. The azimuth variation for the seismic waves indicates that the rupture propagated from



Infrasound from Earthquakes, Tsunamis and Volcanoes, Figure 1

Ray traces **a** and travel time curves **b** for infrasonic waves launched almost horizontally from the epicenter area at 5 km height. The red line indicates the propagation range of the rays from the secondary sources to IS08. The color scale indicates the horizontal trace velocity of each ray. **c** Location of the sources of distant generation of infrasonic waves measured from 20:39 to 21:28 due to the  $M_w$  8.4 earthquake and the  $m_b$  5.8 aftershock (Topography data: USGS DEM & Cornell Andes Project). The colored dots indicate the arrival times (UTC) of the infrasonic waves at the station

the northwestern to the southeastern part of the fault at a speed of 3.3 km/s. However, the predominant source of infrasound is attributed to pressure waves radiated by the Andean Cordillera. This is consistent with the theory that the vibration of mountains can generate infrasonic waves which travel to the station along atmospheric waveguides. By performing basic inversions, the azimuth variation of the infrasonic waves can then be interpreted as a distribution of secondary sources along the highest mountain ranges. Using the azimuth and arrival time determination, the infrasonic radiation zone was estimated to be  $\sim 100$  by 400 km long.

### Case Study 2: $M_s$ 8.1 Mongolia Earthquake Detected by a Single Station

On November 14, 2001, at 09:26:10 UTC, a magnitude  $M_s$  8.1 earthquake rattled the mountainous western Chinese region near the Qinghai-Xinjiang border. The earthquake origin ( $36.0^\circ\text{N}$ ,  $90.5^\circ\text{E}$ , focal depth  $\sim 5$  km) was centered along the northern margin of the Tibetan Plateau at the foot of the Kunlun Mountains where substantial surface fault ruptures have occurred before. Coherent infrasonic waves associated with this event were detected for more than one hour at a distance of 1800 km from the epicenter by IMS station I34MN in Mongolia [49].

Building on the conclusions from the Arequipa earthquake analysis, both an inverse location procedure and a complete simulation of the radiated pressure field are used to locate the distant source regions. The input parameters of the location procedure include the measured signal azimuths and arrival times as well as the origin time and coordinates of the main shock. The propagation model is based on a constant velocity of 3.3 km/s for seismic surface waves propagating from the epicenter area. The atmosphere is specified by sound velocity and wind speed profiles obtained from the time-varying MSISE-90 and HWM-93 empirical reference models [25,38], and the infrasonic wave propagation was performed using 3D ray theory [73,74].

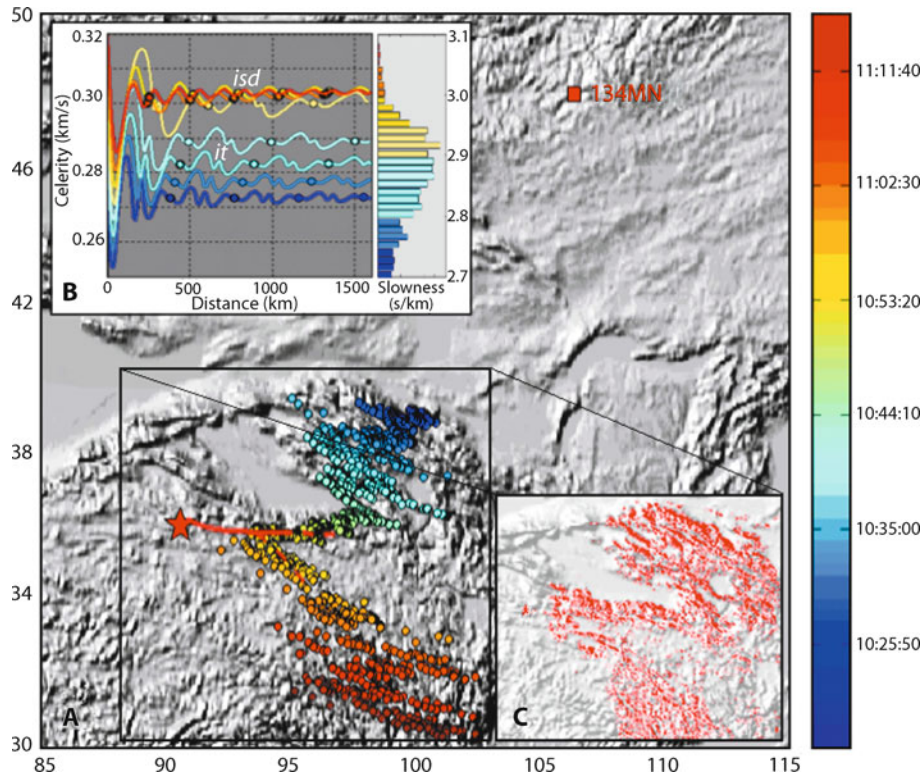
As shown by Fig. 2, two dominant guided wave groups are predicted: (i) thermospheric arrivals refracted below 120 km for slowness ranging from 2.7 to 2.9 s/km (trace velocities of 0.345–0.37 km/s), and (ii) stratospheric ducted waves refracted below 45 km for slowness values ranging from 2.9 to 3.1 s/km (0.32–0.345 km/s). Such trapped waves can be observed when the source is located above the station [75].

The slowness distribution derived from the measured trace velocity presents a maximum between 2.85 and 2.95 s/km. These values correspond to a celerity of 0.28–

0.30 km/s. The component of the wind transverse to the propagation direction deflects the rays from the original launch azimuth by  $\sim 2^\circ$ . This deviation is taken into account by correcting the measured azimuths. Figure 3a reconstructs the distant source regions using a celerity of 0.29 km/s. The spatial extent of the radiating zone is estimated to be  $9^\circ$  in latitude and  $10^\circ$  in longitude. The source distributions fall into line with the Qilian range, then borders the eastern part of the Qaidam basin and join the Kunlun range. To the south of Qaidam basin, more scattered source distributions follow the Bayan Har mounts.

To verify these locations, a simulation of the radiated pressure field was performed. First, the inversion for the rupture propagation along the fault uses a slip patches model developed by [7]. Using this extended model of rupture, synthetic seismograms of surface waves are computed using a discrete wavenumber method [8,70,72] with a one-dimensional regional crust model. The source modeling displays a strong directivity, with most of the seismic energy radiated along the main strike-slip of the fault and a maximum ground velocity placed  $\sim 300$  km to the east of the epicenter [53]. To compute the acoustic radiation of the topography surrounding the fault, the topography is divided into adjacent strip-line sources radiating energy proportional to the simulated ground velocity [41].

Compared to the wavelength of the seismic surface waves ( $\sim 60$  km), the area of each source element ( $3 \times 3 \text{ km}^2$ ) is small enough to consider isophase vibration. Due to the low frequencies of interest ( $kL > 1$ ,  $k$  and  $L$  defining the acoustic wavenumber and the side of each cell, respectively), source elements radiate essentially simultaneously with a pronounced directivity. Based on this assumption, the topography is divided in adjacent strip-line sources of length  $L$  radiating energy proportional to the simulated ground velocity  $V_l$  normal to each surface element  $l$ . Considering a distance of observation  $R_l$  significantly greater than  $L$ , the Fraunhofer approximation of the Helmholtz–Huygens integral yields:  $p_k(t) = iL(k\rho c)/2\pi \sum_{l=1}^N V_l(t_l) \Delta h_l (e^{-ikR_l})/R_l [\sin(k\hat{x}_l L/2)/(k\hat{x}_l L/2)] e^{[-ikco(t-t_l)]}$ , where  $t_l$  is the origin time of each source element,  $p_k(t)$  is the predicted pressure at the arrival time  $t$  ( $t = t_l + R_l/c_{\text{eff}}$ ),  $c_{\text{eff}}$  is the celerity in the atmosphere corresponding to the predicted wave guides (0.29 km/s),  $\rho$  is the air density,  $co$  is the sound speed,  $\Delta h_l$  if the height difference of the radiating surface and  $\hat{x}_l = \sin(\theta_l)$  is given by the angle  $\theta_l$  between the outward unit normal and the source/receiver vector. Note that the sinc function in the predicted pressure reaches its peak when the surface normal and the source-receiver vector are aligned. Thus the Fraunhofer approximation provides a means of evaluating a discretized source pressure term



Infrasonics from Earthquakes, Tsunamis and Volcanoes, Figure 2

**Propagation modeling and location of distant source regions of infrasonic waves (Topography data: ETOPO30). A** The colored dots indicate the sources location according the detected arrival times (UTC) of the infrasonic waves. Taking into account uncertainties due to the measurements and the propagation modeling, a maximum location error of 20 km is estimated for each dot. **B** Predicted celerity models versus slowness and propagation range for a source located at the main shock epicenter. The definition range of the celerity is given by the maximum of the slowness distribution derived from the measured trace velocities (Fig. 2). The circles indicate the locations of the ground reception of both ducted stratospheric (*isd* arrivals) and thermospheric paths (*it* arrivals). **C** Normalized surface pressure distribution along the Kunlun fault

for radiating strip lines of topography, where each strip will yield its largest pressure contribution when its displacement is along the source-receiver direction.

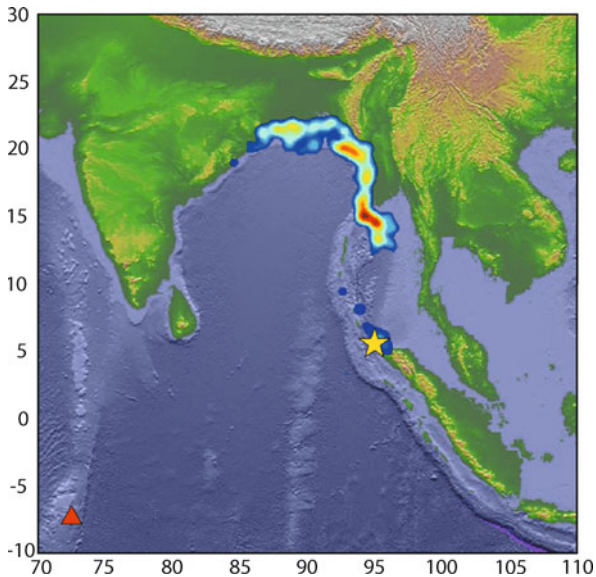
Using the simulated ground velocity and this approximation of the integral formulation, the predicted surface pressure distribution matches reasonably well with those obtained with the inverse location procedure (Fig. 2). Thus, the azimuth variations and the expansion of the signal duration suggest that the Kunlun Mountains acted as sources of infrasonic waves over a radiating zone of  $\sim 1000 \times 1000$  km. The maximum seismo-acoustic coupling is found to the east of the main shock epicenter, which is consistent with the seismic radiation pattern.

### Case Study 3: M 7.8 Chile Earthquake Detected by Multiple Stations

On June 13, 2005, a major earthquake occurred in the mountainous section of the Tarapaca Province (North

Chile) ( $19.93^{\circ}\text{S}$ – $69.03^{\circ}\text{W}$  at 22:44:33 UTC, M7.8, focal depth 117 km, USGS). The epicenter was located deep under the Andes mountain range, near Chile's border with Bolivia. At large distances from the epicenter, coherent infrasonic waves were detected by IMS infrasonic stations I08O-Bolivia, I09BR-Brazilia, and I41PA-Paraguay (410 km, 2300 km, and 1420 km from the epicenter, respectively). The multiple station recordings at different ranges and azimuths from the epicenter allowed a more complete reconstruction of the infrasonic source regions compared to what was obtained using one single station [52].

To invert for the main source regions of infrasonics, the location procedure requires the signal azimuths and arrival times measured independently by each station and the origin time and coordinates of the epicenter. A joint inversion for the source area using data from all three stations simultaneously was not used due to the pronounced directivity of the radiation pattern. Infrasonics



Infrasound from Earthquakes, Tsunamis and Volcanoes, Figure 3 Location distribution of distant source regions of infrasonic waves generated by the December 26, 2004 Sumatra earthquake. Color scales are normalized. The red triangle indicates the location of the IS52 station. Source locations are computed for seismic surface waves and tsunami waves originating from the maximum of coseismic slip

propagation was simulated using a spherical coordinate 3D ray theory formulation which accounts for topography and the spatio-temporal variations of the horizontal wind terms along the ray paths [17]. Atmospheric absorption is integrated using altitude dependent attenuation coefficients [66,67]. The atmospheric conditions of June 13, 2005 are described by the sound velocity and wind speed profiles provided by the time-varying Ground to Space (G2S) atmospheric specifications [19].

As in the previous case study, the slip patches model developed by [72] is used to check the association of regions radiating infrasound with areas of strong ground motion. The simplest extended elliptic source model able to explain the teleseismic seismograms is found, and the first and second order characteristics of the event (location, depth, duration, focal mechanism, and refined kinematic parameters such as spatial slip distribution on the fault and rupture velocity) are calculated from teleseismic body waves. Then, from the resulting extended source model, low frequency synthetic seismograms (period lower than 10 s) are computed on a grid in the vicinity of the epicenter using the discrete wavenumber method and a one-dimensional regional crust model [8]). Finally, the root mean square of the maximum velocity of the vertical and horizontal components of the surface waves is used to reconstruct the areas of strong ground motion.

The reconstructed source regions confirm that most of the energy is radiated by the vibration of land masses near the epicenter, which is consistent with the predicted areas of strong ground motion. No clear signal originates from the Altiplano. Southern high mountain ranges, even far from the epicenter, also generated infrasound. The Central Cordillera extending to altitudes greater than 5000 m efficiently produced infrasound in the direction of I09BR, although the predicted seismic movement is low in this region. Consistent with previous observations, these results suggest an amplification of the ground displacement caused by the topography surrounding the Altiplano. Such site effect could not be predicted from our seismic source modeling since the topography is not considered. The reconstructed source regions extend over  $\sim 800$  km from the Central Cordillera to the Occidental Cordillera. The spatial extent of the radiating zones differs from one station to another, which confirms the influence of shadow zone effects for nearby stations and the directivity of the radiation. As in the previous case study, the topography is modeled as a succession of adjacent strip-line sources, so that mountain ranges radiate energy essentially simultaneously with a pronounced directivity and may generate infrasound arrivals with different azimuths. This suggests that the amount of energy radiated in the direction of the receiver and the duration of the signals also depends on the orientation of the highest mountain ranges around the station.

### Summary of Earthquake Infrasound

The three case studies presented in this section build in complexity and sophistication while producing consistent results. High mountain chains rattled by large earthquakes reliably radiate infrasound and have acoustic radiation patterns that depend on the chain's orientation. Substantial contributions to the sound field are expected from steep topography, which would primarily radiate perpendicular to exposed faces. For large earthquakes occurring in mountainous regions, infrasonic measurements are valuable for the analysis of the remote effects of earthquakes and site effects over broad areas. In remote regions where there is a lack of surface motion instrumentation, infrasonic observations could lead to a rapid determination of the regions where the seismic movements are the largest.

### Tsunami Infrasound

Previous study has shown that significant infrasound is produced by breaking waves (surf) and the complex interaction of open-ocean swells (microbaroms, e. g. [76]).

However, interest in tsunami infrasound emerged when IMS infrasound arrays in the Pacific and Indian Oceans recorded distinct signatures associated with the December 26, 2004 Great Sumatra-Andaman earthquake and tsunami. As in the case of mountains stirred by continental earthquakes, islands which undergo significant surface displacements during submarine earthquakes can also produce infrasound. It also appears that the initiation and propagation of a tsunami may produce low frequency sound near the epicenter as well as along coastlines and basins. In some environments, precursory sound could potentially be used for confirmation and early warning for tsunamis. This field of research is still in its infancy, and our interpretations leave much room for further development. Substantial complexity is involved in the separation of the earthquake-generated infrasound from the sound that may be produced by the genesis and propagation of the tsunami.

### Sumatra Earthquake and Tsunami

The magnitude 9.1 Great Sumatra-Andaman earthquake of December 26, 2004 [63] is the largest earthquake since the 1964 magnitude 9.2 Great Alaska earthquake, and produced the deadliest tsunami in recorded history. In contrast to the Great Alaska earthquake, the Great Sumatra earthquake and tsunami were recorded globally by multiple digital sensor networks in the ground, ocean, atmosphere, and space. Further, the resulting signals were analyzed and interpreted using far more advanced computational capabilities than were available 40 years ago. Although the study of tsunami infrasound rose in the wake of the Sumatra-Andaman event, many fundamental questions on the generation of these deep sounds remain unanswered.

The clearest infrasonic signatures associated with the Sumatra event were captured by the station in Diego Garcia (IS52GB, Fig. 3), which recorded (1) seismic arrivals from the earthquake, (2) tertiary arrivals (T-phases) that propagated along sound channels in the ocean and coupled back into the ground, (3) infrasonic arrivals associated with either the tsunami generation mechanism near the seismic source or the motion of the ground above sea level, and (4) deep infrasound (with a dominant frequency lower than 0.06 Hz) coinciding with the propagation of the tsunami into the Bay of Bengal [30,51]. These signals were all recorded by the pressure sensors in the arrays. The seismic and T-phase recordings are a result of the sensitivity of the microphones to ground vibration (microphonics), whereas the infrasound arrivals correspond to sound propagating through atmospheric waveguides.

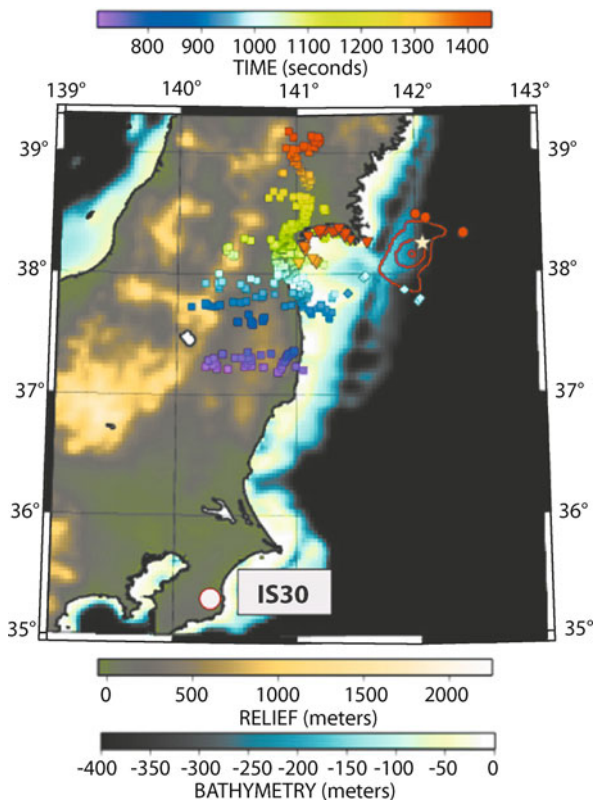
Similar, but not identical arrivals were observed at Diego Garcia (range of  $\sim 2900$  km) during the March 28, 2005 Nias earthquake (M8.7, which produced a non-destructive tsunami). Yet only infrasonic arrivals were observed from the April 10, 2005 Mentawai earthquakes (M6.7 and 6.5, no reported tsunami), indicating that above-water ground motion from submarine earthquakes can produce sound in the Sumatra region. The deep infrasound from the Bay of Bengal region following the Sumatra earthquake suggests that sound can be produced by the interaction of a tsunami with coastal bathymetry.

To reconstruct the main source regions of infrasound recorded at IS52 for this event, the input parameters of the location procedure included the measured signal azimuths and arrival times, and the origin time and coordinates of the epicenter. As described in the previous section, standard velocity models were used to describe the propagation of the seismic surface waves and the propagation of infrasound through the atmosphere in the direction of IS52. However, the speed of propagation of the tsunami from the epicenter needs to be added to the inversion procedure. For the December 26, 2004 earthquake, a velocity of 3.3 km/s is used for the seismic surface waves, and a speed related to the square root of the water depth for the tsunami waves.

Using the infrasonic arrivals to IMS station in Diego Garcia and Palau, the location inversion for the high frequency infrasonic component corresponded to sound originating from the mountains and possibly from the epicenter. However, a unique characteristic of the Sumatra earthquake is that it generated large amplitude coherent waves with a dominant period of  $\sim 30$  s over four hours. The infrasonic source locations derived from the inversion procedure indicate that the tsunami also created infrasonic waves at lower frequencies when it propagated in shallower water as it reached the Bay of Bengal (Fig. 3). Even lower-frequency acoustic gravity waves, with periods of hundreds of seconds, may have been produced by the underwater ground displacement [58]. For wavelengths which are much greater than the water depth, the ocean surface displacement nearly matches the submarine ground displacement above the fault plane (e. g. [44]) and may efficiently radiate long-period pressure waves into the atmosphere.

### Miyagi-Oki Earthquake and Tsunami

The events off the coast of Sumatra were  $\sim 3000$  km from the closest infrasound station in Diego Garcia. The longer ranges, coupled with the fact that all infrasound stations used in those studies were transverse to the axis of Suma-



Infrasound from Earthquakes, Tsunamis and Volcanoes, Figure 4 Estimated infrasound source locations associated with ground vibration, tsunami genesis, and the interaction of the tsunami with the coastline. The squares represent stratospheric arrivals with a celerity of 0.3 km/s. The diamonds are also stratospheric arrivals but with the celerity of 0.32 km/s predicted for that azimuth. The circles are thermospheric arrivals with a celerity of 0.27 km/s. The triangles are stratospheric arrivals with a celerity of 0.3 km/s, but with a delay time attributed to seismic seiche formation. The color of the symbols indicates the arrival time in seconds since the earthquake's origin time. The topography is from NOAA ETOPO2 data

tra, caused uncertainty in the ability to discriminate between sounds potentially produced during tsunami genesis at the ocean surface and the sounds produced by the earthquake-induced vibration of mountains and islands. In contrast, IMS infrasound station IS30 in Japan (Fig. 4) is optimally situated to recognize the different source regions of infrasound associated with the Miyagi-Oki earthquake and tsunami.

The magnitude 7.2 Miyagi-Oki earthquake occurred on August 16th, 2005 at 02:46:28 UTC. The epicenter was off the coast of Japan near Honshu (38.251°N, 142.059°E), with an estimated depth of 36 km. The earthquake caused landslides, ~60 injuries, as well as power and transportation disruptions. A local, nondestructive tsunami was observed on the coast of northern Japan with a wave height

of ~10 cm. Because of its auspicious location, station IS30 can use the angle of arrival information derived from array processing to identify infrasonic arrivals originating from mountain chains, the earthquake epicenter (which is used as the tsunami epicenter), and the coastline of the Bay of Sendai.

In contrast with the Sumatra event, where substantial energy was observed in the deep infrasound band (0.002–0.1 Hz), most of the infrasonic energy for the smaller Miyagi-Oki event was above 0.5 Hz. The arrival azimuths at IS30 range from  $-5^\circ$  to  $28^\circ$ , and do not have a well-defined temporal sequence, suggesting multiple sources, propagation paths, and possible wave types.

Using the ground to space (G2S) atmospheric profiles [19] specific to the station location and time of the event, source locations [30,51] were estimated for the infrasonic arrivals shown in Fig. 4. The first infrasonic arrivals would correspond to acoustic waves coupled to the atmosphere from the seismic vibration of land masses [49]. For ranges less than 330 km and northerly arrivals between  $-5^\circ$  and  $18^\circ$  (measured clockwise from north), the time- and site-specific G2S specifications do not support thermospheric arrivals at the station, so the observed arrivals would propagate in stratospheric waveguides with a celerity of 0.3 km/s. In contrast, for arrival azimuths of  $\sim 18^\circ$  to  $28^\circ$  originating offshore, thermospheric arrivals are also supported for ranges greater than 330 km, although the celerity of stratospheric arrivals is faster.

Multiple wave propagation paths are invoked to produce the temporal and frequency distribution of the infrasonic signals observed at IS30. The squares shown in Fig. 4 correspond to seismic vibrations radiating acoustic energy into stratospheric waveguides, thereafter propagating with a celerity of 0.3 km/s, the diamonds to stratospheric arrivals with a celerity of 0.32 km/s, and the circles correspond to thermospheric arrivals with a celerity of 0.27 km/s. Inclusion of seismic speeds made a negligible difference, as it offsets the locations by <6 km. As expected, many of the arrivals originate from land, and are consistent with the predicted seismic intensity [13,71]. Of more relevance to this paper, clear stratospheric and thermospheric arrivals arrive from the perimeter of the epicenter, which is assumed to be the region of tsunami genesis. These results are consistent with the proposed infrasonic source locations obtained by [30] for the Sumatra earthquakes, and support the idea that ocean surface displacements associated with submarine earthquakes produce infrasound [58]. The possibility that vertical ocean displacements near the epicenter may radiate sound suggests that infrasound signals may be potential discriminants for tsunami genesis.

The later arrivals (shown as triangles in Fig. 4) originate from the mountain regions to the north of the Bay of Sendai when stratospheric and thermospheric arrivals are assumed. Yet, analyses of the Great Sumatra earthquake suggest that the Bay of Bengal produced deep infrasound, and the possibility that the Bay of Sendai may also act as an acoustic source is considered. The triangles in Fig. 3 contour the coastline of the northern Bay of Sendai when a combination of a delay time and stratospheric apparent propagation speeds are considered. For a celerity of 0.3 km/s, the inferred delay time is  $\sim 250$  s. This would require an acoustic generation process that requires  $\sim 4$  min to be established over a coastal region  $\sim 100$  km long. The earthquake occurred in a shallow region, which corresponds to slow tsunami propagation speeds of  $\sim 0.05$  km/s. This would mean the tsunami would have traveled at most 12 km in 250 s, which is not enough time for the tsunami to reach the coastline and produce these signals. Thus the interaction of the small tsunami with the coastline is eliminated as a possible source.

A plausible ensonification mechanism is the local production of coastal waves by the earthquake, as in the generation of seismic seiches [5,56]. Low frequency radiation from enclosed bodies of water have been proposed by other authors [9,46]. For the Miyaki-Oki earthquake, substantial water displacement in shallow regions of the bay is strongly suggested by acoustic sources above the water near the coastline (blue and green arrivals in Fig. 4). Satellite imagery shows that the northern part of the Bay of Sendai has an abundance of lakes, bays, man-made harbor structures and rivers which may sustain seiches. Lower order seiche modes would take minutes to be established, but higher-order modes and coupled oscillations associated with the narrower of the volume dimension could sustain an ensonification process akin to microbarom generation from the ocean (e. g. [76]). Alternatively, the triangular symbols in Fig. 4 would should be shifted  $\sim 75$  km North of their shown location.

### Summary of Tsunami Infrasound

These two case studies strongly suggest that submarine earthquakes and the water level changes they induce can produce low-frequency sound. The sound may be radiated from the ocean surface during the tsunami genesis, produced by the vibration of land masses near the epicenter, or be excited by the interaction of seismic and water waves with the coastline, shallow bathymetry, and harbors.

There is some potential for using infrasound in conjunction with other technologies for remote tsunami monitoring. The effective propagation speeds of tsunami

( $\sim 50$ – $200$  m/s) and sound waves ( $\sim 300$  m/s) yield an advance warning time of at least 1.7 s/km. At 100 km, sound leads the tsunami by at least 170 s, but some of this time would be taken up by signal transmission, processing and identification, leaving less than one minute to issue an alert. However, infrasound may provide early warning in areas within shallow basins or further than a few hundred kilometers from the tsunami source region.

### Volcano Infrasound

The previous sections discussed how earthquake infrasound acoustically couples solids and gases, and tsunami infrasound couples solids, liquids, and gases. Similarly, a volcano can transmit information about its current eruptive state through vibrations induced in the ground, volcanic fluids, and atmosphere. However, there is a great deal of ambiguity about the composition of the acoustically active volcanic fluids, which may vary from vesiculated magma, through gas-ash mixtures, to steam. Although a sealed volcanic conduit can produce faint sounds through earthquakes, once the volcanic plumbing breaches the surface it can broadcast infrasound quite unambiguously.

### Basic Principles

The birth of volcano infrasound research may be traced to the cataclysmic 1883 eruption of Krakatoa, which triggered a series of interdisciplinary, international geophysical studies of the pressure signals produced by the volcano [77]. Barometric records observed throughout the US, Europe, and Russia, and reports of cannon-like sounds in surrounding islands (as far as Diego Garcia and Rodrigues Islands) demonstrated for the first time the ability of low frequency sound to propagate for thousands of kilometers.

Infrasonic signals originating inside magma conduits contain information about the pressure fluctuations driving eruption processes. These signals may be broadly categorized as explosions, long-period events, and tremor. Very large eruptions can also produce acoustic gravity waves [34,68]. Explosions are impulsive and have durations of seconds, long period events are more emergent than explosions, have distinct spectral peaks, and may last seconds to minutes, and tremor signals are sustained atmospheric vibrations that can persist from minutes to years. Decades ago, volcano seismology identified tremor and long-period events as signals indicative of near-surface intrusion of volcanic fluids (e. g. [11]). In contrast, the first infrasonic measurements of tremor from Sakurajima

were published in [64]. As the sensitivity of instrumentation and sophistication of analysis methods increased with advancing technology, the catalogue of known volcanic sounds has expanded and the ability to remotely detect hazardous eruptions has extended. Ongoing monitoring efforts suggest that relatively gentle Hawaiian and mild Strombolian activity may be consistently observed from a range of  $\sim 10$  km [20] and strong Strombolian, Vulcanian and Plinian eruptions from distances of tens to hundreds of kilometers [31,50,54,68].

The difficulty with modeling and interpreting volcanic sounds is that there is a lot of ambiguity in the geometry, composition, thermodynamics, and phase (solid, liquid, gas, or a mixture) of the volcanic interior preceding and during an eruption. In addition, there are many possible ways of exciting volcanic fluids into oscillation through unsteady and transient flow. As examples of these complications, some possible ways of exciting tremor signals in a magma conduit are described. Tremor signals are endemic to volcanoes, and are generally attributed to magma intrusion. Some plausible driving mechanisms of tremor signals and the physical properties of the materials inside volcanic pipes are discussed. In the following example, it is postulated that the harmonic spectrum of tremor events is caused by the acoustic resonance of the fluid within the magma conduit of Arenal volcano, Costa Rica [23]. The possible relationships between observed signals and the gas content of the melt, the physical conditions inside the volcano, and the flow dynamics of the volcanic fluids are discussed.

### Case Study: Tremor Signal at Arenal Volcano

This section hopes to provide an overview of the complexities of modeling volcanic sounds by considering the frequency changes in a harmonic tremor signal recorded at Arenal volcano, Costa Rica. Figure 5 shows a section of the infrasonic tremor signal shown in [23]. This signal is illustrative of most of the tremor recorded during April–May 1997 by a three-element array of infrasonic sensors and a five-element array of seismometers [35].

The prominent features of the Arenal tremor are the harmonic character of the spectra and the oscillation of the spectral bands with time. This oscillation of the spectral bands about an equilibrium value is referred to as gliding. Since steady flow cannot generate acoustic waves, it must be the pressure, volume, and flow fluctuations of the magma injection process that are driving the infrasonic signals. It is possible that the gliding of the spectral bands is caused by time-varying changes in the flow regime or the physical properties of the material in the conduit. The

resonant magma conduit model of [21] is used to explore the ramifications of having changes in the conduit length, flow velocity of a gas and in the void fraction of a magma-gas mixture.

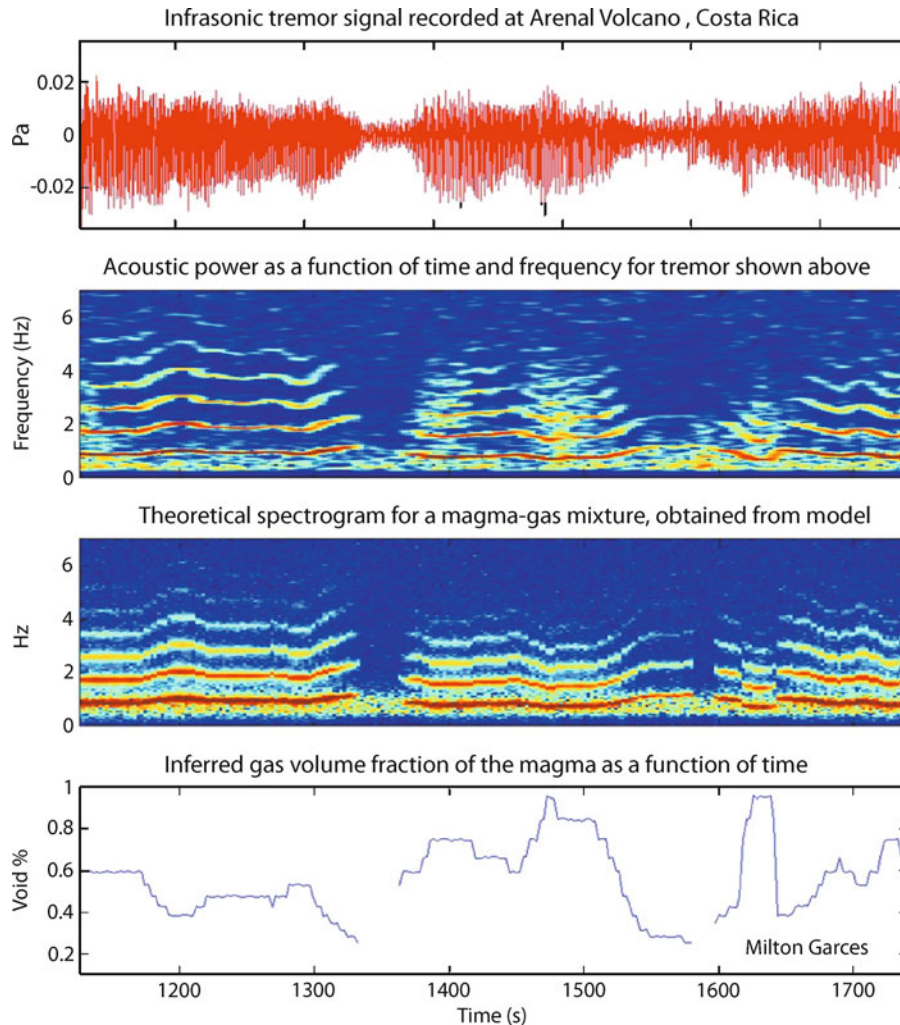
The tremor signal shown in Fig. 5 lasts over ten minutes and is not a transient event. Attempts to model the spectra with a random continuous source, as successfully done for Pavlof volcano [24], failed at Arenal because of the sharpness of the spectral peaks and the rapid rolloff at higher frequencies. The tremor signal appears to be sustained by a source mechanism that is triggering repeatedly in time. The source region may consist of a compliant gas volume within a magma-filled conduit. A bubble may form around a corner or in a constriction in a magma-filled conduit, where fluid acceleration may generate relatively stable gas-rich pockets. These cavities will be sensitive to flow fluctuations, and may act as effective acoustic sources [21].

Lava flows at Arenal and incandescent pieces ejected during explosions strongly suggests an open magma conduit associated with the infrasonic recordings. However, there is some ambiguity on whether the acoustically active part of the conduit is filled with a magma-gas or an ash-gas mixture. The source region is placed at the lower part of the magma conduit, and it is assumed that it ensonifies the magma-gas mixture above it. Although for the Arenal signal there is evidence for a buried tremor source [35], it may also be possible to have efficient atmospheric excitation from jet flow above the vent during more powerful eruptions [31]. For the purposes of our discussion, a two-layer, open vent solution representing a slow-velocity, gas rich magma floating atop a less vesiculated layer is used, as developed in Sect. 4 of [21]. The source region at the bottom of the conduit could correspond to a constriction in the conduit, where cavitation may occur due to fluid acceleration.

The sound speed of a liquid gas mixture is a strong function of the amount of bubbles, or the void fraction, in the mixture. For high void fractions, the sound speed of a magma-gas mixture may reach sound speeds of tens of meters/second, and sound would be heavily attenuated. For a resonant fluid column of length  $L$ ,  $f_n = n(1 - M^2)c/(2L)$ , where  $f_n$  is the resonant frequency of the  $n$ th spectral peak,  $M$  is the Mach number, or ratio of the flow speed to the sound speed, and  $c$  is the sound speed of the fluid.

Assuming a gas-rich magma inside a resonant magma conduit, the changes in the spectral peaks of the tremor signal may be attributed to changes in the amount of exsolved gas in the melt (Fig. 5), which would dramatically change the sound speed  $c$ . The tremor signal would cease





Infrasound from Earthquakes, Tsunamis and Volcanoes, Figure 5  
 Recorded acoustic signal, spectrogram, synthetic spectrogram, and one possible interpretation for a harmonic tremor signal observed at Arenal Volcano, Costa Rica

if the mass flux stops or becomes steady. It is unlikely that there is only gas inside the conduit because the acoustic attenuation in the gas would not be sufficient to explain the rapid amplitude decay with frequency. However, it is possible to have a gas-ash mixture, in which case it may be plausible to also explain the gliding with a change in the Mach number of the flow or a variation in the length of the resonant portion of the conduit [21].

Attenuation coefficients for dusty gases introduce an additional number of poorly characterized physical parameters, adding further ambiguity to our interpretations. It is due to these and many other complexities in modeling volcanic systems that the science of volcanology has evolved towards the synergy of geophysical observations

and geological parameters [37]. Some of the primary goals of volcanic studies are to permit reliable eruption forecasting, hazard assessment, and early warning. Although it is difficult to forecast an eruption without a clear understanding of the physical processes associated with precursory activity, it is possible to rapidly identify a powerful hazardous eruption and provide early warning.

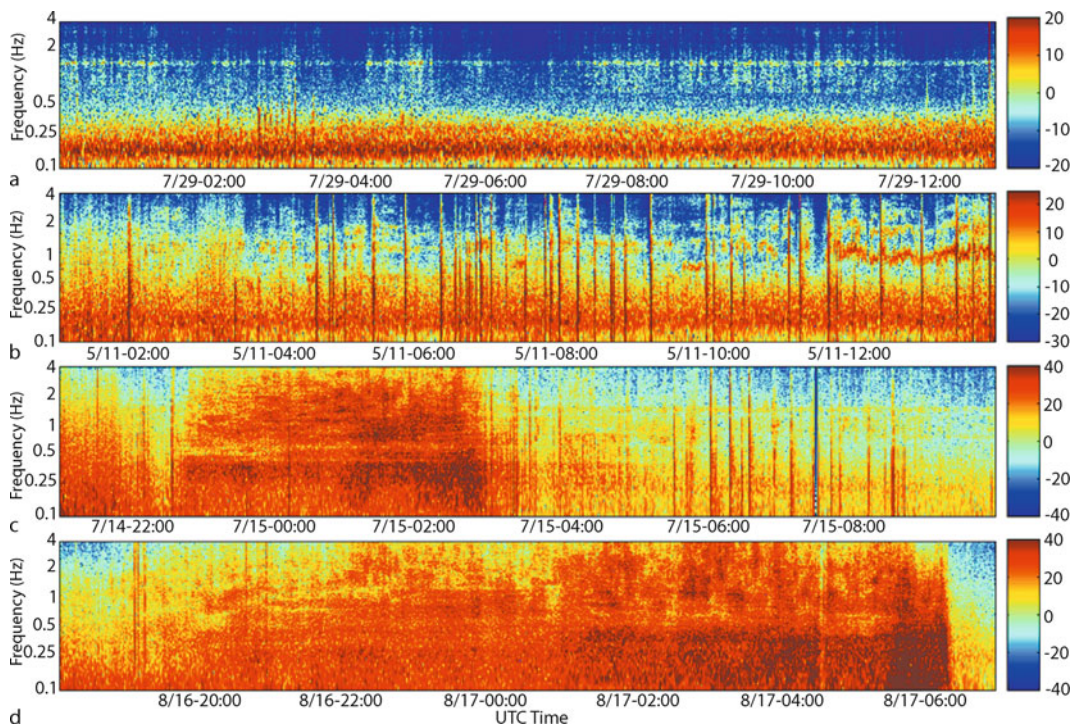
#### Prototype Operational System: ASHE

The Acoustic Surveillance for Hazardous Eruptions (ASHE) proof-of-concept project seeks to develop and evaluate the potential for robust, operational infrasonic remote sensing of volcanic eruptions. In contrast to

other ground-based volcano surveillance systems, the autonomous ASHE arrays are sited tens to hundreds of kilometers away from the devastation zones of erupting volcanoes. The diverse eruption signals produced by Tungurahua Volcano and captured by the ASHE arrays illustrate how acoustic remote sensing may complement seismic observations and satellite remote sensing to improve continuous monitoring of wide regions of potential eruption hazard.

The prototype ASHE stations consist of a four-element infrasound array with an aperture of  $\sim 100$  m, a broadband seismic sensor, and a wind sensor [54]. A satellite dish sends the digital data in real time from the field to the Geological Survey of Canada data center in Ottawa, Canada, which distributes the data in real-time to other collaborating parties. In January of 2006 the ASHE team deployed two infrasound arrays in Ecuador [31]. These two arrays, separated by  $\sim 250$  km, were sited to detect volcanic eruption in Ecuador or Southern Colombia by one of the arrays within 15 min of the eruption, and as early within 5 min for Tungurahua and Sangai volcanoes, which were within 40 km of an array.

Three distinct types of eruption signatures can be identified at Tungurahua (Fig. 6). The first and most common is low ash-producing background tremor with a dominant frequency of 1.4 Hz (Fig. 6a). During mid-May 2006, volcanic activity changed and was temporarily characterized by large explosions followed by harmonic tremor (Fig. 6b), reminiscent of the aforementioned tremor at Arenal. However, no significant ash release was observed, and this second type of eruptive regime was characterized as ash-poor. On July 14th, 2006 a large Vulcanian to sub-Plinian eruption (Fig. 6c) produced dangerous pyroclastic flows, substantial stratospheric ash clouds, and a significant increase in tremor. Between August 16th–17th, 2006 a larger Vulcanian to Plinian eruption occurred (Fig. 6d) and was characterized by lethal pyroclastic flows, a larger stratospheric ash cloud, and considerable infrasonic tremor. As the eruption progressed, the majority of acoustic energy shifted to lower frequencies ( $<0.5$  Hz). The acoustic signatures of the two major eruptions are comparable and easily identifiable in infrasonic records due to the vast amount of energy present over a broader band of frequencies. These two eruptions injected substantial ash



Infrasound from Earthquakes, Tsunamis and Volcanoes, Figure 6

Spectrograms for different eruption styles at Tungurahua volcano. The horizontal axis represents 13 hours of data, and the vertical scale shows frequency in a logarithmic scale from 0.1 to 4 Hz. The color denotes acoustic power in decibels referenced to 1 Pa/Hz. The ocean microbarom shows up in the upper panel two panels as a red band centered about 0.2 Hz, and is not of volcanic origin. Explosive events appear as intense vertical bands

into the stratosphere, suggesting this type of acoustic signature may be used for remote infrasonic monitoring of hazardous eruptions.

Automatic analysis techniques for eruption detection prototyped by the ASHE project produce automated low-latency email notifications coupled with more detailed real-time data products which can be used by responsible operational agencies to disseminate updated information. Based on the acoustic records captured during the Tungurahua eruptions of July and August 2006, source parameters that may be estimated during large eruptions include (but may not be limited to) the start time and duration of an ash cloud injection that could pose a hazard to international aircraft at cruising altitudes.

### Summary of Volcano Infrasound

Volcanic signals incorporate many of the complexities found in earthquake and tsunami signals, and then add some because of the inherently unstable thermodynamic and hydrodynamic conditions that lead to and accompany eruptions. Because infrasound measures the excess pressure change that drives an eruption, it can be used to help unravel the dynamic physical processes driving eruptive activity. Although infrasound may be used with additional technologies in forecasting and hazard assessment, its clearest contribution is in the identification of the intensity and timing of a powerful eruption. The technology and methodology to use infrasound for acoustic remote sensing of hazardous eruptions is mature and ready to implement into operations.

### Future Directions

Although infrasound has manifold applications in near-source studies, recent advances in sensor and analysis techniques have strengthened the potential to develop infrasound as a robust remote sensing tool that may cover observational gaps in distant and inhospitable environments. Although operational acoustic monitoring systems that can help notify aircraft of hazardous eruptions have been demonstrated as viable, much work is still needed to implement these systems and establish clear relationships between ash heights, infrasonic source parameters, and atmospheric conditions. Many parts of Earth's oceans remain undersampled, and there is a great potential for the application of infrasound to the detection of tsunami-genes, the identification and tracking of severe ocean weather, and the monitoring of breaking ocean wave intensity and other coastal hazards. Much fundamental research lies ahead in the construction of functional models for these various source processes. One of the most am-

bitious projects the community has initiated is the passive acoustic tomography of the atmosphere using natural sources [50]. These studies, refined and extended over the next decades, may contribute to our long-term assessment of global climate change.

### Concluding Remarks

There is beauty in complexity, as there is perplexity in beauty. The oscillatory interaction between the solid and fluid phases of our Earth reveals a wealth of information that can be turned into knowledge by the meticulous and persistent use of observation and modeling. Infrasound captures this solid-fluid interaction at the temporal scales of tens of seconds to tens of hertz, offering alluring glimpses into the inner workings of the natural world. The heaving ground and the frothing magmas, borne of the dark depths of Earth, oft conspire with the surging oceans to pervade the skies with deep sound. The field of infrasound strives to interpret the imperceptible yet omnipresent soundscape that permeates the air about us, so as to assuage uncertainty and perchance alleviate our dread of the natural threats posed by earthquakes, tsunamis, and volcanoes.

### Acknowledgments

The authors wish to thank C. Hetzer, P. Caron, and S. McNamara for figures and analyses. Garces also extends his appreciation to M. Protti, M. Haggerty, and S. Schwartz for their contributions to the Arenal studies. D. Fee and R. Matoza provided very useful revisions to this chapter, for which we are grateful. Garces' contributions to this chapter were supported in part by the National Science Foundation (grant EAR-0609669).

### Bibliography

#### Primary Literature

1. Alcoverro B, Le Pichon A (2005) Design and optimization of a noise reduction system for infrasonic measurements using elements with low acoustic impedance. *J Acoust Soc Am* 117:1717–27. doi:10.1121/1.1804966
2. Alcoverro B, Martysevich P, Starovoit Y (2005) Mechanical sensitivity of microbarometers MB2000 (DASE, France) and Chaparral 5 (USA) to vertical and horizontal ground motion. *Inframatrics* 9:1–10 <http://www.inframatrics.org/>. Accessed Mar 2005
3. Bhattacharyya, Bass JH, Drob D, Whitaker R, Revelle D, Sandoval T, Woodward R (2003) Description and Analysis of Infrasound and Seismic Signals Recorded from the Watusi High-explosive Experiment of September 28, 2002. SAIC technical report SAIC-03/2206
4. Bedard AJ (1971) Seismic response of infrasonic microphones. *J Res Nat Bur Stand* 75:41–45

5. Berninghausen WH (1969) Tsunamis and seismic seiches of Southeast Asia. *BSSA* 59:289–297
6. Bolt BA (1964) Seismic air waves from the great 1964 Alaskan earthquake. *Nature* 202:1095–1096
7. Bouchon M (1976) Teleseismic body wave radiation from a seismic source in a layered medium. *Geophys J Int* 47(3):515–530. doi:10.1111/j.1365-246X.1976.tb07099.x
8. Bouchon M (1981) A simple method to calculate Greens functions for elastic layered media. *Bull Seism Soc Am* 71:959–971
9. Bowman HS, Bedard AJ (1971) Observations of Infrasound and Subsonic Disturbances Related to Severe Weather. *Geophys J Int* 26(1–4):215–242. doi:10.1111/j.1365-246X.1971.tb03396.x
10. Brown D, Garcés M (2009) Ray Tracing in an Inhomogeneous Atmosphere with Winds, *Handbook on Signal Processing in Acoustics*. Springer (in press)
11. Chouet B (2003) Volcano Seismology. *Pageoph* 160:739–788
12. Cansi Y (1995) An automatic seismic event processing for detection and location: the PMCC method. *Geophys Res Lett* 22:1021–1024
13. Che I, Lee H, Jeon J, Kang T (2007) An analysis of the infrasound signal from the Miyagi-Oki earthquake in Japan on 16 August 2005. *Earth Planets Space* 59:e9–e12
14. Cook RK (1971) Infrasound radiated during the Montana earthquake of 1959 August 18. *Geophys J R Astr Soc* 26:191–198
15. Cook RK, Young JM (1962) Strange sounds in the atmosphere, Part II. *Sound* 1:25–33
16. Cox EF, Plagge HJ, Reed JW (1954) Meteorology Directs Where Blast Will Strike. *Bull Am Meteorol Soc* 35:95–103
17. Dessa JX, Virieux J, Lambotte S (2005) Infrasound modeling in a spherical heterogeneous atmosphere. *Geophys Res Lett* 32:L12808.1–5 doi:10.1029/2005GL022867
18. Donn WL, Posmentier ES (1964) Ground-coupled air waves from the great Alaskan earthquake. *J Geophys Res* 69:5357–5361
19. Drob DP, Picone MJ, Garces M (2003) Global morphology of infrasound partitioning. *J Geophys Res* 108(D21):4680. doi:10.1029/2002JD003307
20. Fee D, Garcés M (2007) Infrasonic tremor in the diffraction zone. *Geophys Res Lett* 34:L16826. doi:10.1029/2007GL030616
21. Garcés MA (2000) Theory of acoustic propagation in a multiphase stratified liquid flowing within an elastic-walled conduit of varying cross-sectional area. *J Volcanol Geotherm Res* 101:1–17
22. Garces M, Hetzer C (2003) Optimizing the Progressive Multi-Channel Correlation Detector for the Discrimination of Infrasonic Sources. In: *Proceedings of the 25th Seismic Research Review*, Tucson, 23–25 Sept 2003
23. Garcés MA, Hagerty MT, Schwartz SY (1998) Magma acoustics and time-varying melt properties at Arenal Volcano, Costa Rica. *Geophys Res Lett* 25:2293–2296
24. Garcés M, Hansen RA, Lindquist KG (1998) Traveltimes for infrasonic waves propagating in a stratified atmosphere. *Geophys J Int* 135:255–263
25. Garces M, Hetzer C, Lindquist K, Drob D (2002) Source Location Algorithm for Infrasonic Monitoring. 24th Annual DTRA/NNSA Seismic Research Review, Ponte Vedra, 17–19 Sept 2002
26. Garces M, Harris A, Hetzer C, Johnson J, Rowland S, Marchetti E, Okubo P (2003) Infrasonic tremor observed at Kilauea Volcano, Hawaii. *Geophys Res Lett* 30:2023–2027 2003
27. Garces M et al (2004) Forensic studies of infrasound from massive hypersonic sources. *EOS* 85(43):433
28. Garcés M, Willis M, Hetzer C, Le Pichon A, Drob D (2004) On using ocean swells for continuous infrasonic measurements of winds and temperature in the lower, middle, and upper atmosphere. *Geophys Res Lett* 31:L19304
29. Garcés M, Willis M, Hetzer C (2004) The Hunt for Leaky Elevated Infrasonic Waveguides. 26th Seismic Research Review, Orlando
30. Garces M, Caron P, Hetzer C, Le Pichon A, Bass H, Drob D, Bhattacharyya J (2005) Deep infrasound from the Sumatra earthquake and tsunami. *EOS* 86(35):317–320
31. Garces M et al (2008) An acoustic fingerprint of stratospheric ash injection. *EOS* 89(40):377–378
32. Georges TM (1968) *Acoustic Gravity Waves in the Atmosphere*. Proceedings of the ESSA-ARPA Symposium. US Government Printing Office, Washington
33. Gossard EE, Hooke WH (1975) *Waves in the Atmosphere: Atmospheric Infrasound and Gravity Waves – their Generation and Propagation*. Elsevier, London
34. Goerke VH, Young JM, Cook RK (1965) Infrasonic observations of the May 16, 1963, volcanic eruption on the Island of Bali. *J Geophys Res* 70:6017–6022
35. Hagerty M, Schwartz S, Garces M, Protti M (2000) Analysis of seismic and acoustic observations at Arenal Volcano, Costa Rica, 1995–1997. *J Volcanol Geotherm Res* 101:27–65
36. Ham FM, Park S (2002) A Robust Neural Network Classifier for Infrasound Events using Multiple Array. presented at WCCI-2002, (IJCNN-2002), Honolulu, 12–17 May, pp 2615–2619
37. Harris A, Ripepe M (2008) Synergy of multiple geophysical approaches to unravel explosive eruption conduit and source dynamics – A case study from Stromboli. *Chemie der Erde - Geochemistry* 67(1):1–35
38. Hedin AE, Biondi MA, Burnside RG, Hernandez G, Johnson RM, Killeen TL, Mazaudier C, Meriwether JW, Salah JE, Sica RJ, Smith RW, Spencer NW, Wickwar VB, Virdi TS (1996) Revised global model of upper thermospheric winds using satellite and ground-based observations. *J Geophys Res* 96:7657–7688
39. Hedlin M, Garces M, Bass H, Hayward C, Herrin G, Olson JV, Wilson C (2002) Listening to the Secret Sounds of Earth's Atmosphere. *EOS* 83:557, 564–565
40. Hedlin MAH, Alcoverro B, D'Spain G (2003) Evaluation of rosette infrasonic noise-reducing spatial filters. *J Acoust Soc Am* 114:1807–1820
41. Heil C, Urban M (1992) Sound fields radiated by arrayed multiple sound sources. paper presented at the 92nd Convention of the Audio Engineering Society. Preprint no 3269, Vienna, 24–27 March 1992
42. Hercz AR (1987) *Fundamentals of Sound Ranging*. published by Arthur R Hercz
43. Herrin E, Golden P, Negraru P, Andre W, Bass H, Garces M, Hedlin M, McKenna M, Norris D, Osborne D, Whitaker R (2006) Infrasound Calibration Explosions from Rockets Launched at White Sands Missile Range. 28th Seismic Research Review, Orlando, 19–21 Sept 2006
44. Kajiura K (1970) Tsunami source, energy and the directivity of wave radiation. *Bull Earthq Res Inst* 48:835–869
45. Kim TS, Hayward C, Stump B (2004) Local infrasound signals from the Tokachi-Oki earthquake. *Geophys Res Lett* 31:L20605 doi:10.1029/2004GL021178
46. Larson RJ, Craine LB, Thomas JE, Wilson CR (1971) Correlation of winds and geographic features with production of certain

- infrasonic signals in the atmosphere. *Geophys J R Astr Soc* 26:201–214
47. Le Pichon A, Garcés M, Blanc E, Barthélémy M, Drob DP (2002) Acoustic propagation and atmosphere characteristics derived from infrasonic waves generated by the Concorde. *J Acoust Soc Am* 111:629–641
  48. Le Pichon A, Guilbert J, Vega A, Garcés M, Brachet N (2002) Ground-coupled air waves and diffracted infrasonics from the Arequipa earthquake of June 23, 2000. doi:10.1029/2002GL015052
  49. Le Pichon A, Guilbert J, Vallée M, Dessa JX, Ulziibat M (2003) Infrasonic imaging of the Kunlun Mountains for the great 2001 China earthquake. *Geophys Res Lett* 30(15):1814. doi:10.1029/2003GL017581
  50. Le Pichon A, Blanc E, Drob DP, Lambotte S, Dessa JX, Lardy M, Bani P, Vergnolle S (2004) Infrasonic monitoring of volcanoes to probe high altitude winds. *J Geophys Res* 110:D13106 doi:10.1029/2004JD005587
  51. Le Pichon A, Herry P, Mialle P, Vergoz J, Brachet N, Drob D, Garcés M, Ceranna L (2005) Infrasonic associated with large Sumatra earthquakes and tsunami. *Geophys Res Lett* 32:L19802 doi:10.1029/2005GL023893
  52. Le Pichon A, Mialle P, Guilbert J, Vergoz J (2006) Multistation infrasonic observations of the Chilean earthquake of 2005 June 13. *Geophys J Int* 167(2):838–844. doi:10.1111/j.1365-246X.2006.03190.x
  53. Lin A, Fu B, Guo J, Zeng Q, Dang G, He W, Zhao Y (2002) Coseismic strike-slip and rupture length produced by the 2001 Ms 8.1 Central Kunlun earthquake. *Science* 296:2015–2017
  54. Matoza R, Hedlin M, Garcés M (2007) An infrasonic array study of Mount St. Helens. *J Volcanol Geotherm Res* 160:249–262
  55. McKisic JM (1997) Infrasonic and the infrasonic monitoring of atmospheric nuclear explosions. DOE document PL-TR-97-2123
  56. McGarr A, Vorhis RC (1968) Seismic seiches from the March 1964 Alaska earthquake. US Geological Survey Professional Paper 544-E, pp E1–E43, 1 sheet, scale 1:5,000,000
  57. Mikumo T (1968) Atmospheric pressure waves and tectonic deformation associated with the Alaskan earthquake of March 28, 1964. *J Geophys Res* 73:2009–2025
  58. Mikumo T, Shibutani T, Le Pichon A, Tsuyuki T, Watada S, Garcés M, Fee D, Morii W (2008) Low-Frequency Acoustic-Gravity Waves from Tectonic Deformation Associated with the 2004 Sumatra-Andaman Earthquake (Mw=9.2). *J Geophys Res* 113:B12402, doi:10.1029/2008JB005710
  59. Mutschlecner P, Whitaker R (1998) Infrasonic observations of earthquakes. Tech Rep LA-UR-98-2689, Los Alamos Laboratory, New Mexico
  60. Mutschlecner JP, Whitaker RW (2005) Infrasonic from earthquakes. *J Geophys Res* 110:D01108. doi:10.1029/2004JD005067
  61. National Academy of Sciences (2002) Technical Issues Related to the Comprehensive Nuclear Test Ban Treaty. National Academy Press, Washington, International Standard Book Number 0-309-08506-3
  62. Olson JV, Wilson CR, Hansen R (2003) Infrasonic associated with the 2002 Denali fault earthquake, Alaska. *Geophys Res Lett* 30:N0232195. doi:10.1029/2003GL018568
  63. Park J, Song T-RA, Tromp J, Okal E, Stein S, Roullet G, Clevede E, Laske G, Kanamori H, Davis P et al (2005) Earth's Free Oscillations Excited by the 26 December 2004 Sumatra-Andaman Earthquake. *Science* 308:1139–1144
  64. Sakai T, Yamasato H, Uehira K (1996) Infrasonic accompanying C-type tremor at Sakurajima volcano. *Bull Volcanol Soc Japan* 41:181–185
  65. Shumway RH (2001) Detection and location capabilities of multiple infrasonic arrays. Proceedings of the 23rd NNSA Research Review: Worldwide Monitoring of Nuclear Explosions, pp 160–167
  66. Sutherland L, Bass H (2004) Atmospheric absorption in the atmosphere up to 160 km. *J Acoust Soc Am* 115:1012–1032
  67. Sutherland L, Bass H (2006) Erratum: Atmospheric absorption in the atmosphere up to 160 km. *J Acoust Soc Am* 120:2985
  68. Tahira M, Nomura M, Sawada Y, Kamo K (1996) Infrasonic and acoustic-gravity waves generated by the Mount Pinatubo eruption of June 15, 1991. In: Newhall C, Punongbayan R (ed) Fire and Mud -Eruption and Lahars of Mount Pinatubo, Philippines. Univ. Washington Press, Seattle
  69. Takahashi Y, Koyama Y, Isei T (1994) In situ measured infrasonic at Sapporo associated with an earthquake occurring offshore in southwest Hokkaido on July 12, 1993. *J Acoust Soc Jpn* 15:409–411
  70. Thatcher W (1990) Order and diversity in the modes of circum-Pacific earthquake recurrence. *J Geophys Res* 95:2609–2623
  71. Tsuda K, Steidl J, Archuleta R, Assimaki D (2006) Site-Response Estimation for the 2003 Miyagi-Oki Earthquake Sequence Considering Nonlinear Site. *Response Bull Seismol Soc Am* 96(4A):1474–1482. doi:10.1785/0120050160
  72. Vallée M, Bouchon M (2004) Imaging coseismic rupture in far field by slip patches. *Geophys J Int* 156:615–630
  73. Virieux J, Farra V (1991) Ray-tracing in 3D complex isotropic media: an analysis of the problem. *Geophysics* 56:2057–2069
  74. Virieux J, Garnier N, Blanc E, Dessa JX (2004) Paraxial ray-tracing for atmospheric wave propagation. *Geophys Res Lett* 31:L20106. doi:10.1029/2004GL020514
  75. Weber ME, Donn WL (1982) Ducted propagation of Concorde-generated shock waves. *J Acoust Soc Am* 71:340–347
  76. Willis M, Garcés M, Hetzer C, Businger S (2004) Infrasonic observations of open ocean swells in the Pacific: Deciphering the song of the sea. *Geophys Res Lett* 31:L19303
  77. Winchester S (2004) Krakatoa, The Day the World Exploded. Harper Perennial
  78. Young JM, Greene GE (1982) Anomalous infrasonic generated by the Alaskan earthquake of 28 March 1964. *J Acoust Soc Am* 71:334–339
  79. Yamasato H (1998) Nature of infrasonic pulse accompanying low frequency earthquake at Unzen Volcano. *Japan Bull Volcanol Soc Japan* 43:1–13

## Books and Reviews

- Bouchon M, Bouin M-P, Karabulut H, Toksöz MN, Dietrich M, Rosakis AJ (2001) How fast is the rupture during an Earthquake? New insights from the 1999 Turkey Earthquakes. *Geophys Res Lett* 28:2723–2726
- Donn WL, Balachandran NK (1981) Mount St. Helens eruption of 18 May 1980: Air waves and explosive yield. *Science* 213:539–541
- Garcés M, Iguchi M, Ishihara K, Morrissey M, Sudo Y, Tsutsui T (1999) Infrasonic precursors to a Vulcanian eruption at Sakurajima volcano, Japan. *Geophys Res Lett* 26:2537–2540

- Garcés MA, Hansen RA, McNutt SR, Eichelberger J (2000) Application of wave-theoretical seismoacoustic models to the interpretation of explosion and eruption tremor signals radiated by Pavlof volcano, Alaska. *J Geophys Res* 105:3039–3058
- Hagerty M, Schwartz SY, Protti M, Garcés M, Dixon T (1997) Observations at Costa Rican volcano offers clues to causes of eruptions. *EOS Trans Am Geophys Union* 78:565–571
- Harkrider DG (1964) Theoretical and observed acoustic-gravity waves from explosive sources in the atmosphere. *J Geophys Res* 69:5295–5321
- Johnson JB, Aster RC, Ruiz MC, Malone SD, McChesney PJ, Lees JM, Kyle PR (2003) Interpretation and utility of infrasonic records from erupting volcanoes. *J Volc Geotherm Res* 121:15–63
- Kamo K, Ishihara K, Tahira M (1994) Infrasonic and seismic detection of explosive eruptions at Sakurajima Volcano, Japan, and the PEGASAS-VE early warning system. In: Casadevall T (ed) *Proceedings, 1st International Symposium on Volcanic Ash and Aviation Safety*. 8–12 July 1991. US Geological Survey Bulletin 2047, pp 357–365
- Ripepe M, Poggi P, Braun T, Gordeev E (1996) Infrasonic waves and volcanic tremor at Stromboli. *Geophys Res Lett* 23:181–184
- Tahira M (1982) A study of the infrasonic wave in the atmosphere. II, Infrasonic waves generated by the explosions of the volcano Sakura-jima. *J Meteorol Soc Jpn* 60:896–907
- Tahira M, Ishihara K, Iguchi M (1988) Monitoring volcanic eruptions with infrasonic waves. In: *Proceedings of the Kagoshima International Conference on Volcanoes, Tokyo, 19–23 July 1998*. National Institute for Research Advancement, and Kagoshima, Kagoshima Prefectural Government, p 530–533

## Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of

KARIN A. DAHMEN<sup>1</sup>, YEHUDA BEN-ZION<sup>2</sup>

<sup>1</sup> Department of Physics, University of Illinois at Urbana-Champaign, Urbana, USA

<sup>2</sup> Department of Earth Sciences, University of Southern California, Los Angeles, USA

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Models](#)

[Theoretical Results](#)

[Summary](#)

[Future Directions](#)

[Acknowledgments](#)

[Bibliography](#)

### Glossary

**Mean field theory** A theoretical approximation with an interaction field that has constant strength and infinite range. In mean field approximation every domain interacts equally strongly with every other domain, regardless of their relative distance.

**Critical point** A (phase transition) point in the parameter space of a physical system where the length scale characteristic of its structure, called the correlation length  $\xi$ , becomes infinite and the system displays power law scaling behavior on all available scales. The associated critical power law exponents are universal, i. e. they are independent of the microscopic details of the system.

**Universality** Power law scaling exponents and scaling functions near a critical point are the same for a class of systems, referred to as universality class, independent of the microscopic details. Universal aspects typically depend only on a few basic physical attributes, such as symmetries, range of interactions, dimensions, and dynamics.

**Tuning parameters** Parameters such as disorder, temperature, pressure, driving force etc. that span phase diagrams. Critical values of the tuning parameters describe critical points of the phase diagrams.

**Renormalization group (RG)** A set of mathematical tools and concepts used to describe the change of physics with the observation scale. Renormalization

Group techniques can be used to identify critical points of a system as fixed points under a coarse graining transformation, and to calculate the associated critical power law exponents and the relevant tuning parameters. They can also be used to determine what changes to the system will leave the scaling exponents unchanged, and thus to establish the extent of the associated universality class of the critical point.

**Earthquake quantities** The most common form of earthquake data consists of seismic catalogs that list the time, location, and size of earthquakes in a given space-time domain. The size of earthquakes is usually specified by magnitudes associated with spectral amplitudes of seismograms at a given frequency and site-instrument conditions. The seismic potency and moment provide better physical characterizations for the overall size of earthquakes. Additional important quantities are the geometry of faulting (e. g., strike slip), stress drop at the source region, and radiated seismic energy.

**Seismic potency** A physical measure for the size of earthquakes given by the integral of slip over the rupture area during a seismic event.

**Seismic moment** A physical measure of earthquakes given by the rigidity at the source region times the seismic potency.

**Strike slip fault** A style of faulting involving pure horizontal tangential motion, predicted for situations where the maximum and minimum principal stresses are both horizontal. Prominent examples include the San Andreas fault in California, the Dead Sea transform in the Levant and the North Anatolian fault in Turkey.

### Definition of the Subject

Observations indicate that earthquakes and avalanches in magnetic systems (Barkhausen Noise) exhibit broad regimes of power law size distributions and related scale-invariant quantities. We review results of simple models for earthquakes in heterogeneous fault zones and avalanches in magnets that belong to the same universality class, and hence have many similarities. The studies highlight the roles of tuning parameters, associated with dynamic effects and property disorder, and the existence of several general dynamic regimes. The models suggest that changes in the values of the tuning parameters can modify the frequency size event statistics from a broad power law regime to a distribution of small events combined with characteristic system size events (characteristic distribution). In a certain parameter range, the earthquake model exhibits mode switching between both dis-

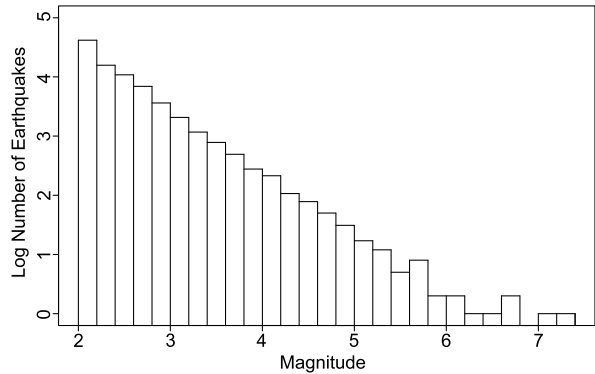
tributions. The properties of individual events undergo corresponding changes in different dynamic regimes. Universal scaling functions for the temporal evolution of individual events provide similar predictions for the earthquake and magnet systems. The theoretical results are generally in good agreement with observations. Additional developments may lead to improved understanding of the dynamics of earthquakes, avalanches in magnets, and the jerky response to slow driving in other systems.

## Introduction

### Global Statistics and Power Law Scaling

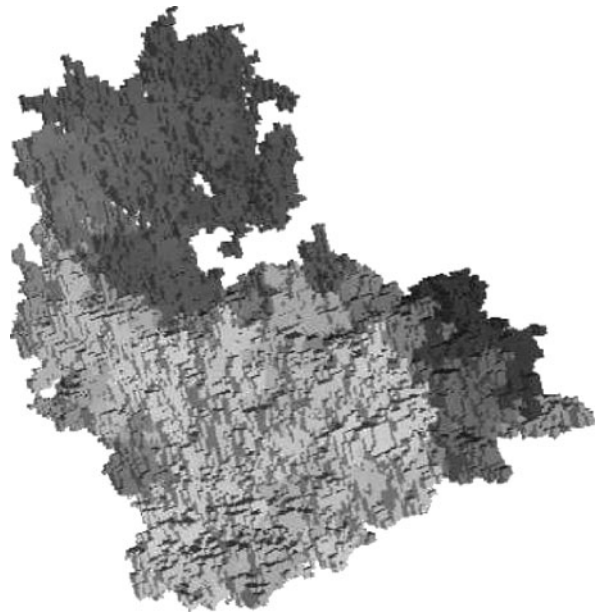
Earthquakes occur in a broad spectrum of sizes, ranging from unnoticeable tremors to catastrophic events. While short term earthquake prediction is still beyond reach, understanding the statistics of earthquakes might facilitate longer term prediction of large earthquakes and statistical estimates of seismic hazard. Gutenberg and Richter [29] found that the frequency of observed regional and global earthquakes versus magnitude forms a regular function over a very large range of scales (see Fig. 1). When the measure for the earthquake size is the seismic potency or moment (see “Glossary”), the frequency size statistics of regional and global earthquakes follow a power law distribution. Precise definitions and details on the seismic potency and moment are given in [1] and [3]. (In this paper we assume a unit nominal rigidity and will therefore use potency and moment interchangeably.) Omori [57] found that the decay rate of aftershocks with time follows a power law distribution. One would expect that there might be a simple explanation for why earthquakes occur in a broad range of sizes and follow regular statistical patterns!

In the last two decades it has become increasingly evident that there are many other systems that respond to slowly changing external conditions with events on extremely large ranges of scales (“crackling noise”). An example of particular interest here involves magnets, which respond to a slowly varying external field by changing their magnetization in a series of bursts or “avalanches” called Barkhausen Noise. Just like earthquakes, these avalanches come in many sizes, ranging from microscopic to macroscopic and are distributed according to a regular function over the entire range. The spectra of the source time function of earthquakes is approximately flat up to a corner frequency related to the rupture size, followed by a power law decay at higher frequencies [1,3]. Similarly, the spectra of the number of spins flipping per time during an avalanche in magnets has high frequency power law decay with a low frequency roll off [68]. For certain values of tuning parameters, earthquake and magnet quantities are



Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of, Figure 1

Histogram of earthquakes with magnitude 2.0 or larger recorded by the Southern California network during 1984–2002. The earthquake catalog is available at <http://www.data.scec.org/research/SHLK.html>



Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of, Figure 2

Fractal spatial structure of a medium sized avalanche of 282 785 domain flips in the 3 dimensional random field Ising model [66]. Fractal structures and power laws are characteristic of systems at their critical point. The shading represents time of the domain flips: the first domains to flip are at the right end of the avalanche, the last towards the left. The short range of the ferromagnetic interactions causes the avalanche to be spatially connected (see [66])

associated with scale invariant functions (power laws). In such cases each individual magnetic avalanche or earthquake slip has fractal spatial structure (see Fig. 2 for mag-



Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of, Table 1

Some scaling features that are similar for magnets and earthquakes. More details on the various properties are given in the indicated sections

	Earthquake system	Magnetic system
Frequency size statistics	Power law near criticality, characteristic distribution away from criticality (Sect. "Results on the Monotonic Version of the Model")	Same as earthquakes (Sect. "Results on the Monotonic Version of the Model")
Scaling of source shape functions	Parabola for moment rate shape of events with fixed duration $T$ in simulations, scaling function skewed to the left for observational data (Sect. "Moment Rate Shapes for Monotonic Models")	Same as earthquakes (Sect. "Moment Rate Shapes for Monotonic Models")
Spatial properties of individual events	Fractal near criticality, compact away from criticality (Sect. "Non-monotonic Models" and Fig. 5)	Same as earthquakes (Sects. "Non-monotonic Models" and "Phase Diagram" and Fig. 2)
Spectral decay of source function of individual events	Flat up to a corner frequency followed by power law decay (Sect. "Introduction")	Same as earthquakes (Sect. "Introduction")

nets and Fig. 5a for earthquakes). Other systems with similar "collective events" of all available sizes include, among others, superconductors, charge density waves, and group decision making [66].

While there are several interesting recent reviews, pointing out the similarities between systems with power law event size distributions, the goal of this paper is to develop in detail some of the connections and analysis methods in earthquake and magnetic systems. Expanding on some of our earlier results, we focus especially on the role of disorder and dynamic changes in the strength threshold as potential tuning parameters to drive the system toward power law scaling behavior or away from it. Table 1 summarizes some of the similarities between magnets and earthquakes that are discussed in this review.

In Sect. "Models" we describe several magnet and earthquake models that are simple enough to make the connections transparent and easy to recognize. In Sect. "Theoretical Results" we review theoretical results obtained from these models and their comparison to experimental or observational data. Finally in Sect. "Summary" we summarize the results and discuss future work, both observationally and theoretically, that can help to improve our understanding of the dynamics of earthquakes and magnets.

## Models

### Models for Barkhausen Noise in Magnets

Hysteresis and avalanches in disordered magnetic materials have been modeled using several variants of the non-equilibrium, zero-temperature random-field Ising model (RFIM), which is one of the simplest models of magnetism, with applications far beyond magnetic systems

(for a review, see [66], and also [15,56,58]). In contrast to some other hysteresis models, like the Preisach model [44] and the Stoner–Wohlfarth model [33], where interactions between the individual hysteretic units (grains) are not included and collective behavior in the form of avalanches is not addressed, in the RFIM the inter grain coupling is an essential feature and cause for hysteresis and avalanche effects.

**The Random Field Ising Model (RFIM)** The *equilibrium* RFIM was originally introduced to study disordered magnetic materials in thermal equilibrium. We study the *nonequilibrium* version, to model hysteresis and avalanches observed *far from thermal equilibrium*. Even though the model is a toy version of the microscopic details in a magnet, near the critical point it correctly describes the large scale behavior of systems with the same general properties such as symmetries, dimensions, interaction ranges and dynamics [15], as follows from renormalization group arguments.

In the RFIM, to each site  $i$  in a simple cubic lattice is assigned a variable  $s_i$ , a so called "spin", which can take two different values,  $s_i = +1$  ("up") or  $s_i = -1$  ("down"). (This corresponds to a real magnet where a crystal anisotropy prefers the magnetic moments or elementary domains, represented by the spins, to point along a certain easy axis.) Each spin interacts with its nearest neighbors on the lattice through a positive exchange interaction,  $J_{nn}$ , which favors parallel alignment. (For the behavior on large scales the exact range of the microscopic interaction is irrelevant, so long as it is finite.) Some variations of the RFIM also include *long range* interactions due to the demagnetizing field and the dipole-dipole interactions. A general form of the Hamiltonian can be written

as [37]

$$\mathcal{H} = - \sum_{nn} J_{nn} s_i s_j - \sum_i H s_i - \sum_i h_i s_i + \sum_i \frac{J_{inf}}{N} s_i - \sum_{\{i,j\}} J_{dipole} \frac{3 \cos(\theta_{ij}) - 1}{r_{ij}^3} s_i s_j, \quad (1)$$

where  $H$  is the homogeneous external magnetic driving field,  $h_i$  is a local, uncorrelated random field, that models the disorder in the system,  $J_{inf}$  is the strength of an infinite range demagnetizing field,  $N$  is the total number of spins in the system, and  $J_{dipole}$  is the strength of the dipole-dipole interactions. The power laws of generated events are independent of the particular choice for the distribution  $\rho(h_i)$  of random fields, for a large variety of distributions. Usually a Gaussian distribution of random fields is used, with a standard deviation (“disorder”)  $R$ . As a simple approximation the model is studied at zero temperature, far from equilibrium, to describe materials with sufficiently high barriers to equilibration, so that temperature fluctuations are negligible on experimental time scales. As the magnetic field is adiabatically slowly changed between  $H = -\infty$  to  $H = +\infty$  two different *local* dynamics have been considered:

(1) in the first (“bulk”) dynamics, each spin  $s_i$  flips while decreasing its own energy. We have studied this dynamics for the original RFIM without long range interactions, i. e. for  $J_{inf} = J_{dipole} = 0$  [15,58]. This dynamics allows for *both* domain nucleation (when a spin  $s_i$  surrounded by equal valued spins flips in the opposite direction), *and* for domain wall motion (when a spin flips on the surface of a preexisting cluster of uniform spins in a background of opposite valued spins). A spin flip can trigger neighboring (or more generally, coupled) spins to flip as well, leading to an avalanche of spin flips, analogous to a real Barkhausen pulse. During an avalanche the external field is kept constant until the avalanche is finished, in accordance with the assumed adiabatic limit. The model is completely deterministic – two successive sweeps through the hysteresis loop produce the exact same sequence of avalanches (since the temperature is set to zero). This dynamics may be appropriate to describe for example hard magnetic materials with strong anisotropies. The analogue earthquake system may be associated with fault regions or fault networks that have strong geometrical and material heterogeneities.

(2) The second dynamics is a “front propagation dynamics” in which only the spins on the edge of an existing front (interface between up and down spins) flip if that decreases their energy. This dynamics can be used to model soft magnetic materials with a single or several noninter-

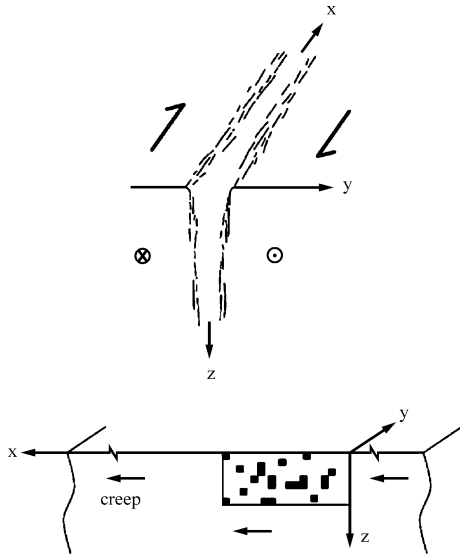
acting advancing domain walls and negligible new domain nucleation, due to antiferromagnetic demagnetizing fields. The front propagation model without long range interactions ( $J_{inf} = J_{dipole} = 0$ ) was originally introduced by Robbins et al. to model fluids invading porous media [43]. The analogue earthquake system for this case may be associated with a single fault zone.

### Simple Models for Inhomogeneous Earthquake Faults

Much of the previous work on simple earthquake models has involved variants of the Burridge–Knopoff (or “slid-erblock”) model, in which complex behavior is generated in a system with many degrees of freedom, and where inertia, friction laws and inherent discreteness play important roles [11,38,61]. These systems appear to exhibit power-law statistics over some range with a cutoff beyond some magnitude, and with most of the slip occurring in larger system-size events. However the understanding of the origin of the power law behavior is limited. Our approach here is to obtain an analytic understanding of a class of models and then to add in various additional features by analytic scaling arguments using tools from the theory of phase transition and the renormalization group, aided by numerical studies. There are interesting related studies using tools from statistical physics [12,63]. Some studies suggest that the power law scaling is connected to a spinodal [34]. Various cellular automata models have also been used for modeling earthquakes [40]. Rather than reviewing a large number of models, we will focus on a subgroup of models, that we found particularly well suited to clarify the connection between earthquake and magnetic systems with a jerky response to slowly changing external conditions.

**The Ben-Zion and Rice Model** A representative of the class of models that we consider is a model developed originally by Ben-Zion and Rice [2,5,6], referred to below as the BZR model. The model assumes that a narrow irregular strike-slip fault zone of horizontal length  $L$  and vertical depth  $W$  may be represented by an array of  $N \sim LW$  cells in a two dimensional planar region of length  $L$  and width  $W$ , with long range interaction, abrupt transitions in the threshold dynamics during failure, and constitutive parameters that vary from cell to cell to model the disorder (offsets etc.) of the fault zone structure (Fig. 3).

The cells represent brittle patches on the interface between two tectonic blocks that move with slow transverse velocity  $v$  in the  $x$  direction at a great distance from the fault. The interaction between cells during slip events is governed by 3-D elasticity and falls off with a distance  $r$



Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of, Figure 3

**Illustration of the Ben-Zion and Rice (BZR) model: projection of a 3D fault zone (top) onto a 2D interface embedded in a 3D elastic halfspace (bottom). The geometrical inhomogeneities of the physical fault zone are modeled by spatially varying constitutive parameters of the brittle patches (see [46])**

from the failure zone as  $1/r^3$ . The cells remain stuck while the stress  $\tau_i$  on each cell is increased gradually as a result of the external loading which grows adiabatically (that is we take the limit  $\nu \rightarrow 0$ ). When the stress on a cell  $i$  reaches its local failure threshold  $\tau_{s,i}$ , the cell slips until the stress is reduced to its local arrest stress  $\tau_{a,i}$ . Both failure stress and arrest stress are distributed according to some bounded probability distribution. The stress drop resulting from a cell failure is redistributed to the other cells according to the long range elastic stress transfer function. The resulting stress increase on the other cells can cause some of them to slip as well, leading to an avalanche of cell slips, or a model earthquake. A review of extensive numerical simulations with various versions of the BZR model, in relation to observed features of seismicity, criticality, and other dynamic regimes, is given in [75].

**Dynamical Weakening** The model includes dynamic weakening effects during the failure process [2,5,6]: after an initial slip in an earthquake, the strength of a failed cell is reduced to a *dynamical* value:

$$\tau_{d,i} \equiv \tau_{s,i} - \epsilon(\tau_{s,i} - \tau_{a,i}), \quad (2)$$

with  $0 \leq \epsilon \leq 1$  parametrizing the relative importance of the dynamical weakening in the system. This weakening

represents the transition from static friction to dynamic friction during the rupture. The strength of a failed cell remains at its dynamic value throughout the remainder of the earthquake. In the time intervals between earthquakes all failure thresholds heal back to their static value  $\tau_{s,i}$ .

**Dynamical Strengthening** The model can be expanded further to include dynamic strengthening represented by  $\epsilon < 0$ . Multidisciplinary observations indicate [7] that brittle failure of rock has an initial transient phase associated with strengthening, distributed deformation, and creation of new structures. Detailed frictional studies also show an initial strengthening phase associated with the creation of a new population of asperity contacts [3,19]. In mean field studies of our model (Fig. 3) discussed in Sect. “Results on Aftershocks”, we associate  $\epsilon < 0$  with regions off the main fault segments that are in an early deformation stage. The events that are triggered as the failure stresses are lowered back in the following weakening period are referred to as *aftershocks*. The Omori law [3,69,70] is obtained if we assume that the increased failure stress thresholds  $\tau_{f,i}$  are slowly lowered with time as  $\log(t)$  towards their earlier static values  $\tau_{s,i}$ , and that the stresses are distributed over a wide range of values [46].

**Related General Continuum Equations of Motion** The above model is a special case of a more general class of models for infinite systems driven by a constant drive force  $F$  [27]. We consider general equations of motion of the form:

$$\eta \partial u(\mathbf{r}, t) / \partial t = F + \sigma(\mathbf{r}, t) - f_R[u(\mathbf{r}, t), \mathbf{r}, \{u(\mathbf{r}, t' < t)\}] \quad (3)$$

where

$$\sigma(\mathbf{r}, t) = \int_{-\infty}^t dt' \int d^d r' J(\mathbf{r} - \mathbf{r}', t - t') \cdot [u(\mathbf{r}', t') - u(\mathbf{r}, t)] \quad (4)$$

is the stress and  $f_R$  is a quenched random “pinning” force crudely representing inhomogeneities in the friction, asperities, stepovers etc., which in general can depend on the local past history (e. g. as in velocity dependent friction). The dynamical variables  $u(\mathbf{r}, t)$  are assumed to represent the discontinuity across the fault plane in the component of the displacement in the direction of slip. The dynamics depend on the local history dependence of the pinning force, the stress transfer function  $J(\mathbf{r}, t)$ , and the coefficient  $\eta$  that represents the fault impedance. (In an elastic medium, the impedance depends on mass density, the

elastic parameters, and directional parameters [1].) Equation (3) can be considered a continuum description of the rules of the BZR model. Integrating out the degrees of freedom due to the bulk material on either side of the  $d=2$  dimensional fault plane leaves us with effective long range static stress transfer:  $J_s(r) \equiv \int dt J(r, t) \sim 1/r^{d+\Gamma} \sim 1/r^3$ . For a planar fault in an elastic half space,  $d = 2$  and  $\Gamma = 1$  [2,6]. The correlations in  $f_R$  are generally assumed to be short-range in  $u$  and  $\mathbf{r}$ . (For results on the BZR model with long range correlations in the disorder, see [2,27,77] and Sect. “Theoretical Results”.) In a version of the BZR earthquake model with a constant driving force  $F$ , the loading may be replaced by driving through a weak spring with spring constant  $K \sim 1/L$  coupled to the slowly moving continents far away (i. e. replacing  $F$  in Eq. (3) by  $F(\mathbf{r}, t) = K[v t - u(\mathbf{r}, t)]$ , with  $v \rightarrow 0$ ).

**Monotonic Models** Substantial simplifications occur if  $f_R$  is history independent and  $J(\mathbf{r}, t) \geq 0$  for all  $(\mathbf{r}, t)$ ; leading to *monotonic* models [27]. Related monotonic models have been studied extensively in various other contexts [25,53]. Examples include elastic depinning models for contact lines, vortex lines, liquids invading porous materials, and elastic charge density waves. Their crucial simplifying feature is that the steady state velocity  $\bar{v} \equiv \langle \partial u / \partial t \rangle$  is a history independent function of  $F$  [48]. In the context of the BZR model this corresponds to the case with zero weakening ( $\epsilon = 0$ ) and non-negative  $J$ . A crucial feature of monotonic models is that the slip profile  $\Delta u(\mathbf{r})$  of a quake is *independent of the dynamics* [48]. However, several interesting dynamical issues discussed below are associated with the effects left out of the monotonic models that can make this feature break down.

**Non-monotonic Models** (a) *Weakening*: We first consider including some weakening effects of sections which have already slipped in a given quake. This is best studied in the discrete model. In analogy to the dynamic weakening in the BZR model discussed above, we choose [27]

$$f_R = \tilde{f}_R[u(\mathbf{r}), \mathbf{r}] \{1 - \epsilon \Theta[u(\mathbf{r}, t) - u(\mathbf{r}, t - T)]\} \quad (5)$$

with  $T$  a cutoff time much longer than the duration of the largest quakes, but much smaller than the interval between the quakes. Here  $\Theta(x)$  is the Heavyside step function. As mentioned, the case  $\epsilon > 0$  represents the difference between static and dynamic friction. The effects of small weakening ( $\epsilon > 0$ ) can be analyzed perturbatively (see Sect. “Theoretical Results”).

(b) *Stress Pulses*: A similar but more subtle effect can be caused by stress pulses that result from non-posi-

itive  $J(\mathbf{r}, t)$ ; these arise naturally when one includes elastodynamic effects. We consider

$$J(\mathbf{r}, t) \sim \frac{\delta(t - \frac{r}{c})}{r^{d+\Gamma}} + \frac{\alpha \delta'(t - \frac{r}{c})}{c r^{d+\gamma}} \quad (6)$$

with  $c$  the sound speed,  $\delta(t)$  the Dirac delta distribution, and  $\delta'(t) = d\delta(t)/dt$ . The scalar approximation to elasticity in a half space corresponds to  $d = 2$ ,  $\Gamma = 1$ ,  $\gamma = 0$ , and  $\alpha = 1$  [27]. If a region slips forward, the stress at another point first has a short pulse at the sound arrival time from the second term in Eq. (6), and then settles down to its smaller static value, i. e. it is non-monotonic. The magnitude of these stress pulses and their duration is set by various aspects of the models; for example larger  $\eta$  in Eq. (3) implies weaker stress pulses as the local motion will be slower.

## Theoretical Results

Both the magnet and earthquake models of the previous section are capable of producing a large range of power law scaling of event sizes, and related scale invariant quantities in response to a slowly varying driving force or field. This section highlights similarities between these different physical systems and attempts to explain them.

### The Universality Class of the BZR Model

We first review results for the simplified monotonic case, starting with scaling relations for driving with fixed force and far field plate motion and continuing with moment rate shapes. We then discuss additional results associated with non monotonic versions of the model, including mode-switching and aftershocks.

### Results on the Monotonic Version of the Model

**General results: Depinning transition** As mentioned above, substantial simplifications occur for the monotonic version of the model, i. e. if  $f_R$  is history independent and  $J(\mathbf{r}, t) \geq 0$  for all  $(\mathbf{r}, t)$ . In [25,27,53] it is shown that for  $F$  greater than a critical force  $F_c$  the displacement grows continuously in a “sliding state” for which the mean velocity  $\bar{v} \equiv \langle \partial u / \partial t \rangle \sim (F - F_c)^\beta$ . Here  $\beta$  is a universal exponent that is independent of the microscopic details of the system. It only depends on a few fundamental properties, such as symmetries, spatial dimensions  $d$ , range of interactions, etc. [53]. Long time dynamic properties such as  $\beta$  depend in addition on the small  $\omega$  dependence of  $J(\mathbf{q}, \omega)$  [54].

For  $F$  less than the critical force  $F_c$ , the mean velocity is  $\bar{v} \equiv 0$ . If  $F$  is adiabatically slowly increased to

towards  $F_c$ , the system moves from one metastable configuration to another by a sequence of “quakes” of various sizes. The “quakes” can be characterized by their radius  $R$ , the  $d$ -dimensional area  $A$  which slips (by more than some small cutoff), their potency or moment  $M \equiv \int_A d^d \mathbf{r} \Delta u(\mathbf{r})$ , a typical displacement  $\Delta u \sim M/A$ , and a duration  $\tau$ . The critical force  $F_c$  marks a second order phase transition point. Such phase transitions are typically associated with power law scaling behavior.

In the class of earthquake models with long range interactions along the fault involving the static stress transfer  $J_s(r) \equiv \int dt J(r, t) \sim 1/r^3$ , the equations are very similar to those of a model for contact line depinning studied in ref. [25]. Using renormalization group methods it was shown in [25] that for a physical two dimensional interface (or “fault”) in a 3 dimensional elastic half space, these long range interactions are so long that the scaling behavior near  $F_c$  is correctly described by mean field theory (up to logarithmic corrections, since  $d = 2$  is the “upper critical dimension”). The main assumption in mean field theory is that the spatial and temporal fluctuations in the displacement field  $u(\mathbf{r}, t)$  are so small that the local displacement  $u(\mathbf{r}, t)$  can be replaced by a time dependent spatial average  $u(t)$ , which then needs to be determined self consistently from the behavior of the neighboring regions that contribute to the stress at a chosen point  $\mathbf{r}$  [26]. The same mean field equations are obtained when the long range interaction is approximated to be constant in space  $J(\mathbf{r}, t) = J_{mft}(t)/(LW)$ . With this approximation Eqs. (3) and (4) become

$$\eta \partial u(\mathbf{r}, t) / \partial t = F + \sigma_{mft}(\mathbf{r}, t) - f_R[u(\mathbf{r}, t), \mathbf{r}, \{u(\mathbf{r}, t') < t\}] \quad (7)$$

where

$$\sigma_{mft}(\mathbf{r}, t) = \int_{-\infty}^t dt' J_{mft}(t - t') [u(t') - u(\mathbf{r}, t)] \quad (8)$$

and the self consistency requirement is

$$\int u(\mathbf{r}, t) d^2 r / (LW) = u(t) \quad (9)$$

Many scaling exponents and scaling functions can be calculated exactly in mean field theory by solving these simplified equations of the model. In [15,26,42], several illustrative examples are given for solving similar self consistent mean field theories. There are various approaches that one may use, ranging from numerical simulations to analytical expansion and scaling analysis near a phase transition point where universal power law scaling occurs. The

approach of choice to solve the mean field equations depends on the quantity under consideration. To obtain exact results for the scaling behavior of the frequency size statistics of earthquake or avalanche events, a fairly simple approach is to use a discrete version of the model in which we treat the fault as a discrete set of dislocation patches, coupled to a mean displacement and an external driving force that slowly increases with time. (The stress  $\tau_i$  at each patch is given by Eq. (16) of Sect. “Mode-Switching” below.) As shown in [17], the sequence that describes the distance from failure of the rescaled stress variables resembles a biased random walk. The scaling behavior of the resulting random walk is known exactly from the literature. Using this mapping it then becomes straightforward to derive universal scaling predictions for the mean field earthquake frequency size distribution [17].

Furthermore, as shown in [13,25], their (and thus also our) model have the same scaling behavior as a front propagation model for a two dimensional domain wall in a soft magnet with long range dipolar magnetic interactions, driven by a slowly changing external field (see Sect. “Models”). A flipping spin in the magnet model corresponds to a slipping dislocation patch in the earthquake model. The long range elastic interactions in the earthquake model are similar to the long range dipolar magnetic interactions in the magnet model. The driven two dimensional magnetic domain wall in the (three dimensional) magnet model corresponds to the driven two dimensional earthquake fault in a three dimensional elastic half space. Since the scaling behavior of the earthquake model and that of [25] and [73] are identical, we may simply copy their results and translate them into quantities that can be extracted from seismic data. Using tools from phase transitions, such as the renormalization group (RG), near the critical force the following scaling results were derived by [13,25,53,54,73] and others:

$$\begin{aligned} \Delta u &\sim R^\xi, \\ A &\sim R^{d_f} \text{ with } d_f \leq 2 \text{ a fractal dimension,} \\ M &\sim R^{d_f + \xi}, \\ \text{and } \tau &\sim R^z. \end{aligned}$$

The differential distribution  $P(M)$  of moments  $M$  is shown in [25,54] and [26,27] to scale as

$$P(M) dM \sim dM / M^{1+B} \rho_\infty(M/\hat{M}) \quad (10)$$

with  $\rho_\infty$  a universal scaling function which decays exponentially for large argument. The cutoff  $\hat{M}$  for large moments is characterized by a correlation length – the largest likely radius –  $\xi \sim 1/(F_c - F)^\nu$  with  $\hat{M} \sim \xi^{d_f + \xi}$ .

In the same references it is shown that in mean-field theory,  $B = 1/2$ ,  $1/\nu = 1$ ,  $z = 1$  and the quakes are fractal with displacements of order the range of correlations in  $f_R(u)$ , i. e.  $\xi = 0$ .

These mean-field exponents are valid for a  $d = 2$  dimensional planar fault in a three dimensional elastic half space [7], since the physical fault operates at the upper critical dimension. As usual, at the upper critical dimension, there are logarithmic corrections to mean-field results. Using renormalization group methods one can calculate these corrections [27] and finds barely fractal quakes with  $A \sim R^2/\ln R$  so that the fraction of the area slipped decreases only as  $1/\ln r$  away from the ‘‘hypocenter’’. The typical slip is  $\Delta u \sim (\ln R)^{1/3}$  so that  $M \sim R^2/(\ln R)^{2/3}$ . The scaling form of  $P(M)$  is the same as Eq. (10) with the mean-field  $\rho_\infty$ , although for  $M \ll \hat{M}$ ,  $P(M) \sim (\ln M)^{1/3}/M^{3/2}$  so that  $B$  will be virtually indistinguishable from  $1/2$  [27]. A similar form of moment distribution and exponent value  $B = 1/2$  were obtained also for a critical stochastic branching model [71].

*More realistic driving of a fault* We now consider more realistic drive and finite-fault-size effects. As mentioned, driving the fault by very slow motion far away from the fault is roughly equivalent to driving it with a weak spring, i. e. replacing  $F$  in Eq. (3) by  $F(\mathbf{r}, t) = K[\nu t - u(\mathbf{r}, t)]$ . With  $\nu \rightarrow 0$  the system must then operate with the spring stretched to make  $F(\mathbf{r}, t) \lesssim F_c$  at least on average, to ensure  $\bar{v} = 0$ ; depending on the stiffness of the spring, it will actually operate just below  $F_c$ , as shown below. If in constant force drive the force is increased by a small amount  $\Delta F$ , the average resulting slip per area,  $\langle \Delta u \rangle \equiv \sum_i \Delta u_i/(LW)$  is given by the total potency/moment per total area  $M \equiv \int d^d \mathbf{r} \Delta u(\mathbf{r})/(LW)$ . The total moment per area observed in response to a small force increase equals the number  $n\Delta F$  of earthquakes per area that are triggered by the increase  $\Delta F$ , multiplied with the average observed moment of a single earthquake  $\langle M \rangle = \int MP(M)dM$ . The result is

$$\langle \Delta u \rangle = n\Delta F \int MP(M)dM \quad (11)$$

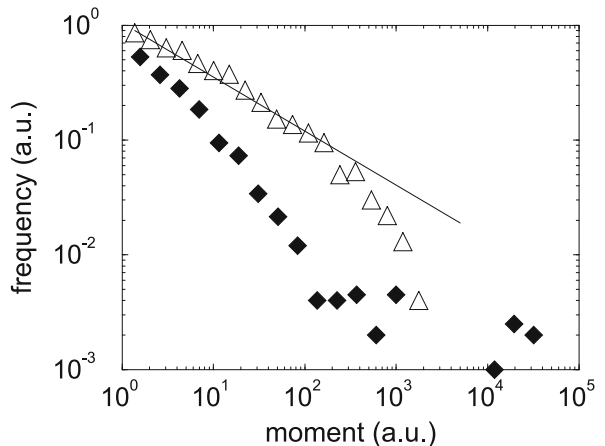
where  $n$  is the number of quakes per unit area per force increase  $\Delta F$ . It has been shown that  $n(F)$  is non-singular at  $F_c$  [53], so it can be treated like a constant near  $F_c$ . Plugging in Eq. (10) and the scaling laws written above and below that equation, we obtain

$$\langle \Delta u \rangle \sim \Delta F \xi^{(2\tilde{F} + \xi)(1-B)} \sim \Delta F \xi \quad (12)$$

for our case where mean field results can be used for the critical exponents. For consistency, we must have in steady

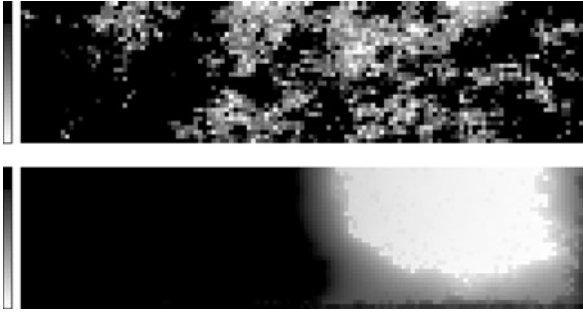
state with the spring drive,  $K\nu\Delta t = \Delta F = K\Delta u$  so that the system will operate with a correlation length  $\xi \sim 1/K^{1/\tilde{F}}$ , i. e.  $1/K$  for our case. For a fault section with linear dimensions of order  $L$ , drive either from uniformly moving fault boundaries or from a distance  $\sim L$  perpendicularly away from the fault plane will be like  $K \sim 1/L$  so the power-law quake distribution will extend out to roughly the system size  $\xi \sim L$ . For smaller quakes, i. e.  $R \ll L$ , the behavior will be the same as in the infinite system with constant  $F$  drive, but the cutoff of the distribution of moments will be like Eq. (10) with a different cutoff function  $\rho$  that depends on the shape of the fault, how it is driven, and the boundary conditions.

We have tested these conclusions numerically by simulating the BZR model, which is a discrete space, time, and displacement version of a monotonic Eq. (3), with quasistatic stress transfer appropriate for an elastic half space [2,5]. The slip,  $u$ , is purely in the horizontal direction along the fault and  $f_R[u(\mathbf{r})]$  is a series of equal height spikes with spacings which are a random function of  $\mathbf{r}$ . When  $\sigma(\mathbf{r}, t) > f_R[u(\mathbf{r}, t)]$ ,  $u(\mathbf{r})$  jumps to the next spike. This provides a way of implementing the random stress drops of the BZR model. The boundary conditions on the bottom and sides are uniform creep or slip ( $u = \nu t$ ) with infinitesimal  $\nu$  – and stress free on the top (Fig 3). The statistics of the moments of the quakes are shown by the triangles in Fig. 4. Although the uncertainties are appreciable, relatively good agreement is found with the



Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of, Figure 4

Histograms of moments for a simulation of a rectangular fault with  $32 \times 128$  cells for the discrete monotonic quasistatic model (with arbitrary units (a.u.)). Triangles: without dynamical weakening ( $\epsilon = 0$ ). Diamonds: with dynamic weakening of  $\epsilon = 0.95$ . ( $\epsilon$  is defined in Eq. (5).) The straight line indicates the predicted slope  $B = 1/2$  (from [27])



Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of, Figure 5

Distribution of horizontal slip,  $u$ , along a fault with  $32 \times 128$  cells for a *single* large quake event. Lighter shading represents larger slip during the quake. *Top*: almost fractal quake with a total moment of 1750 (and 1691 cells failing) for the monotonic model without any dynamical effects ( $\epsilon = 0$ ). *Bottom*: “crack like” quake with a total moment of 16922 (and 2095 cells failing) for the model with dynamic weakening ( $\epsilon = 0.95$ ). In both cases the system is driven by horizontally creeping fault boundaries (*sides and bottom*) while the top boundary is free (from [27])

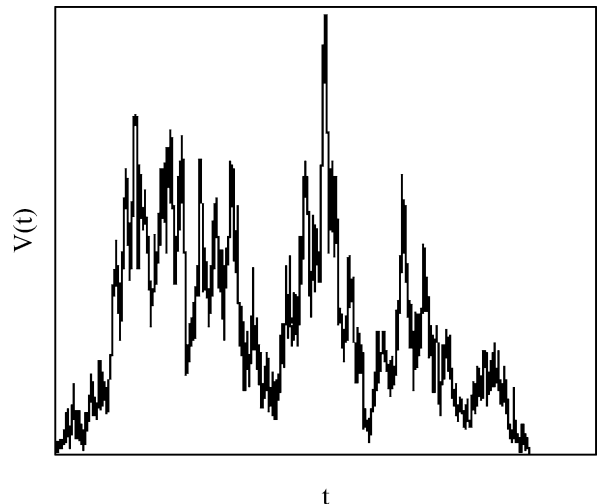
prediction  $B = 1/2$ . One typical large quake is illustrated in Fig. 5 (top); it appears almost fractal as predicted, and tends to stay away from the bottom and sides due to the specific loading that we chose. The ratios of the moments of quakes to their areas have been studied and found to grow only very slowly with the area, as predicted from the logarithmic corrections listed below Eq. (3). This is in striking contrast to earthquakes in conventional crack models which are compact (Fig. 5 (bottom)) and have  $\Delta u \sim R$  (i. e.  $\zeta = 1$ ), so that  $M/A \sim \sqrt{A}$ . As discussed by [8], however, the scaling  $M \sim A$  appears to be consistent with observational results for small earthquakes which presumably propagate and are arrested in rough stress fields. More observational data on the scaling of the moment  $M$  with the slipping area  $A$  for smaller earthquakes would be highly desirable to test this prediction more precisely.

Because the system is at its critical dimension, the cut-off function  $\rho$  of the moment distribution appropriate to the boundary conditions, as well as various aspects of the shapes and dynamics of quakes can be computed using tools from the theory of phase transitions [26,27]. For quasistatic stress transfer,  $J(\mathbf{r}, t) \sim \delta(t)/r^3$ , in the infinite system the quake durations are found to scale as  $\tau \sim R^z$  with  $z = 1$  for a  $d = 2$  dimensional fault, with logarithmic corrections [25]. (A more physical dynamics with sound-travel-time delay has slower growth of the quakes with  $z = 1$  in all dimensions.) Due to the geometrical disorder included in the model, in either case the growth will be very irregular – including regions starting and stop-

ping – in contrast to crack models and what is often assumed in seismological analysis of earthquakes on more regular faults. Similar fractal-like quakes were simulated by Zöller et al. [75,76], for a quasi-dynamic version of the BZR model that includes stress redistribution with a finite communication speed.

**Moment Rate Shapes for Monotonic Models** In both magnet and earthquake models it has been shown that there are not just universal scaling exponents but also some experimentally accessible universal scaling functions [66]. By comparing theoretical predictions for these functions to experiments or observations, one can often test models much more accurately than by merely comparing a finite set of discrete exponents. Two such functions were first discovered for Barkhausen Noise in magnets [47,66]. The analogy between magnets and earthquakes then lead to the development of the corresponding functions for earthquakes. For slowly driven magnets, consider the time history  $V(t)$  of the number of domains flipping per unit time (Barkhausen train). It is called  $V$  because it is usually measured as a voltage in a pickup coil. An example of a Barkhausen train for a single avalanche is shown in Fig. 6.

The voltage function  $V(t)$  in magnets is the analogue of the moment rate  $dm/dt(t)$ , or the slip per unit time for earthquakes. Recent analysis allowed researchers to ob-



Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of, Figure 6

**Voltage train of a typical large avalanche.** Note that the voltage fluctuates drastically and the avalanche nearly stopped several times (from [37]). The analogous moment rate time trace for earthquakes (though measured with lower resolution) is shown in the right inset marked “RAW” of Fig. 8

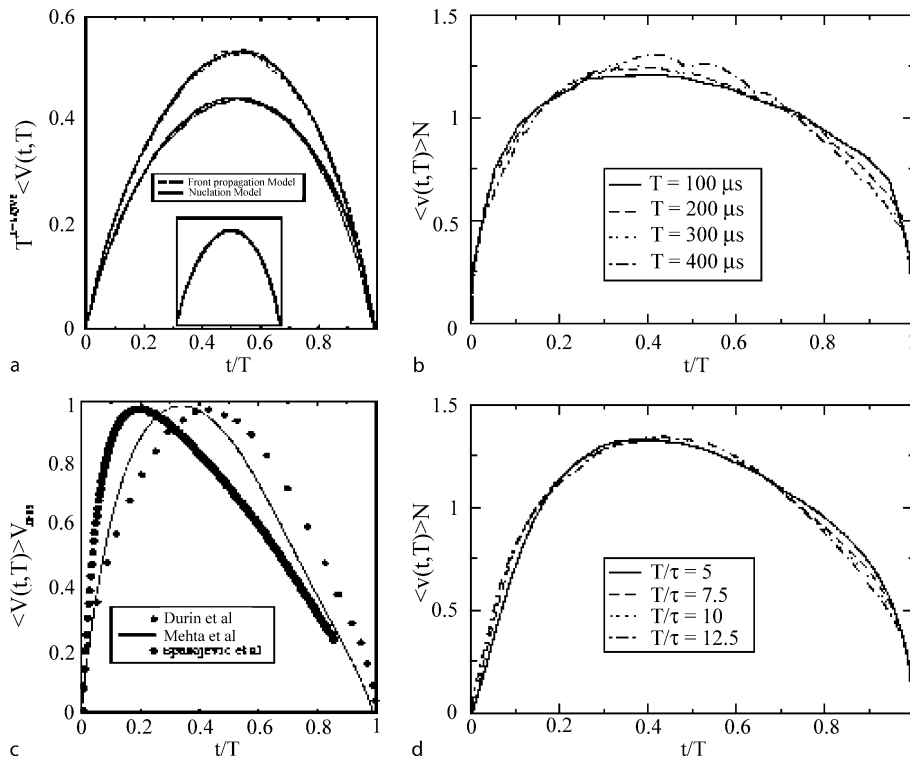
tain the moment rate  $dm_0(t)/dt$ , during the propagation of earthquake rupture for hundreds of large seismic events recorded on global networks [9,31]. The moment rates shown below are derived from inversions of teleseismically recorded seismograms on a global seismic network [62]. (The frequency-moment distribution,  $D(M_0) \sim M_0^{-1-\beta}$  of the observed data [9] has three decades of scaling and an exponent of  $\beta = 1/2 \pm 0.05$ , in close agreement with the BZR model near  $\epsilon = 0$  [46].)

For both magnets and earthquakes there are large fluctuations in  $V(t)$  and  $dm/dt(t)$  respectively (Fig. 6). However averaging the signal over many avalanches, leads to typical shapes. Figure 7 shows the average over all avalanches of fixed duration  $T$ ,  $\langle V \rangle(T, t)$  obtained from

simulations of two variants of the RFIM (a), and from three different Barkhausen noise experiments (b). Figure 8 shows  $\langle dm/dt \rangle(T, t)$  obtained for the BZR earthquake model and derived from earthquake observations respectively. The renormalization group and scaling theory [66] predict that for a self similar system at a critical point with power law size and duration distributions for avalanches, there are self similar average avalanche profiles. As shown in [46,66] one finds

$$\langle dm/dt \rangle(T, t) \sim T^{b'} g(t/T) \quad (13)$$

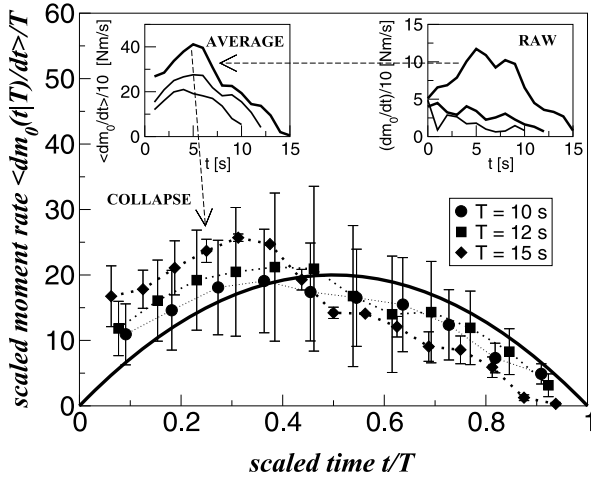
where the function  $g(x)$  is a universal scaling prediction and  $b' \equiv 1/(\sigma \nu z) - 1 = 1$  for the BZR earthquake model (as obtained from mean field theory). The corresponding



Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of, Figure 7

**a Theoretical average avalanche shape scaling functions** for fixed avalanche durations  $T$  denoted with  $g(t/T)$  in the text, for the nucleation and the front propagation RFIM [47]. The overall height is non universal, the curves for the two models are otherwise extremely similar. The front propagation model has  $1/\sigma \nu z = 1.72$  and, the nucleation model has  $1/\sigma \nu z = 1.75$  in this collapse. The *inset* shows the two curves rescaled to the same (non universal) height: the two curves are quantitatively different, but far more similar one to another than either is to the experimental curve in b. **b Experimental average pulse shapes from three different experiments** for fixed pulse duration, as measured by three different groups [21,23,47,67]. Notice that both theory curves are much more symmetric than those of the experiments. Notice also that the three experiments do not agree. At first this result represented a serious challenge to the idea about universality of the dynamics of crackling noise [66]. **c Pulse shape asymmetry experiment** [72]. Careful experiments show a weak but systematic duration dependence in the collapse of the average Barkhausen pulse shape. The longer pulses (larger avalanches) are systematically more symmetric (approaching the theoretical prediction). **d Pulse shape asymmetry theory** [72]. Incorporating the delay effects of eddy currents into the theoretical model produces a similar systematic effect. The non-universal effects of eddy currents are in principle irrelevant for extremely large avalanches (from [64])





Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of, Figure 8

A collapse of averaged earthquake pulse shapes,  $\langle dm_0(t|M_0)/dt \rangle$  with a duration of  $T$  (seconds) within 10% (given in legend), is shown. The collapse was obtained using the mean field scaling relation [37]:  $\langle dm_0(t|T)/dt \rangle \sim g(t/T)$ . In order to obtain each collapsed pulse shape, two to ten earthquakes were averaged for each value of  $T$ . In our mean field theory the universal scaling function is  $g_{mf}(x) = Ax(1-x)$  with  $x = t/T$ . We plot this functional form (bold curve) with  $A = 80$ . Note the apparent asymmetry to the left in the observed data while the theoretical curve is symmetric around its maximum. *Inset*: The raw data and the averaged data (before collapsed) (from [46])

value for  $b'$  for magnets in three dimensions is smaller – the values used for the corresponding collapses can be read off for the different versions of the RFIM from the caption of Fig. 7.

Based on universality one would expect these theoretical predictions to agree with experimental results, apart from an overall shift in time and voltage or moment rate scales. For the moment rate of earthquakes this means

$$\langle dm/dt \rangle_{\text{observation}}(T, t) = A \langle dm/dt \rangle_{\text{theory}}(T/B, t/B) \quad (14)$$

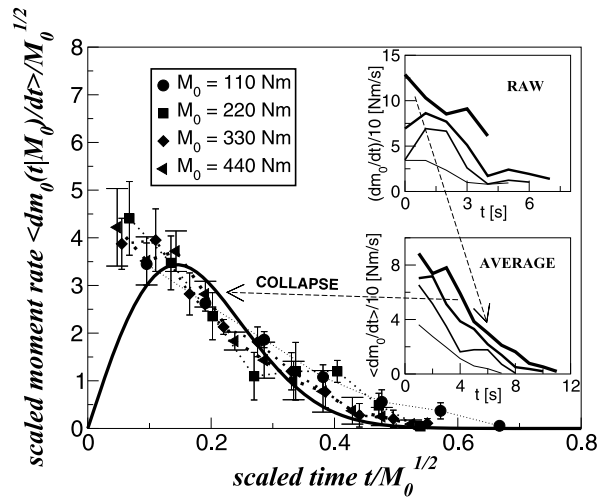
for some rescaling factors  $A$  and  $B$ , and similarly for the average voltage  $\langle V \rangle(t, T)$  in magnets. In both cases the theory predicts a symmetric looking profile. The mean field prediction for  $g(x)$  is in fact a parabola [46,66] – the theoretical prediction thus is that events grow as quickly as they decay. As seen in Figs. 7b and 8 the experimental/observational profile in both cases, however, appear skewed – the real events tend to grow more quickly than they decay! A similar asymmetry has also been observed in avalanches associated with plastic deformation [39].

For magnets this apparent disagreement has been resolved by taking greater account of a microscopic detail involving eddy currents that had been neglected by previ-

ous models. Eddy currents are transient current loops that arise in conducting magnets in response to the reorientation of a magnetic domain. These currents temporarily prevent neighboring domains from being triggered to realign in the same direction in an avalanche of domain reversals. The eddy currents decay after a microscopic time  $\tau$  given by the resistance of the material. Their delay effect thus also decays after a time  $\tau$ . If the avalanche duration is large compared to  $\tau$  this effect is negligible and the mean profile approaches the predicted symmetrical shape (see Fig. 7c and d).

The source of asymmetry in the mean moment rate profile may be similar for earthquakes [16]. It has been suggested that triggering delays – arising from a noticeable earthquake nucleation time, or an increase in the failure threshold during the formation of new cracks and subsequent weakening as rock damage increases – could be responsible for aftershocks that often follow large earthquakes [46]. On long time scales a large mainshock with smaller aftershocks can be seen as a similar asymmetry to that seen in magnets, possibly with a similar explanation.

There is a second scaling function that may be extracted from the same data: Fig. 9 shows the average over



Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of, Figure 9

A collapse of averaged earthquake pulse shapes,  $\langle dm_0(t|M_0)/dt \rangle$ , with the size of the moment  $M_0$  in Newton meters within 10% of each size given in the legend respectively. In order to obtain each collapsed moment rate shape, five to ten earthquakes were averaged for each value of  $M_0$ . The collapse was obtained using the mean field scaling relation [27]:  $\langle dm_0(t|M_0)/dt \rangle / M_0^{1/2} \sim f(t/M_0^{1/2})$ . In our mean field theory the universal scaling function is  $f_{mf}(x) = Axe^{-Bx^2/2}$  where  $x = t/M_0^{1/2}$ . We plot this functional form (bold curve) with  $A = 4$  and  $B = 4.9$ . *Inset*: The raw data and the averaged data (before collapsed) (see [46])

all earthquakes of fixed total moment  $M(dm/dt)(M, t)$ , both for observations and the BZR model prediction. As shown in [46,66] the theory predicts

$$(dm/dt)(M, t) \sim M^{1/2}q(t/M^{1/2}) \tag{15}$$

where the universal scaling function  $q(x) = Ax \exp -Bx^2/2$  and the universal exponents are obtained from the mean field theory for the BZR earthquake model. A comparison between prediction and observational results for this scaling function is shown in Fig. 9.

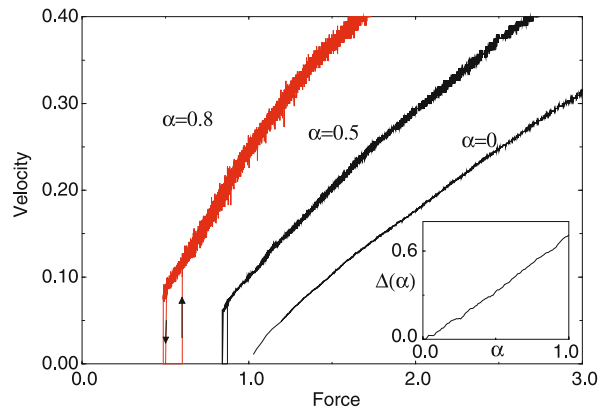
Clearly, more data, especially for small earthquakes, are needed to decrease the statistical error bars of the observational data and determine the degree of agreement between theory and observations. An alternative scaling approach to moment rate data was given by [31] and a comparison between both approaches is discussed in [46].

**Non-Monotonic Models** We first consider including weakening of the cell failure threshold by an amount  $\epsilon$  for sections which have already slipped in a given quake. This crudely models the difference in static versus dynamic friction (see Sect. “Models”, Eq. (5)). In between quakes all thus weakened thresholds heal back to their static strength. The effects of small weakening can be analyzed perturbatively.

With  $\epsilon = 0$ , consider a quake of diameter  $R_1 (\ll L$  or  $\xi)$ , with moment  $M_1$  and area  $A_1$ : i. e.  $A_1$  sites have slipped. If a small  $\epsilon$  is turned on at the end of the quake, all slipped sites that are within  $\epsilon$  of slipping will now slip again – this will be  $N_2^{ex} \sim \epsilon A_1$  sites. The simplest justifiable guess is that each of these will cause an approximately independent secondary quake. The total moment of these secondary quakes will be dominated by the largest one, so the extra moment will be  $M_2^{ex} \sim (\epsilon A_1)^{1/B}$ . (For a very large or infinite fault this is obtained from  $1 = N_2^{ex} \int_{M_2^{ex}}^{\infty} P(M)dM$ , and inserting Eq. (10).) If  $M_2^{ex} \ll M_1$  this process can continue but will not increase the total moment substantially. If  $M_2^{ex} \sim M_1$ , however, the process can continue with a larger area  $A_2$  and hence a larger  $M^{ex}$ , leading to a catastrophic runaway event. From the above exponent relations and scaling laws we obtain  $B = 1/2$  and  $A \sim M$ , so that for any  $\epsilon$ , for large enough  $M_1$ ,  $M_1 \gtrsim M_D \sim \epsilon^{-2}$ ,  $M_2^{ex}$  will be comparable to  $M_1$  and the quake will become much larger (runaway). In the force driven infinite system for  $F \lesssim F_c$ , quakes of size  $\xi$  will runaway and become infinite if  $\xi > \epsilon^{-1}$ . Since  $\xi \sim (F - F_c)^{-\nu}$  and  $1/\nu = 1$ , this will occur for  $F_c - F < C_w \epsilon$  with some constant  $C_w$ . This result is very intuitive and justifies a posteriori the assumptions leading to it: Since on slipping, the random pinning forces,  $f_R$  in a region are reduced by order  $\epsilon$ , the effective critical

force  $F_c$  for continuous slip will have been reduced by order  $\epsilon$ ; thus if  $F > F_c(\epsilon) = F_c - C_w \epsilon$ , the mean velocity  $\bar{v}$  will be nonzero. A similar effect can be caused by stress pulses associated with Eq. (6). By considering which of the sites in a long quake with  $\alpha = 0$  can be caused to slip further by such stress pulses one finds that runaway will occur for  $M \geq M_D \sim \alpha^{-4}$  for the physical case [27]. This has been checked in  $d = 1$  with  $\Gamma = 1$  and  $\gamma = 0$ , finding the predicted reduced critical force  $F_c(\alpha) \sim F_c - C_p \alpha^2$  as shown in Fig. 10 [27]. These 1-d simulations also reveal a hysteretic  $\bar{v}(F)$  curve in finite systems. This is expected to also occur with the model with weakening discussed above. Related higher dimensional systems are discussed in [60] and in [30].

We can now understand what should happen with either weakening or stress pulses in finite systems driven with a weak spring or with slowly moving boundaries. As the system is loaded, quakes of increasing size are observed. If the system is small enough that it cannot sustain quakes with  $M > M_D(\epsilon, \alpha)$ , i. e. even events within the power law scaling regime of the event size distribution, with  $M \leq M_D(\epsilon, \alpha)$ , are system spanning, then the behavior will not be much different from the monotonic case with  $\epsilon = \alpha = 0$ . In both cases there is a power law event size distribution all the way to the largest events, that are determined by the system size. This will occur if the dominant linear system size  $L$  is less than the maximum possible linear extent of an earthquake that does not become a runaway event:  $L < R_D(\epsilon, \alpha) \sim M_D^{1/2} \sim \max(C_\alpha/\alpha^2, C_\epsilon/\epsilon)$



Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of, Figure 10

Mean velocity vs. force for one dimensional system with a non-monotonic kernel  $J(x, t) = \delta(t - x)/x^2 + \alpha \delta'(t - x)/x$  for  $\alpha = 0.8, 0.5, 0$ . A spring or boundary loaded system will traverse the hysteresis loops in the direction indicated. Inset: the threshold force,  $F_c^\uparrow(\alpha)$ , on increasing the load;  $\Delta_\alpha = [1 - F_c^\uparrow(\alpha)/F_c^\uparrow(\alpha = 0)]^{1/2}$  is plotted vs.  $\alpha$  (from [27])

with appropriate coefficients  $C_\alpha$ ,  $C_\epsilon$ , which will depend on the amount of randomness in the fault. On the other hand, if  $L > R_D$ , quakes of size of order  $R_D$  will runaway and most of the system will slip, stopping only when the load has decreased enough to make the loading forces less than the lower end of the hysteresis loop in  $\bar{v}(F)$  (as in Fig. 10).

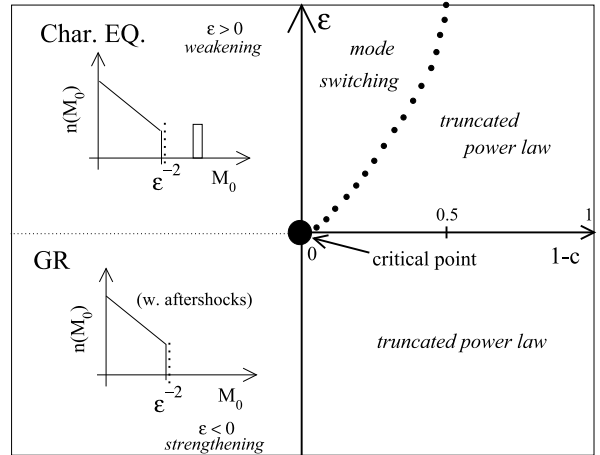
Because of the tendency of regions that have already slipped to slip further, and the consequent buildup of larger stresses near the boundaries of the slipped regions, large events in systems with dynamic weakening will be much more crack like than in monotonic models, probably with  $\Delta u \sim L$ . Statistics of quakes with weakening,  $\epsilon$ , reasonably large, but no stress pulses ( $\alpha = 0$ ) are shown in Fig. 4 and in [2,5,6]; note the absence of quakes with intermediate moments. A typical large event in this case is shown in Fig. 5b; it appears to be crack-like.

In this section we have shown that simple models of heterogeneous faults – with the dimensionality and long-range elastic interactions properly included – can give rise to either power-law statistics of earthquake moments or a distribution of small events combined with characteristic system size events. Which behavior – or intermediate behavior – obtains is found to depend on a number of physical properties such as frictional weakening and dynamic stress transfer, analogs of which should definitely exist in real systems. In the power-law-regime the conventionally defined Gutenberg–Richter exponent  $b \equiv 3B/2$  [3] is found to be  $b = 3/4$ . This is close to the observed  $b$ -value of global strike-slip earthquakes at depth less than 50 km [28].

**Mode-Switching** In [17] the mean field approximation of infinite range elastic interaction in the BZR model with  $N = LW$  (with  $W \sim L$ ) geometrically equal cells on the fault, is used to write the local stress  $\tau_i$  on cell  $i$  as

$$\begin{aligned} \tau_i &= J/N \sum_j (u_j - u_i) + K_L(vt - u_i) \\ &= J\bar{u} + K_L vt - (K_L + J)u_i, \end{aligned} \tag{16}$$

where  $u_i$  is the total fault offset of cell  $i$  in the horizontal ( $x$ ) direction,  $\bar{u} = (\sum_j u_j)/N$ ,  $J/N$  is the elastic coupling between cells in the mean-field approximation, and  $K_L$  is the effective loading stiffness of the bulk material surrounding the fault patch. Instead of the loading spring stiffness  $K_L$ , a conservation parameter  $c \equiv J/(K_L + J)$  is introduced, which equals the fraction of the stress drop of the failing cell, that is retained in the system after the slip. There, it is shown that for the physical loading spring stiffness  $K_L \sim 1/L$ , one has  $1 - c \sim O(1/\sqrt{N})$ . A value  $c < 1$  for a large system would be physically realized if the ex-



Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of, Figure 11  
**Phase Diagram of the BZR model, described in the text. The range  $\epsilon > 0$  represents dynamic weakening, while  $\epsilon < 0$  represents strengthening. The parameter  $1 - c$  quantifies the deviation from stress conservation in the mean field approximation of the model**

ternal drive is closer to the fault than its linear extent. To be precise, mean field theory only gives the correct physical scaling behavior near the critical point at zero weakening  $\epsilon \rightarrow 0$  and for  $c \rightarrow 1$ . In [17] it is shown, however, that in a certain parameter regime for  $\epsilon > 0$  and  $0.5 < c < 1$  indicated in the phase diagram of Fig. 11 one finds a mode switching behavior between Gutenberg–Richter statistics and characteristic earthquake statistics. Similar mode switching behavior has also been seen in a more realistic three dimensional model for coupled evolution of earthquakes and faults [4,41], and in numerical simulations with the BZR model that includes elastic stress transfer [76]. In the mean field BZR model, the activity switching results from episodic global reorganization of the mode of strain energy release of the fault system, reflected in a “configurational entropy” of stress states on the fault [17]. This is associated with a statistical competition between a tendency of a synchronized behavior leading to clusters of large earthquakes and the characteristic earthquake distribution, and a tendency for disordered response leading to Gutenberg–Richter type statistics without a preferred event size. Mode switching happens when these two opposite tendencies are roughly equal in strength. Some possible observational evidence for mode switching in earthquake data are discussed in [4].

**Results on Aftershocks** As mentioned in Sect. “Simple Models for Inhomogeneous Earthquake Faults”, we asso-

ciate regions off the main fault segments that are in an early deformation stage with dynamic strengthening  $\epsilon < 0$ . To capture basic aspects of brittle deformation on such regions in the three-dimensional volume around the main fault (Fig. 3), we change the model as follows: when any cell  $i$  slips during an earthquake, and thereby reduces its stress by  $\Delta\tau_i \equiv \tau_{f,i} - \tau_{a,i}$ , the failure stress  $\tau_{f,j}$  of every cell  $j = 1, \dots, N$  is *strengthened* by an amount  $|\epsilon|\Delta\tau_i/N$ . Once the earthquake is complete, the failure stress of each cell is slowly lowered back to its original value. This represents in a simple way the brittle deformation that occurs during an earthquake in the off-fault regions, which are first in a strengthening regime, compared to the main fault, and then have a weakening process. The events that are triggered as the failure stresses are lowered in the weakening period are referred to as *aftershocks*. The occurrence of aftershocks in this version of the model for off-fault regions is in agreement with the observation that a large fraction of observed aftershocks typically occur in off-fault regions [70]. For this version of the model with  $\epsilon < 0$ , both the primary earthquakes (i. e., mainshocks) and the triggered aftershocks are distributed according to the Gutenberg–Richter distribution, up to a cutoff moment scaling as  $1/\epsilon^2$ . Assuming that the increased failure stress thresholds  $\tau_{f,i}$  are slowly lowered with time as  $\log(t)$  towards their earlier static values  $\tau_{s,i}$ , and that the stresses are distributed over a wide range of values, we show analytically in [46] that the temporal decay of aftershock rates at long times is proportional to  $1/t$ , as in the modified Omori law  $\Delta N/\Delta t K/(t+c)^p$  with  $p = 1$  [3,69,70], where  $N$  is the cumulative number of aftershocks,  $t$  is the time after the mainshock, and  $K$ ,  $c$ , and  $p$  are empirical constants.

Remarkably, the long length scale behavior of this model can be shown [45] to be the same as the behavior of the mean field BZR model given in Eq. (16) with an added “antiferroelastic” term  $(-\epsilon|J\bar{u})$ :

$$\tau_i = J\bar{u} + K_L v t - (K_L + J)u_i - \epsilon|J\bar{u}. \quad (17)$$

In Eq. (17) every time a cell fails, it slips by an amount  $\Delta u_i$  that leads to stress loading of the other cells, lessened by  $|\epsilon|J\Delta u_i/N$  compared to our original model (Eq. (16)). On the other hand, in the global strengthening model (described above) when a cell slips the failure stresses of all cells are strengthened by  $|\epsilon|J\Delta u_i/N$ . On long length scales the global strengthening of the failure stress has equivalent effects on the earthquake statistics as the dissipation of the redistributed stress, up to corrections of order  $O(1/N)$ , so the scaling behavior for large events of both models are the same. Moreover, Eq. (17) can be rewritten as:

$$\tau_i = J[1 - |\epsilon|][\bar{u} - u_i] + K_L v t - [K_L + J|\epsilon|]u_i. \quad (18)$$

We can now absorb  $|\epsilon|$  by defining  $J' = J(1 - |\epsilon|)$  and  $K'_L = K_L + J|\epsilon|$ . Rewriting Eq. (18) with the new definitions, and dropping the  $|\epsilon|$  contribution in  $[K'_L - J|\epsilon|]v t$  since  $v \rightarrow 0$ , we find:

$$\tau_i = J'\bar{u} + K'_L v t - (K'_L + J')u_i. \quad (19)$$

Therefore we recover Eq. (16) with  $J \rightarrow J'$  and  $K_L \rightarrow K'_L$ . This amounts to changing the stress conservation parameter  $c$  (from reference [17]). For Eq. (19):

$$c = J'/(K'_L + J') = 1 - |\epsilon| \quad (20)$$

where  $K_L \rightarrow 0$  since we are concerned with the adiabatic limit. We also know (from reference [17]) that the cutoff  $S_{cf}$  for the Gutenberg–Richter distribution scales as  $S_{cf} \sim 1/(1-c)^2$ . Thus, from Eq. (20) we find that the cutoff for Eq. (17) will scale as  $\sim 1/|\epsilon|^2$ .

**Mapping to Single Interface Magnet Model** The mean field version of the single interface magnet model with infinite range antiferromagnetic interactions is given by [22,74]:

$$\dot{h}_i(t) = J[\bar{h} - h_i(t)] + H(t) - k\bar{h} + \eta_i(h) \quad (21)$$

where  $h_i(t)$  is the position of the domain wall,  $H(t)$  is the external driving field,  $k$  is the coefficient of the antiferromagnetic term, and  $\eta_i(h)$  is the pinning field. In the paper by Fisher et al. [27] it has been shown that the scaling behavior on long length scales resulting from Eq. (5), without the  $-|\epsilon|J\bar{u}$  term, is same as that of Eq. (21) without the antiferromagnetic term  $-k\bar{h}$ . Furthermore, upon inspection we see the following correspondence between the single interface magnet model (Eq. (21)), and the mean field earthquake model (Eq. (17)):

$$-k\bar{h} \iff -|\epsilon|J\bar{u} \quad (22)$$

In other words, the coefficient of the antiferromagnetic term  $k$  plays the same role in the magnet model (Eq. (21)), as the coefficient of strengthening  $|\epsilon|J$  does in the earthquake model (Eq. (17)).

## Summary

### Phase Diagram

The regimes with various statistics produced by the model are summarized by the phase diagram given in Fig. 11. The range  $\epsilon > 0$  corresponds to “mature” localized faults with a weakening rheology and characteristic earthquake statis-

tics. The value  $\epsilon = 0$  corresponds to “immature” strongly inhomogeneous fault zones and fault networks with power law statistics and scale invariant rupture properties. The range  $\epsilon < 0$  corresponds to the fracture and fault networks around large rupture zones, characterized by strengthening due to the creation of new structures and associated emerging aftershocks. The right side of the diagram summarizes the mean field theory results on mode switching described in Sect. “Mode-Switching”. The left side of the phase diagram resembles the phase diagram for avalanches in the nucleation RFIM for magnets [65]. There, too, increasing the disorder from small to large (compared to the ferromagnetic coupling between the individual domains) drives the system from a characteristic avalanche size distribution to a truncated power law, with a disorder induced critical point separating the two regimes.

It may be surprising that the discussed simple BZR model can capture many of the essential general features of earthquake statistics (or other systems with avalanches, such as driven magnetic domain walls). This can be understood through the renormalization group [10,66], a powerful mathematical tool to coarse grain a system and extract its effective behavior on long space-time scales. Many microscopic details of a system are averaged out under coarse graining, and universal aspects of the behavior on long scales depend only on a few basic properties such as symmetries, dimensions, range of interactions, weakening/strengthening, etc. When a model correctly captures those basic features, the results provide proper predictions for statistics, critical exponents, and universal scaling functions near the critical point. Consequently, many models that are in the same universality class lead to the same statistics and exponents [10,17,27,66].

## Conclusions

The phenomenology of earthquakes and avalanches in magnets exhibit a number of power law distributions and scale-invariant functions (Table 1). In search of basic model ingredients that can explain these results, we have focused on models that are rich enough to produce a diversity of observed features, while being simple enough to allow analytical predictions on long spatio-temporal scales. For the earthquake system we use the BZR model for a heterogeneous fault with threshold dynamics and long range stress-transfer interactions [2,5,6]. For the magnet system we use variants of the RFIM model with threshold dynamics and both short range and long range interactions [37,65,66,74]. In both classes of models, changes in the property disorder and dynamic effects lead to different dynamic regimes (Fig. 11). For different ranges of param-

eters, the earthquake model produces fractal and crack-like slip functions, power law frequency-size statistics, characteristic earthquake distribution, mode switching, and aftershocks. Similar features are found with the magnet models. We discussed two universal scaling functions of moment rates near criticality as a stronger test of the theory against observations than mere scaling exponents that have large error bars. As in magnetic systems, we find that our analysis for earthquakes provides a good overall agreement between theory and observations, but with a potential discrepancy in one particular universal scaling function for mean moment-rate shapes at fixed duration. The discrepancy has an interesting precedent in the context of avalanches in magnetic systems, and has been explained there in terms of non-universal time retardation effects due to eddy currents. Similar retardation effects may be due to triggering delays or strengthening effects that are responsible for aftershocks in earthquake faults. More observational data, in particular on small earthquakes would be needed to test some of the predictions in detail.

## Future Directions

We have highlighted some interesting connections between earthquake and magnet systems with a jerky response to a slowly varying driving force. Future useful studies include analysis of factors controlling nucleation processes, transitions to instabilities and final event sizes, along with a more detailed analysis of the effects of geometrical heterogeneities in the fault structure on the statistics of earthquakes. Additional observational data, particularly for small earthquakes, are needed to test predictions for the scaling of the earthquake duration and rupture area with moment, and for accurately testing our mean field predictions for moment rate shapes. Developing analytical corrections to the mean field earthquake models can provide additional important insights. Testing similar ideas in other systems with crackling noise would improve and deepen our understanding of universal behavior in disordered nonequilibrium systems.

## Acknowledgments

We thank Daniel S. Fisher, James R. Rice, James P. Sethna, Michael B. Weissman, Deniz Ertas, Matthias Holschneider, Amit Mehta, Gert Zöller and many others for very helpful discussions. K.D. acknowledges support from the National Science Foundation, the NSF funded Materials Computation Center, and IBM. YBZ acknowledges support from the National Science Foundation, the United States Geological Survey, and the Southern California Earthquake Center.

## Bibliography

### Primary Literature

1. Aki K, Richards PG (2002) *Quantitative Seismology*, 2nd edn. University Science Books, Sausalito
2. Ben-Zion Y (1996) Stress slip and earthquakes in models of complex single-fault systems incorporating brittle and creep deformations. *J Geophys Res* 101:5677–5706
3. Ben-Zion Y (2003) Appendix 2, Key Formulas in Earthquake Seismology. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part B*. Academic Press, San Diego, pp 1857–1875
4. Ben-Zion Y, Dahmen K, Lyakhovsky V, Ertaş D, Agnon A (1999) Self Driven Mode Switching of Earthquake Activity on a Fault System. *Earth Planet Sci Lett* 172(1–2):11–21
5. Ben-Zion Y, Rice JR (1993) Earthquake failure sequences along a cellular fault zone in a three-dimensional elastic solid containing asperity and nonasperity regions. *J Geophys Res* 98:14109–14131
6. Ben-Zion Y, Rice JR (1995) Slip patterns and earthquake populations along different classes of faults in elastic solids. *J Geophys Res* 100:12959–12983
7. Ben-Zion Y, Sammis CG (2003) Characterization of Fault Zones. *Pure Appl Geophys* 160:677–715
8. Ben-Zion Y, Zhu L (2002) Potency-magnitude Scaling Relations for Southern California Earthquakes with  $1.0 < ML < 7.0$ . *Geophys J Int* 148:F1–F5
9. Bilek SL (2001) Earthquake rupture processes in circum-Pacific subduction zones. Ph D thesis, University of California
10. Binney JJ, Dowrick NJ, Fisher AJ, Newman MEJ (1993) *The theory of critical phenomena*. Oxford University Press
11. Carlson JM, Langer JS, Shaw BE (1994) Dynamics of earthquake faults. *Rev Mod Phys* 66:658–70, and references therein
12. Chen K, Bak P, Obukhov SP (1991) *Phys Rev A* 43:625
13. Cizeau P, Zapperi S, Durin G, Stanley HE (1997) *Phys Rev Lett* 79:4669–4672
14. Cowie PA, Vanette C, Sornette D (1993) *J Geophys Res* 98:21809
15. Dahmen K (1995) Hysteresis, Avalanches, and Disorder Induced Critical Scaling: A Renormalization Group Approach, Ph D Thesis, Cornell University
16. Dahmen K (2005) *Nature Physics* 1:13–14
17. Dahmen K, Ertaş D, Ben-Zion Y (1998) Gutenberg–Richter and Characteristic Earthquake behavior in a Simple Mean-Field Model of Heterogeneous Faults. *Phys Rev E* 58:1494–1501
18. Dahmen KA, Sethna JP (1996) Hysteresis, Avalanches, and Disorder Induced Critical Scaling: A Renormalization Group Approach. *Phys Rev B* 53:14872
19. Dieterich JH (1979) *J Geophys Res* 84:2161–2168
20. Dieterich JH (1981) *Amer Geophys Union Monog* 24:103–120
21. Durin G, Zapperi S (2000) Scaling exponents for barkhausen avalanches in polycrystalline and amorphous ferromagnets. *Phys Rev Lett* 84:4705–4708
22. Durin G, Zapperi S (2001) *J Magn Mat* 1085:242–245
23. Durin G, Zapperi S (2002) Low field hysteresis in disordered ferromagnets. *Phys Rev B* 65:144441
24. Ertaş D, Kardar M (1994) Critical dynamics of contact line depinning. *Phys Rev E* 49:R2532–5
25. Ertaş D, Kardar M (1994) *Phys Rev E* 49:R2532. (1994) *Phys Rev Lett* 73:1703
26. Fisher DS (1998) *Phys Rep* 301:113
27. Fisher DS, Dahmen K, Ramanathan S, Ben-Zion Y (1997) *Phys Rev Lett* 78:4885–4888
28. Frohlich C, Davis SD (1993) *J Geophys Res* 98:631
29. Gutenberg B, Richter CF (1954) *Seismicity of Earth and Associated Phenomena*. Princeton Univ. Press, Princeton
30. Hillers G, Mai PM, Ben-Zion Y, Ampuero J-P (2007) Statistical Properties of Seismicity Along Fault Zones at Different Evolutionary Stages. *Geophys J Int* 169:515–533. doi: 10.1111/j.1365-246X.2006.03275.x
31. Houston H (2001) Influence of depth, focal mechanism, and tectonic setting on the shape and duration of earthquake source time functions. *J Geophys Res* 106(B6):11137–11150
32. Ji H, Robbins MO (1992) Percolative, self-affine, and faceted domain growth in random three-dimensional magnets. *Phys Rev B* 46:14519–27
33. Jiles D (1991) *Introduction to Magnetism and Magnetic Materials*. Chapman and Hall
34. Klein W, Rundle JB, Ferguson CD (1997) Scaling and Nucleation in Models of Earthquake Faults. *Phys Rev Lett* 78:3793–3796
35. Koiller B, Ji H, Robbins MO (1992) Fluid wetting properties and the invasion of square networks. *ibid* 45:7762–7
36. Koiller B, Ji H, Robbins MO (1992) Effect of disorder and lattice type on domain-wall motion in two dimensions. *Phys Rev B* 46:5258–65
37. Kuntz MC, Sethna JP (2000) *Phys Rev B* 62:11699–11708
38. Langer JS, Carlson JM, Myers CR, Shaw BE (1996) Slip complexity in dynamic models of earthquake faults. *Proc Natl Acad Sci* 93:3825–3829
39. Laurson L, Alava MJ (2006)  $1/f$  noise and avalanche scaling in plastic deformation. *Phys Rev E* 74:066106
40. Lomnitz-Adler J (1993) Automaton models of seismic fracture: constraints imposed by the magnitude-frequency relation. *J Geophys Res* 98:17745–17756
41. Lyakhovsky V, Ben-Zion Y, and Agnon A (2001) Earthquake Cycle, Fault Zones, and Seismicity Patterns in a Rheologically Layered Lithosphere. *J Geophys Res* 106:4103–4120
42. Marchetti MC, Middleton AA, Prellberg T (2000) Viscoelastic Depinning of Driven Systems: Mean-Field Plastic Scallop. *Phys Rev Lett* 85:1104–1107
43. Martys N, Robbins MO, Cieplak M (1991) Scaling relations for interface motion through disordered media: application to two-dimensional fluid invasion. *Phys Rev B* 44:12294–306
44. Mayergoyz ID (1991) *Mathematical Models of Hysteresis*. Springer
45. Mehta AP (2005) Ph D Thesis, University of Illinois at Urbana Champaign
46. Mehta AP, Dahmen KA, Ben-Zion Y (2006) Universal mean moment rate Profiles of earthquake ruptures. *Phys Rev E* 73:056104
47. Mehta AP, Mills AC, Dahmen KA, Sethna JP (2002) *Phys Rev E* 65:46139/1–6
48. Middleton AA (1992) *Phys Rev Lett* 68:670
49. Miltenberger P, Sornette D, Vanette C (1993) *Phys Rev Lett* 71:3604
50. Myers CR, Sethna JP (1993) Collective dynamics in a model of sliding charge-density waves. I. Critical behavior. *Phys Rev B* 47:11171–93

51. Myers CR, Sethna JP (1993) Collective dynamics in a model of sliding charge-density waves. II. Finite-size effects. *Phys Rev B* 47:11194–203
52. Narayan O, Fisher DS (1992) Critical behavior of sliding charge-density waves in  $4 - \epsilon$  dimensions, *Phys Rev B* 46:11520–49
53. Narayan O, Fisher DS (1992) Dynamics of sliding charge-density waves in  $4 - \epsilon$  dimensions. *Phys Rev Lett* 68:3615–8
54. Narayan O, Fisher DS (1993) Threshold critical dynamics of driven interfaces in random media. *Phys Rev B* 48:7030–42
55. Narayan O, Middleton AA (1994) Avalanches and the renormalization group for pinned charge-density waves. *Phys Rev B* 49:244
56. Nattermann T (1997) Theory of the Random Field Ising Model. In: Young AP (ed) *Spin Glasses and Random Fields*. World Scientific, Singapore
57. Omori F (1894) On the aftershocks of earthquakes. *J Coll Sci Imp Univ Tokyo* 7:111–200
58. Perković O, Dahmen K, Sethna JP (1995) Avalanches, Barkhausen Noise, and Plain Old Criticality. *Phys Rev Lett* 75:4528–31
59. Perković O, Dahmen K, Sethna JP (1999) Disorder-Induced Critical Phenomena in Hysteresis: Numerical Scaling in Three and Higher Dimensions. *Phys Rev B* 59:6106–19
60. Ramanathan S, Fisher DS (1998) *Phys Rev B* 58:6026
61. Rice JR, Ben-Zion Y (1996) Slip complexity in earthquake fault models. *Proc Natl Acad Sci* 93:3811–3818
62. Ruff LJ, Miller AD (1994) *Pure Appl Geophys* 142:101
63. Schwarz JM, Fisher DS (2001) Depinning with Dynamic Stress Overshoots: Mean Field Theory. *Phys Rev Lett* 87:096107/1–4
64. Sethna JP (2006) Les Houches Summer School notes. Crackling Noise and Avalanches: Scaling, Critical Phenomena, and the Renormalization Group. e-print at <http://xxx.lanl.gov/pdf/cond-mat/0612418>
65. Sethna JP, Dahmen K, Kartha S, Krumhansl JA, Roberts BW, Shore JD (1993) Hysteresis and Hierarchies: Dynamics of Disorder Driven First Order Phase Transformations. *Phys Rev Lett* 70:3347
66. Sethna JP, Dahmen KA, Myers CR (2001) *Nature* 410:242–250
67. Spasojevic D, Bukvic S, Milosevic S, Stanley HE (1996) Barkhausen noise: Elementary signals, power laws, and scaling relations. *Phys Rev E* 54:2531–2546
68. Travesset A, White RA, Dahmen KA (2002) *Phys Rev B* 66:024430
69. Utsu T (2002) Statistical features of seismology. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part A*. pp 719–732
70. Utsu Y, Ogata Y, Matsu'ura RS (1995) The centenary of the Omori Formula for a decay law of aftershock activity. *J Phys Earth* 43:1–33
71. Vere-Jones D (1976) A branching model for crack propagation. *Pure Appl Geophys* 114(4):711–726
72. Zapperi S, Castellano C, Calaiori F, Durin G (2005) Signature of effective mass in crackling-noise asymmetry. *Nature Phys* 1:46–49
73. Zapperi S, Cizeau P, Durin G, Stanley HE (1998) Dynamics of a ferromagnetic domain wall: Avalanches, depinning transition, and the Barkhausen effect. *Phys Rev B* 58:6353–66
74. Zapperi S, Cizeau P, Durin G, Stanley HE (1998) *Phys Rev B* 58(10):6353–6366
75. Zöller G, Hainzl S, Ben-Zion Y, Holschneider M (2009) Critical states of seismicity: From models to practical seismic hazard estimates. In: *Encyclopedia of Complexity and System Science*
76. Zöller G, Holschneider M, Ben-Zion Y (2004) Quasi-static and Quasi-dynamic modeling of earthquake failure at intermediate scales. *Pure Appl Geophys* 161:2103–2118
77. Zöller G, Holschneider M, Ben-Zion Y (2005) The role of heterogeneities as a tuning parameter of earthquake dynamics. *Pure Appl Geophys* 162:1027 V1049. doi: 10.1007/s00024-004-2660-9

# Pedestrian, Crowd and Evacuation Dynamics

DIRK HELBING<sup>1,2</sup>, ANDERS JOHANSSON<sup>1</sup>

<sup>1</sup> ETH Zurich, Zurich, Switzerland

<sup>2</sup> Institute for Advanced Study, Collegium Budapest, Budapest, Hungary

## Article Outline

Glossary

Definition of the Subject

Introduction

Pedestrian Dynamics

Crowd Dynamics

Evacuation Dynamics

Future Directions

Acknowledgments

Bibliography

## Glossary

**Collective intelligence** Emergent functional behavior of a large number of people that results from *interactions* of individuals rather than from individual reasoning or global optimization.

**Crowd** Agglomeration of many people in the same area at the same time. The density of the crowd is assumed to be high enough to cause continuous interactions with or reactions to other individuals.

**Crowd turbulence** Unanticipated and unintended irregular motion of individuals into different directions due to strong and rapidly changing forces in crowds of extreme density.

**Emergence** Spontaneous establishment of a qualitatively new behavior through non-linear interactions of many objects or subjects.

**Evolutionary optimization** Gradual optimization based on the effect of frequently repeated random mutations and selection processes based on some success function (“fitness”).

**Faster-is-slower effect** This term reflects the observation that certain processes (in evacuation situations, production, traffic dynamics, or logistics) take more time if performed at high speed. In other words, waiting can often help to coordinate the activities of several competing units and to speed up the average progress.

**Freezing-by-heating effect** Noise-induced blockage effect caused by the breakdown of direction-segregated walking patterns (typically two or more “lanes” characterized by a uniform direction of motion). “Noise”

means frequent variations of the walking direction due to nervousness or impatience in the crowd, e. g. also frequent overtaking maneuvers in dense, slowly moving crowds.

**Panic** Breakdown of ordered, cooperative behavior of individuals due to anxious reactions to a certain event. Often, panic is characterized by attempted escape of many individuals from a real or perceived threat in situations of a perceived struggle for survival, which may end up in trampling or crushing of people in a crowd.

**Self-organization** Spontaneous organization (i. e. formation of ordered patterns) not induced by initial or boundary conditions, by regulations or constraints. Self-organization is a result of non-linear interactions between many objects or subjects, and it often causes different kinds of spatio-temporal patterns of motion.

**Social force** Vector describing acceleration or deceleration effects that are caused by social interactions rather than by physical interactions or fields.

## Definition of the Subject

The modeling of pedestrian motion is of great theoretical and practical interest. Recent experimental efforts have revealed quantitative details of pedestrian interactions, which have been successfully cast into mathematical equations. Furthermore, corresponding computer simulations of large numbers of pedestrians have been compared with the empirically observed dynamics of crowds. Such studies have led to a deeper understanding of how collective behavior on a macroscopic scale emerges from individual human interactions. Interestingly enough, the non-linear interactions of pedestrians lead to various complex, spatio-temporal pattern-formation phenomena. This includes the emergence of lanes of uniform walking direction, oscillations of the pedestrian flow at bottlenecks, and the formation of stripes in two intersecting flows. Such self-organized patterns of motion demonstrate that efficient, “intelligent” collective dynamics can be based on simple, local interactions. Under extreme conditions, however, coordination may break down, giving rise to critical crowd conditions. Examples are “freezing-by-heating” and “faster-is-slower” effects, but also the transition to “turbulent” crowd dynamics. These observations have important implications for the optimization of pedestrian facilities, in particular for evacuation situations.

## Introduction

The emergence of new, functional or complex collective behaviors in social systems has fascinated many scientists. One of the primary questions in this field is how cooper-



ation or coordination patterns originate based on elementary individual interactions. While one could think that these are a result of intelligent human actions, it turns out that much simpler models assuming automatic responses can reproduce the observations very well. This suggests that humans are using their intelligence primarily for more complicated tasks, but also that simple interactions can lead to intelligent patterns of motion. Of course, it is reasonable to assume that these interactions are the result of a previous learning process that has optimized the automatic response in terms of minimizing collisions and delays. This, however, seems to be sufficient to explain most observations.

In this contribution, we will start with a short history of pedestrian modeling and, then, introduce a simplified model of pedestrian interactions, the “social force model”. Furthermore, we will discuss its calibration using video tracking data. Next, we will turn to the subject of crowd dynamics, as one typically finds the formation of large-scale spatio-temporal patterns of motion, when many pedestrians interact with each other. These patterns will be discussed in some detail before we will turn to evacuation situations and cases of extreme densities, where one can sometimes observe the breakdown of coordination. Finally, we will address possibilities to design improved pedestrian facilities, using special evolutionary algorithms.

## Pedestrian Dynamics

### Short History of Pedestrian Modeling

Pedestrians have been empirically studied for more than four decades [1,2,3]. The evaluation methods initially applied were based on direct observation, photographs, and time-lapse films. For a long time, the main goal of these studies was to develop a *level-of-service concept* [4], *design elements* of pedestrian facilities [5,6,7,8], or *planning guidelines* [9,10]. The latter have usually the form of *regression relations*, which are, however, not very well suited for the prediction of pedestrian flows in pedestrian zones and buildings with an exceptional architecture, or in challenging evacuation situations. Therefore, a number of simulation models have been proposed, e.g. *queuing models* [11], *transition matrix models* [12], and *stochastic models* [13], which are partly related to each other. In addition, there are models for the *route choice behavior* of pedestrians [14,15].

None of these concepts adequately takes into account the self-organization effects occurring in pedestrian crowds. These are the subject of recent experimental studies [8,16,17,18,19,20]. Most pedestrian models, however, were formulated before. A first modeling approach that

appears to be suited to reproduce spatio-temporal patterns of motion was proposed by Henderson [21], who conjectured that pedestrian crowds behave similar to gases or fluids (see also [22]). This could be partially confirmed, but a realistic gas-kinetic or fluid-dynamic theory for pedestrians must contain corrections due to their particular interactions (i.e. avoidance and deceleration maneuvers) which, of course, do not obey momentum and energy conservation. Although such a theory can be actually formulated [23,24], for practical applications a direct simulation of *individual* pedestrian motion is favorable, since this is more flexible. As a consequence, pedestrian research mainly focuses on *agent-based models* of pedestrian crowds, which also allow one to consider local coordination problems. The “social force model” [25,26] is maybe the most well-known of these models, but we also like to mention *cellular automata* of pedestrian dynamics [27,28,29,30,31,32,33] and *AI-based models* [34,35].

### The Social Force Concept

In the following, we shall shortly introduce the social force concept, which reproduces most empirical observations in a simple and natural way. Human behavior often seems to be “chaotic”, irregular, and unpredictable. So, why and under what conditions can we model it by means of forces? First of all, we need to be confronted with a phenomenon of motion in some (quasi-)continuous space, which may be also an abstract behavioral space such as an opinion scale [36]. Moreover, it is favorable to have a system where the fluctuations due to unknown influences are not large compared to the systematic, deterministic part of motion. This is usually the case in pedestrian traffic, where people are confronted with standard situations and react “automatically” rather than taking complicated decisions, e.g. if they have to evade others.

This “automatic” behavior can be interpreted as the result of a *learning process* based on trial and error [37], which can be simulated with *evolutionary algorithms* [38]. For example, pedestrians have a preferred side of walking, since an asymmetrical avoidance behavior turns out to be profitable [25,37]. The related *formation of a behavioral convention* can be described by means of *evolutionary game theory* [25,39].

Another requirement is the vectorial additivity of the separate force terms reflecting different environmental influences. This is probably an approximation, but there is some experimental evidence for it. Based on quantitative measurements for animals and test persons subject to separately or simultaneously applied stimuli of different nature and strength, one could show that the behavior in

conflict situations can be described by a superposition of forces [40,41]. This fits well into a concept by Lewin [42], according to which behavioral changes are guided by so-called *social fields* or *social forces*, which has later on been put into mathematical terms [25,43]. In some cases, social forces, which determine the amount and direction of systematic behavioral changes, can be expressed as gradients of dynamically varying potentials, which reflect the social or behavioral fields resulting from the interactions of individuals. Such a social force concept was applied to opinion formation and migration [43], and it was particularly successful in the description of collective pedestrian behavior [8,25,26,37].

For reliable simulations of pedestrian crowds, we do not need to know whether a certain pedestrian, say, turns to the right at the next intersection. It is sufficient to have a good estimate what percentage of pedestrians turns to the right. This can be either empirically measured or estimated by means of route choice models [14]. In some sense, the uncertainty about the individual behaviors is averaged out at the macroscopic level of description. Nevertheless, we will use the more flexible microscopic simulation approach based on the social force concept. According to this, the temporal change of the location  $\mathbf{r}_\alpha(t)$  of pedestrian  $\alpha$  obeys the equation of motion

$$\frac{d\mathbf{r}_\alpha(t)}{dt} = \mathbf{v}_\alpha(t). \quad (1)$$

Moreover, if  $\mathbf{f}_\alpha(t)$  denotes the sum of social forces influencing pedestrian  $\alpha$  and if  $\boldsymbol{\xi}_\alpha(t)$  are individual fluctuations reflecting unsystematic behavioral variations, the velocity changes are given by the *acceleration equation*

$$\frac{d\mathbf{v}_\alpha}{dt} = \mathbf{f}_\alpha(t) + \boldsymbol{\xi}_\alpha(t). \quad (2)$$

A particular advantage of this approach is that we can take into account the flexible usage of space by pedestrians, requiring a continuous treatment of motion. It turns out that this point is essential to reproduce the empirical observations in a natural and robust way, i. e. without having to adjust the model to each single situation and measurement site. Furthermore, it is interesting to note that, if the fluctuation term is neglected, the social force model can be interpreted as a particular *differential game*, i. e. its dynamics can be derived from the minimization of a special utility function [44].

### Specification of the Social Force Model

The social force model for pedestrians assumes that each individual  $\alpha$  is trying to move in a desired direction  $\mathbf{e}_\alpha^0$

with a desired speed  $v_\alpha^0$ , and that it adapts the actual velocity  $\mathbf{v}_\alpha$  to the desired one,  $\mathbf{v}_\alpha^0 = v_\alpha^0 \mathbf{e}_\alpha^0$ , within a certain relaxation time  $\tau_\alpha$ . The systematic part  $\mathbf{f}_\alpha(t)$  of the acceleration force of pedestrian  $\alpha$  is then given by

$$\mathbf{f}_\alpha(t) = \frac{1}{\tau_\alpha} (v_\alpha^0 \mathbf{e}_\alpha^0 - \mathbf{v}_\alpha) + \sum_{\beta(\neq\alpha)} \mathbf{f}_{\alpha\beta}(t) + \sum_i \mathbf{f}_{\alpha i}(t), \quad (3)$$

where the terms  $\mathbf{f}_{\alpha\beta}(t)$  and  $\mathbf{f}_{\alpha i}(t)$  denote the repulsive forces describing attempts to keep a certain safety distance to other pedestrians  $\beta$  and obstacles  $i$ . In very crowded situations, additional physical contact forces come into play (see Subsect. “**Force Model for Panicking Pedestrians**”). Further forces may be added to reflect attraction effects between members of a group or other influences. For details see [37].

First, we will assume a simplified interaction force of the form

$$\mathbf{f}_{\alpha\beta}(t) = \mathbf{f}(\mathbf{d}_{\alpha\beta}(t)), \quad (4)$$

where  $\mathbf{d}_{\alpha\beta} = \mathbf{r}_\alpha - \mathbf{r}_\beta$  is the distance vector pointing from pedestrian  $\beta$  to  $\alpha$ . Angular-dependent shielding effects may be furthermore taken into account by a prefactor describing the anisotropic reaction to situations in front of as compared to behind a pedestrian [26,45], see Subsect. “**Angular Dependence**”. However, we will start with a **circular specification** of the distance-dependent interaction force,

$$\mathbf{f}(\mathbf{d}_{\alpha\beta}) = A_\alpha e^{-d_{\alpha\beta}/B_\alpha} \frac{\mathbf{d}_{\alpha\beta}}{\|\mathbf{d}_{\alpha\beta}\|}, \quad (5)$$

where  $d_{\alpha\beta} = \|\mathbf{d}_{\alpha\beta}\|$  is the distance. The parameter  $A_\alpha$  reflects the *interaction strength*, and  $B_\alpha$  corresponds to the *interaction range*. While the dependence on  $\alpha$  explicitly allows for a dependence of these parameters on the single individual, we will assume a homogeneous population, i. e.  $A_\alpha = A$  and  $B_\alpha = B$  in the following. Otherwise, it would be hard to collect enough data for parameter calibration.

**Elliptical Specification** Note that it is possible to express Eq. (5) as gradient of an exponentially decaying potential  $V_{\alpha\beta}$ . This circumstance can be used to formulate a generalized, elliptical interaction force via the potential

$$V_{\alpha\beta}(b_{\alpha\beta}) = AB e^{-b_{\alpha\beta}/B}, \quad (6)$$

where the variable  $b_{\alpha\beta}$  denotes the semi-minor axis  $b_{\alpha\beta}$  of the elliptical equipotential lines. This has been specified according to

$$2b_{\alpha\beta} = \sqrt{\frac{(\|\mathbf{d}_{\alpha\beta}\| + \|\mathbf{d}_{\alpha\beta} - (\mathbf{v}_\beta - \mathbf{v}_\alpha)\Delta t\|)^2}{-\|(\mathbf{v}_\beta - \mathbf{v}_\alpha)\Delta t\|^2}}, \quad (7)$$

so that both pedestrians  $\alpha$  and  $\beta$  are treated symmetrically. The repulsive force is related to the above potential via

$$\begin{aligned} f_{\alpha\beta}(\mathbf{d}_{\alpha\beta}) &= -\nabla_{\mathbf{d}_{\alpha\beta}} V_{\alpha\beta}(b_{\alpha\beta}) \\ &= -\frac{dV_{\alpha\beta}(b_{\alpha\beta})}{db_{\alpha\beta}} \nabla_{\mathbf{d}_{\alpha\beta}} b_{\alpha\beta}(\mathbf{d}_{\alpha\beta}), \end{aligned} \quad (8)$$

where  $\nabla_{\mathbf{d}_{\alpha\beta}}$  represents the gradient with respect to  $\mathbf{d}_{\alpha\beta}$ . Considering the chain rule,  $\|z\| = \sqrt{z^2}$ , and  $\nabla_z \|z\| = z/\sqrt{z^2} = z/\|z\|$ , this leads to the explicit formula

$$\begin{aligned} f_{\alpha\beta}(\mathbf{d}_{\alpha\beta}) &= Ae^{-b_{\alpha\beta}/B} \cdot \frac{\|\mathbf{d}_{\alpha\beta}\| + \|\mathbf{d}_{\alpha\beta} - \boldsymbol{\gamma}_{\alpha\beta}\|}{2b_{\alpha\beta}} \\ &\quad \cdot \frac{1}{2} \left( \frac{\mathbf{d}_{\alpha\beta}}{\|\mathbf{d}_{\alpha\beta}\|} + \frac{\mathbf{d}_{\alpha\beta} - \boldsymbol{\gamma}_{\alpha\beta}}{\|\mathbf{d}_{\alpha\beta} - \boldsymbol{\gamma}_{\alpha\beta}\|} \right) \end{aligned} \quad (9)$$

with  $\boldsymbol{\gamma}_{\alpha\beta} = (\mathbf{v}_\beta - \mathbf{v}_\alpha)\Delta t$ . We used  $\Delta t = 0.5$  s. For  $\Delta t = 0$ , we regain the expression of Eq. (5).

The elliptical specification has two major advantages compared to the circular one: First, the interactions depend not only on the distance, but also on the relative velocity. Second, the repulsive force is not strictly directed from pedestrian  $\beta$  to pedestrian  $\alpha$ , but has a lateral component. As a consequence, this leads to less confrontative, smoother (“sliding”) evading maneuvers. Note that further velocity-dependent specifications of pedestrian interaction forces have been proposed [7,26], but we will restrict to the above specifications, as these are sufficient to demonstrate the method of evolutionary model calibration.

### Evolutionary Calibration with Video Tracking Data

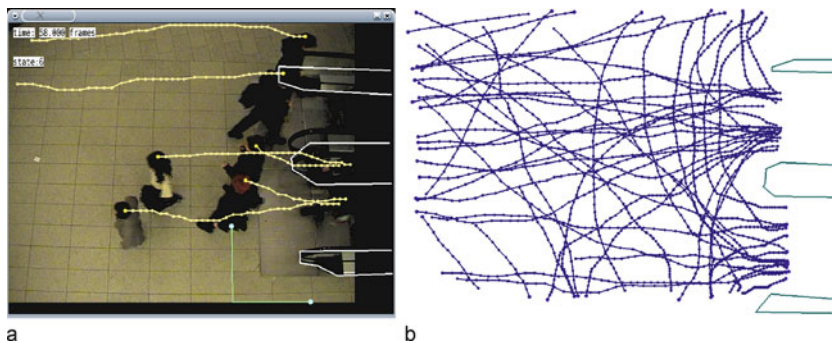
For parameter calibration, several video recordings of pedestrian crowds in different natural environments have been used. The dimensions of the recorded areas were

known, and the floor tiling or environment provided something like a “coordinate system”. The heads were automatically determined by searching for round moving structures, and the accuracy of tracking was improved by comparing actual with linearly extrapolated positions (so it would not happen so easily that the algorithm interchanged or “lost” close by pedestrians). The trajectories of the heads were then projected on two-dimensional space in a way correcting for distortion by the camera perspective. A representative plot of the resulting trajectories is shown in Fig. 1. Note that trajectory data have been obtained with infra-red sensors [47] or video cameras [48,49] for several years now, but algorithms that can simultaneously handle more than one thousand pedestrians have become available only recently [87].

For model calibration, it is recommended to use a hybrid method fusing empirical trajectory data and microscopic simulation data of pedestrian movement in space. In corresponding algorithms, a virtual pedestrian is assigned to each tracked pedestrian in the simulation domain. One then starts a simulation for a time period  $T$  (e.g. 1.5 s), in which one pedestrian  $\alpha$  is moved according to a simulation of the social force model, while the others are moved exactly according to the trajectories extracted from the videos. This procedure is performed for all pedestrians  $\alpha$  and for several different starting times  $t$ , using a fixed parameter set for the social force model.

Each simulation run is performed according to the following scheme:

1. Define a starting point and calculate the state (position  $\mathbf{r}_\alpha$ , velocity  $\mathbf{v}_\alpha$ , and acceleration  $\mathbf{a}_\alpha = d\mathbf{v}_\alpha/dt$ ) for each pedestrian  $\alpha$ .
2. Assign a desired speed  $v_\alpha^0$  to each pedestrian, e.g. the maximum speed during the pedestrian tracking time. This is sufficiently accurate, if the overall pedestrian



Pedestrian, Crowd and Evacuation Dynamics, Figure 1

Video tracking used to extract the trajectories of pedestrians from video recordings close to two escalators (after [45]). **a** Illustration of the tracking of pedestrian heads. **b** Resulting trajectories after being transformed onto the two-dimensional plane

density is not too high and the desired speed is constant in time.

3. Assign a desired goal point for each pedestrian, e. g. the end point of the trajectory.
4. Given the tracked motion of the surrounding pedestrians  $\beta$ , simulate the trajectory of pedestrian  $\alpha$  over a time period  $T$  based on the social force model, starting at the actual location  $\mathbf{r}_\alpha(t)$ .

After each simulation run, one determines the relative distance error

$$\frac{\|\mathbf{r}_\alpha^{\text{simulated}}(t+T) - \mathbf{r}_\alpha^{\text{tracked}}(t+T)\|}{\|\mathbf{r}_\alpha^{\text{tracked}}(t+T) - \mathbf{r}_\alpha^{\text{tracked}}(t)\|}. \quad (10)$$

After averaging the relative distance errors over the pedestrians  $\alpha$  and starting times  $t$ , 1 minus the result can be taken as measure of the goodness of fit (the “fitness”) of the parameter set used in the pedestrian simulation. Hence, the best possible value of the “fitness” is 1, but any deviation from the real pedestrian trajectories implies lower values.

One result of such a parameter optimization is that, for each video, there is a broad range of parameter combinations of  $A$  and  $B$  which perform almost equally well [45]. This allows one to apply additional goal functions in the parameter optimization, e. g. to determine among the best performing parameter values such parameter combinations, which perform well for *several* video recordings, using a fitness function which equally weights the fitness reached in each single video. This is how the parameter values listed in Table 1 were determined. It turns out that, in order to reach a good model performance, the pedestrian interaction force must be specified velocity dependent, as in the elliptical model.

Note that our evolutionary fitting method can be also used to determine interaction laws without prespecified

Pedestrian, Crowd and Evacuation Dynamics, Table 1  
Interaction strength  $A$  and interaction range  $B$  resulting from our evolutionary parameter calibration for the circular and elliptical specification of the interaction forces between pedestrians (see main text). The calibration was based on three different video recordings, one for low crowd density, one for medium, and one for high density. The parameter values are specified as mean value  $\pm$  standard deviation. The best fitness value obtained with the elliptical specification for the video with the lowest crowd density was as high as 0.9

Model	A	B	“Fitness”
Extrapolation	0	–	0.34
Circular	0.11 $\pm$ 0.06	0.84 $\pm$ 0.63	0.35
Elliptical	4.30 $\pm$ 3.91	1.07 $\pm$ 1.35	0.53

interaction functions. For example, one can obtain the distance dependence of pedestrian interactions without a prespecified function. For this, one adjusts the values of the force at given distances  $d_k = kd_1$  (with  $k \in \{1, 2, 3, \dots\}$ ) in an evolutionary way. To get some smoothness, linear interpolation is applied. The resulting fit curve is presented in Fig. 2 (left). It turns out that the empirical dependence of the force with distance can be well fitted by an exponential decay.

### Angular Dependence

A closer study of pedestrian interactions reveals that these are not isotropic, but dependent on the angle  $\varphi_{\alpha\beta}$  of the encounter, which is given by the formula

$$\cos(\varphi_{\alpha\beta}) = \frac{\mathbf{v}_\alpha}{\|\mathbf{v}_\alpha\|} \cdot \frac{-\mathbf{d}_{\alpha\beta}}{\|\mathbf{d}_{\alpha\beta}\|}. \quad (11)$$

Generally, pedestrians show little response to pedestrians behind them. This can be reflected by an angular-dependent prefactor  $w(\varphi_{\alpha\beta})$  of the interaction force [45]. Empirical results are represented in Fig. 2 (right). Reasonable results are obtained for the following specification of the prefactor:

$$w(\varphi_{\alpha\beta}(t)) = \left( \lambda_\alpha + (1 - \lambda_\alpha) \frac{1 + \cos(\varphi_{\alpha\beta})}{2} \right), \quad (12)$$

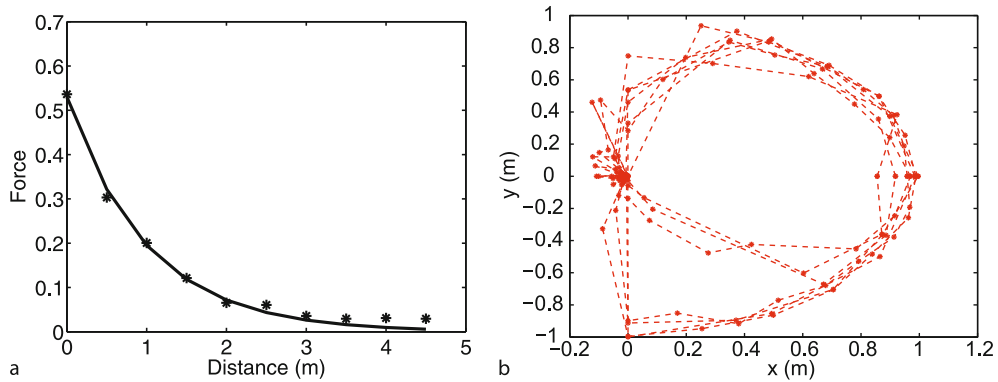
where  $\lambda_\alpha$  with  $0 \leq \lambda_\alpha \leq 1$  is a parameter which grows with the strength of interactions from behind. An evolutionary parameter optimization gives values  $\lambda \approx 0.1$  [45], i. e. a strong anisotropy. With such an angle-dependent prefactor, the “fitness” of the elliptical force increases from 0.53 to 0.61, when calibrated to the same set of videos. Other angular-dependent specifications split up the interaction force between pedestrians into a component against the direction of motion and another one perpendicular to it. Such a description allows for even smoother avoidance maneuvers.

### Crowd Dynamics

#### Analogies with Gases, Fluids, and Granular Media

When the density is low, pedestrians can move freely, and the observed crowd dynamics can be partially compared with the behavior of gases. At medium and high densities, however, the motion of pedestrian crowds shows some striking analogies with the motion of fluids:

1. Footprints of pedestrians in snow look similar to streamlines of fluids [15].



Pedestrian, Crowd and Evacuation Dynamics, Figure 2

Results of an evolutionary fitting of pedestrian interactions. **a** Empirically determined distance dependence of the interaction force between pedestrians (after [45]). An exponential decay fits the empirical data quite well. The dashed fit curve corresponds to Eq. (5) with the parameters  $A = 0.53$  and  $B = 1.0$ . **b** Angular dependence of the influence of other pedestrians. The direction along the positive  $x$ -axis corresponds to the walking direction of pedestrians,  $y$  to the perpendicular direction

2. At borderlines between opposite directions of walking one can observe “viscous fingering” [50,51].
3. The emergence of pedestrian streams through standing crowds [7,37,52] appears analogous to the formation of river beds [53,54].

At high densities, however, the observations have rather analogies with driven granular flows. This will be elaborated in more detail in Sects. “Force Model for Panicking Pedestrians” and “Collective Phenomena in Panic Situations”. In summary, one could say that fluid-dynamic analogies work reasonably well in normal situations, while granular aspects dominate at extreme densities. Nevertheless, the analogy is limited, since the self-driven motion and the violation of momentum conservation imply special properties of pedestrian flows. For example, one usually does not observe eddies, which typically occur in regular fluids at high enough Reynolds numbers.

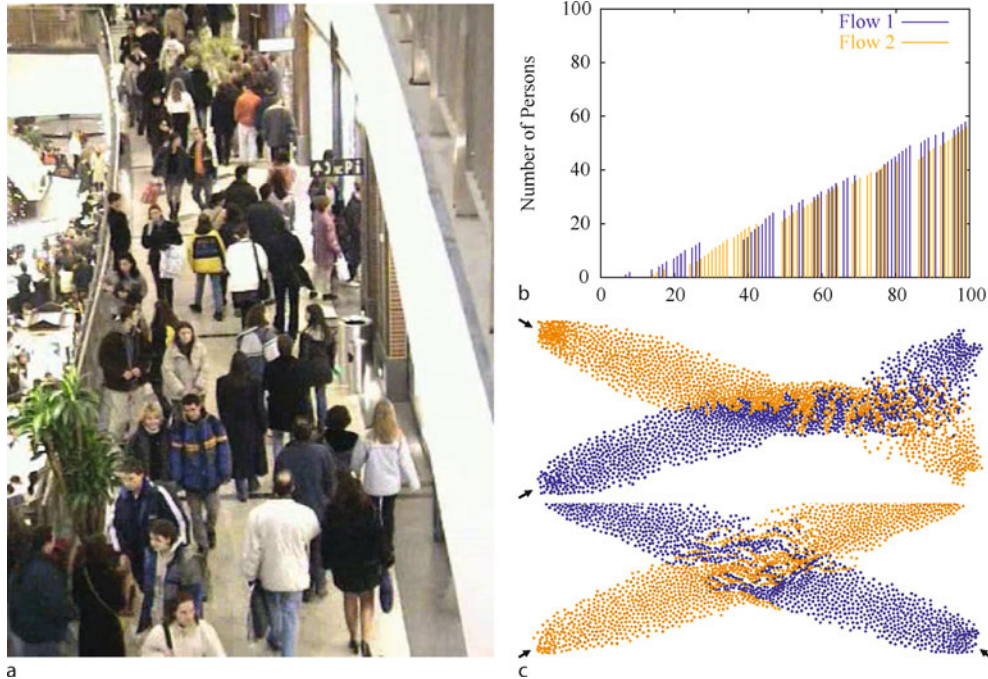
### Self-Organization of Pedestrian Crowds

Despite its simplifications, the social force model of pedestrian dynamics describes a lot of observed phenomena quite realistically. Especially, it allows one to explain various self-organized spatio-temporal patterns that are not externally planned, prescribed, or organized, e. g. by traffic signs, laws, or behavioral conventions. Instead, the spatio-temporal patterns discussed below emerge due to the non-linear interactions of pedestrians even without assuming strategical considerations, communication, or imitative behavior of pedestrians. Despite this, we may still interpret the forming cooperation patterns as phenomena that establish social order on short time scales. It is actu-

ally surprising that strangers coordinate with each other within seconds, if they have grown up in a similar environment. People from different countries, however, are sometimes irritated about local walking habits, which indicates that learning effects and cultural backgrounds still play a role in social interactions as simple as random pedestrian encounters. Rather than on particular features, however, in the following we will focus on the common, internationally reproducible observations.

**Lane Formation** In pedestrian flows one can often observe that oppositely moving pedestrians are forming lanes of uniform walking direction (see Fig. 3). This phenomenon even occurs when there is not a large distance to separate each other, e. g. on zebra crossings. However, the width of lanes increases (and their number decreases), if the interaction continues over longer distances (and if perturbations, e. g. by flows entering or leaving on the sides, are low; otherwise the phenomenon of lane formation may break down [55]).

Lane formation may be viewed as *segregation phenomenon* [56,57]. Although there is a weak preference for one side (with the corresponding behavioral convention depending on the country), the observations can only be well reproduced when repulsive pedestrian interactions are taken into account. The most relevant factor for the lane formation phenomenon is the higher relative velocity of pedestrians walking in opposite directions. Compared to people following each other, oppositely moving pedestrians have more frequent interactions until they have segregated into separate lanes by stepping aside whenever another pedestrian is encountered. The most long-lived pat-



Pedestrian, Crowd and Evacuation Dynamics, Figure 3

**Self-organization of pedestrian crowds.** **a** Photograph of lanes formed in a shopping center. Computer simulations reproduce the self-organization of such lanes very well. **b** Evaluation of the cumulative number of pedestrians passing a bottleneck from different sides. One can clearly see that the narrowing is often passed by groups of people in an oscillatory way rather than one by one. **c** Multi-agent simulation of two crossing pedestrian streams, showing the phenomenon of stripe formation. This self-organized pattern allows pedestrians to pass the other stream without having to stop, namely by moving sideways in a forwardly moving stripe. (After [8])

terns of motion are the ones which change the least. It is obvious that such patterns correspond to lanes, as they minimize the frequency and strength of avoidance maneuvers. Interestingly enough, as computer simulations show, lane formation occurs also when there is no preference for any side.

Lanes minimize frictional effects, accelerations, energy consumption, and delays in oppositely moving crowds. Therefore, one could say that they are a pattern reflecting “collective intelligence”. In fact, it is not possible for a single pedestrian to reach such a collective pattern of motion. Lane formation is a self-organized collaborative pattern of motion originating from simple pedestrian interactions. Particularly in cases of no side preference, the system behavior cannot be understood by adding up the behavior of the single individuals. This is a typical feature of complex, self-organizing systems and, in fact, a widespread characteristics of social systems. It is worth noting, however, that it does not require a conscious behavior to reach forms of social organization like the segregation of oppositely moving pedestrians into lanes. This organiza-

tion occurs automatically, although most people are not even aware of the existence of this phenomenon.

**Oscillatory Flows at Bottlenecks** At bottlenecks, bidirectional flows of moderate density are often characterized by oscillatory changes in the flow direction (see Fig. 3). For example, one can sometimes observe this at entrances of museums during crowded art exhibitions or at entrances of staff canteens during lunch time. While these oscillatory flows may be interpreted as an effect of friendly behavior (“you go first, please”), computer simulations of the social force model indicate that the collective behavior may again be understood by simple pedestrian interactions. That is, oscillatory flows occur even in the absence of communication. Therefore, they may be viewed as another self-organization phenomenon, which again reduces frictional effects and delays. That is, oscillatory flows have features of “collective intelligence”.

While this may be interpreted as result of a learning effect in a large number of similar situations (a “repeated game”), our simulations suggest an even simpler, “many-

particle” interpretation: Once a pedestrian is able to pass the narrowing, pedestrians with the same walking direction can easily follow. Hence, the number and “pressure” of waiting, “pushy” pedestrians on one side of the bottleneck becomes less than on the other side. This eventually increases their chance to occupy the passage. Finally, the “pressure difference” is large enough to stop the flow and turn the passing direction at the bottleneck. This reverses the situation, and eventually the flow direction changes again, giving rise to oscillatory flows.

**Stripe Formation in Intersecting Flows** In intersection areas, the flow of people often appears to be irregular or “chaotic”. In fact, it can be shown that there are several possible collective patterns of motion, among them rotary and oscillating flows. However, these patterns continuously compete with each other, and a temporarily dominating pattern is destroyed by another one after a short time. Obviously, there has not evolved any social convention that would establish and stabilize an ordered and efficient flow at intersections.

Self-organized patterns of motion, however, are found in situations where pedestrian flows cross each other only in two directions. In such situations, the phenomenon of

stripe formation is observed [58]. Stripe formation allows two flows to penetrate each other without requiring the pedestrians to stop. For an illustration see Fig. 3. Like lanes, stripes are a segregation phenomenon, but not a stationary one. Instead, the stripes are density waves moving into the direction of the sum of the directional vectors of both intersecting flows. Naturally, the stripes extend sideways into the direction which is perpendicular to their direction of motion. Therefore, the pedestrians move forward *with* the stripes and sideways *within* the stripes. Lane formation corresponds to the particular case of stripe formation where both directions are exactly opposite. In this case, no intersection takes place, and the stripes do not move systematically. As in lane formation, stripe formation allows to minimize obstructing interactions and to maximize the average pedestrian speeds, i. e. simple, repulsive pedestrian interactions again lead to an “intelligent” collective behavior.

### Evacuation Dynamics

While the previous section has focused on the dynamics of pedestrian crowds in normal situations, we will now turn to the description of situations in which extreme crowd

Pedestrian, Crowd and Evacuation Dynamics, Table 2

Incomplete list of major crowd disasters since 1970 after J. F. Dickie in [59], <http://www.crowddynamics.com/Main/Crowddisasters.html>, [http://SportsIllustrated.CNN.com/soccer/world/news/2000/07/09/stadium\\_disasters\\_ap/](http://SportsIllustrated.CNN.com/soccer/world/news/2000/07/09/stadium_disasters_ap/), and other internet sources, excluding fires, bomb attacks, and train or plane accidents. The number of injured people was usually a multiple of the fatalities

Date	Place	Venue	Deaths	Reason
1971	Ibrox, UK	Stadium	66	Collapse of barriers
1974	Cairo, Egypt	Stadium	48	Crowds break barriers
1982	Moscow, USSR	Stadium	340	Re-entering fans after last minute goal
1988	Katmandu, Nepal	Stadium	93	Stampede due to hailstorm
1989	Hillsborough, Sheffield, UK	Stadium	96	Fans trying to force their way into the stadium
1990	New York City	Bronx	87	Illegal happy land social club
1990	Mena, Saudi Arabia	Pedestrian Tunnel	1426	Overcrowding
1994	Mena, Saudi Arabia	Jamarat Bridge	266	Overcrowding
1996	Guatemala City, Guatemala	Stadium	83	Fans trying to force their way into the stadium
1998	Mena, Saudi Arabia		118	Overcrowding
1999	Kerala, India	Hindu Shrine	51	Collapse of parts of the shrine
1999	Minsk, Belarus	Subway Station	53	Heavy rain at rock concert
2001	Ghana, West Africa	Stadium	> 100	Panic triggered by tear gas
2004	Mena, Saudi Arabia	Jamarat Bridge	251	Overcrowding
2005	Wai, India	Religious Procession	150	Overcrowding (and fire)
2005	Bagdad, Iraq	Religious Procession	> 640	Rumors regarding suicide bomber
2005	Chennai, India	Disaster Area	42	Rush for flood relief supplies
2006	Mena, Saudi Arabia	Jamarat Bridge	363	Overcrowding
2006	Pilippines	Stadium	79	Rush for game show tickets
2006	Ibb, Yemen	Stadium	51	Rally for Yemeni president

densities occur. Such situations may arise at mass events, particularly in cases of urgent egress. While most evacuations run relatively smoothly and orderly, the situation may also get out of control and end up in terrible crowd disasters (see Table 2). In such situations, one often speaks of “panic”, although, from a scientific standpoint, the use of this term is rather controversial. Here, however, we will not be interested in the question whether “panic” actually occurs or not. We will rather focus on the issue of crowd dynamics at high densities and under psychological stress.

### Evacuation and Panic Research

Computer models have been also developed for emergency and evacuation situations [32,60,61,62,63,64,65,66,67,68]. Most research into panic, however, has been of empirical nature (see, e. g. [69,70,71,72]), carried out by social psychologists and others.

With some exceptions, panic is observed in cases of scarce or dwindling resources [69,73], which are either required for survival or anxiously desired. They are usually distinguished into escape panic (“stampedes”, bank or stock market panic) and acquisitive panic (“crazes”, speculative manias) [74,75], but in some cases this classification is questionable [76].

It is often stated that panicking people are obsessed by short-term personal interests uncontrolled by social and cultural constraints [69,74]. This is possibly a result of the reduced attention in situations of fear [69], which also causes that options like side exits are mostly ignored [70]. It is, however, mostly attributed to social contagion [69,71,73,74,75,76,77,78,79,80,81], i. e., a transition from individual to mass psychology, in which individuals transfer control over their actions to others [75], leading to conformity [82]. This “herding behavior” is in some sense irrational, as it often leads to bad overall results like dangerous overcrowding and slower escape [70,75,76]. In this way, herding behavior can increase the fatalities or, more generally, the damage in the crisis faced.

The various socio-psychological theories for this contagion assume hypnotic effects, rapport, mutual excitation of a primordial instinct, circular reactions, social facilitation (see the summary by Brown [80]), or the emergence of normative support for selfish behavior [81]. Brown [80] and Coleman [75] add another explanation related to the prisoner’s dilemma [83,84] or common goods dilemma [85], showing that it is reasonable to make one’s subsequent actions contingent upon those of others. However, the socially favorable behavior of walking orderly is unstable, which normally gives rise to rushing by everyone. These thoughtful considerations are well compatible

with many aspects discussed above and with the classical experiments by Mintz [73], which showed that jamming in escape situations depends on the reward structure (“payoff matrix”).

Nevertheless and despite of the frequent reports in the media and many published investigations of crowd disasters (see Table 2), a quantitative understanding of the observed phenomena in panic stampedes was lacking for a long time. In the following, we will close this gap.

### Situations of “Panic”

Panic stampede is one of the most tragic collective behaviors [71,72,73,74,75,77,78,79,80,81], as it often leads to the death of people who are either crushed or trampled down by others. While this behavior may be comprehensible in life-threatening situations like fires in crowded buildings [69,70], it is hard to understand in cases of a rush for good seats at a pop concert [76] or without any obvious reasons. Unfortunately, the frequency of such disasters is increasing (see Table 2), as growing population densities combined with easier transportation lead to greater mass events like pop concerts, sport events, and demonstrations. Nevertheless, systematic empirical studies of panic [73,86] are rare [69,74,76], and there is a scarcity of quantitative theories capable of predicting crowd dynamics at extreme densities [32,60,61,64,65,68]. The following features appear to be typical [46,55]:

1. In situations of escape panic, individuals are getting nervous, i. e. they tend to develop blind actionism.
2. People try to move considerably faster than normal [9].
3. Individuals start pushing, and interactions among people become physical in nature.
4. Moving and, in particular, passing of a bottleneck frequently becomes incoordinated [73].
5. At exits, jams are building up [73]. Sometimes, intermittent flows or arching and clogging are observed [9], see Fig. 4.
6. The physical interactions in jammed crowds add up and can cause dangerous pressures up to 4,500 Newtons per meter [59,70], which can bend steel barriers or tear down brick walls.
7. The strength and direction of the forces acting in large crowds can suddenly change [87], pushing people around in an uncontrollable way. This may cause people to fall.
8. Escape is slowed down by fallen or injured people turning into “obstacles”.
9. People tend to show herding behavior, i. e., to do what other people do [69,78].





Pedestrian, Crowd and Evacuation Dynamics, Figure 4  
Panicking football fans trying to escape the football stadium in Sheffield. Because of a clogging effect, it is difficult to pass the open door

10. Alternative exits are often overlooked or not efficiently used in escape situations [69,70].

The following quotations give a more personal impression of the conditions during crowd panic:

1. "They just kept pushin' forward and they would just walk right on top of you, just trample over ya like you were a piece of the ground." (After the panic at "The Who Concert Stampede" in Cincinnati.)
2. "People were climbin' over people ta get in ... an' at one point I almost started hittin' 'em, because I could not believe the animal, animalistic ways of the people, you know, nobody cared." (After the panic at "The Who Concert Stampede".)
3. "Smaller people began passing out. I attempted to lift one girl up and above to be passed back ... After several tries I was unsuccessful and near exhaustion." (After the panic at "The Who Concert Stampede".)
4. "I couldn't see the floor because of the thickness of the smoke." (After the "Hilton Hotel Fire" in Las Vegas.)
5. "The club had two exits, but the young people had access to only one", said Narend Singh, provincial minister for agriculture and environmental affairs. However, the club's owner, Rajan Naidoo, said the club had four exits, and that all were open. "I think the children panicked and headed for the main entrance where they initially came in," he said." (After the "Durban Disco Stampede".)
6. "At occupancies of about 7 persons per square meter the crowd becomes almost a fluid mass. Shock waves can be propagated through the mass, sufficient to ... propel them distances of 3 meters or more. ... People

may be literally lifted out of their shoes, and have clothing torn off. Intense crowd pressures, exacerbated by anxiety, make it difficult to breathe, which may finally cause compressive asphyxia. The heat and the thermal insulation of surrounding bodies cause some to be weakened and faint. Access to those who fall is impossible. Removal of those in distress can only be accomplished by lifting them up and passing them overhead to the exterior of the crowd." (J. Fruin in [88].)

7. "It was like a huge wave of sea gushing down on the pilgrims" (P. K. Abdul Ghafour, Arab News, after the sad crowd disaster in Mena on January 12, 2006).

### Force Model for Panicking Pedestrians

Additional, physical interaction forces  $f_{\alpha\beta}^{\text{ph}}$  come into play when pedestrians get so close to each other that they have physical contact (i. e.  $d_{\alpha\beta} < r_{\alpha\beta} = r_{\alpha} + r_{\beta}$ , where  $r_{\alpha}$  means the "radius" of pedestrian  $\alpha$ ). In this case, which is mainly relevant to panic situations, we assume also a "body force"  $k(r_{\alpha\beta} - d_{\alpha\beta})\mathbf{n}_{\alpha\beta}$  counteracting body compression and a "sliding friction force"  $\kappa(r_{\alpha\beta} - d_{\alpha\beta})\Delta v_{\beta\alpha}^t \mathbf{t}_{\alpha\beta}$  impeding relative tangential motion. Inspired by the formulas for granular interactions [89,90], we assume

$$\mathbf{f}_{\alpha\beta}^{\text{ph}}(t) = k\Theta(r_{\alpha\beta} - d_{\alpha\beta})\mathbf{n}_{\alpha\beta} + \kappa\Theta(r_{\alpha\beta} - d_{\alpha\beta})\Delta v_{\beta\alpha}^t \mathbf{t}_{\alpha\beta}, \quad (13)$$

where the function  $\Theta(z)$  is equal to its argument  $z$ , if  $z \geq 0$ , otherwise 0. Moreover,  $\mathbf{t}_{\alpha\beta} = (-n_{\alpha\beta}^2, n_{\alpha\beta}^1)$  means the tangential direction and  $\Delta v_{\beta\alpha}^t = (\mathbf{v}_{\beta} - \mathbf{v}_{\alpha}) \cdot \mathbf{t}_{\alpha\beta}$  the tangential velocity difference, while  $k$  and  $\kappa$  represent large constants. (Strictly speaking, friction effects already set in before pedestrians touch each other, because of the psychological tendency not to pass other individuals with a high relative velocity, when the distance is small.)

The interactions with the boundaries of walls and other obstacles are treated analogously to pedestrian interactions, i. e., if  $d_{\alpha i}(t)$  means the distance to obstacle or boundary  $i$ ,  $\mathbf{n}_{\alpha i}(t)$  denotes the direction perpendicular to it, and  $\mathbf{t}_{\alpha i}(t)$  the direction tangential to it, the corresponding interaction force with the boundary reads

$$\mathbf{f}_{\alpha i} = \{A_{\alpha} \exp[(r_{\alpha} - d_{\alpha i})/B_{\alpha}] + k\Theta(r_{\alpha} - d_{\alpha i})\} \times \mathbf{n}_{\alpha i} - \kappa\Theta(r_{\alpha} - d_{\alpha i})(\mathbf{v}_{\alpha} \cdot \mathbf{t}_{\alpha i})\mathbf{t}_{\alpha i}. \quad (14)$$

Finally, fire fronts are reflected by repulsive social forces similar those describing walls, but they are much stronger. The physical interactions, however, are qualitatively different, as people reached by the fire front become injured and immobile ( $\mathbf{v}_{\alpha} = \mathbf{0}$ ).

### Collective Phenomena in Panic Situations

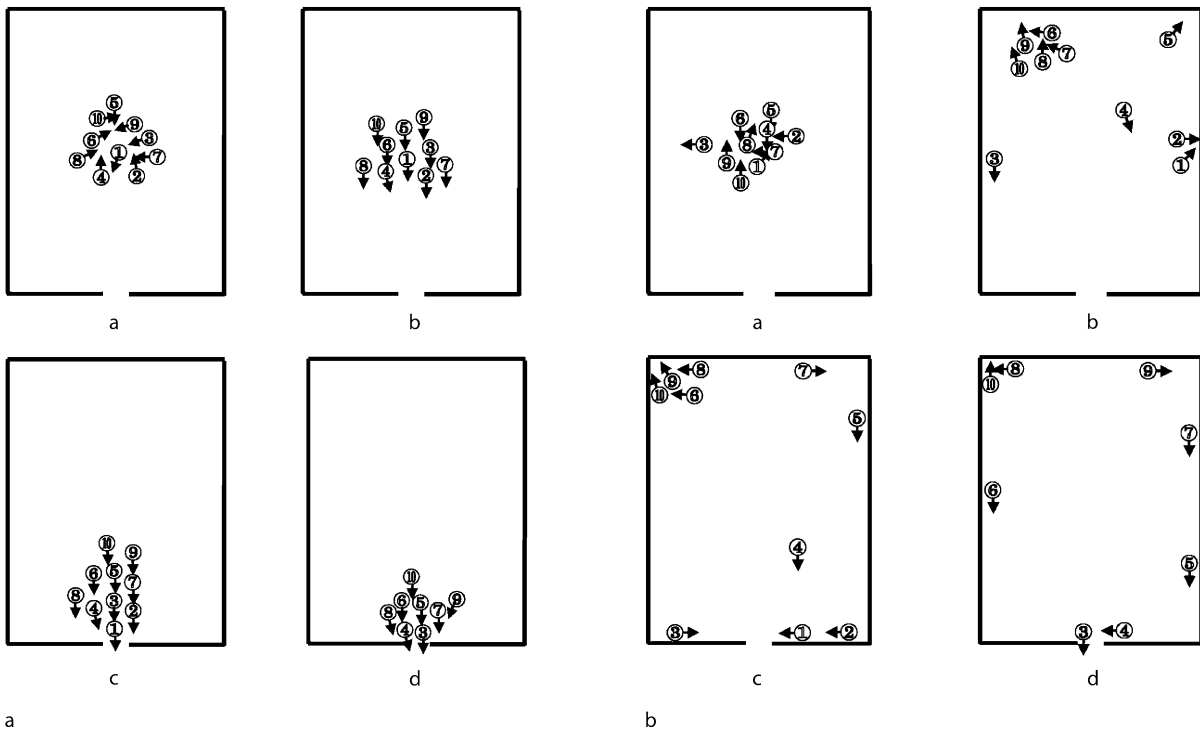
In panic situations (e. g. in some cases of emergency evacuation) the following characteristic features of pedestrian behavior are often observed:

1. People are getting nervous, resulting in a higher level of fluctuations.
2. They are trying to escape from the source of panic, which can be reflected by a significantly higher desired velocity  $v_{\alpha}^0$ .
3. Individuals in complex situations, who do not know what is the right thing to do, orient at the actions of their neighbors, i. e. they tend to do what other people do. We will describe this by an additional herding interaction.

We will now discuss the fundamental collective effects which fluctuations, increased desired velocities, and herding behavior can have. In contrast to other approaches, we do not assume or imply that individuals in panic or emergency situations would behave relentless and asocial, although they sometimes do.

**Herding and Ignorance of Available Exits** If people are not sure what is the best thing to do, there is a tendency to show a “herding behavior”, i. e. to imitate the behavior of others. Fashion, hypes and trends are examples for this. The phenomenon is also known from stock markets, and particularly pronounced when people are anxious. Such a situation is, for example, given if people need to escape from a smoky room. There, the evacuation dynamics is very different from normal leaving (see Fig. 5).

Under normal visibility, everybody easily finds an exit and uses more or less the shortest path. However, when the exit cannot be seen, evacuation is much less efficient and may take a long time. Most people tend to walk relatively straight into the direction in which they suspect an exit, but in most cases, they end up at a wall. Then, they usually move along it in one of the two possible directions, until they finally find an exit [18]. If they encounter others, there is a tendency to take a decision for one direction and move collectively. Also in case of acoustic signals, people may be attracted into the same direction. This can lead to over-crowded exits, while other exits are ignored. The same can happen even for normal visibility, when people



Pedestrian, Crowd and Evacuation Dynamics, Figure 5  
**a** Normal leaving of a room, when the exit is well visible. **b** Escape from a room with no visibility, e. g. due to dense smoke or a power blackout. (After [18])

are not well familiar with their environment and are not aware of the directions of the emergency exits.

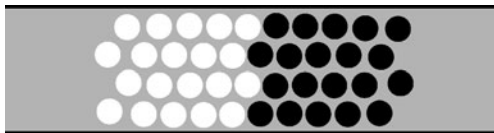
Computer simulations suggest that neither individualistic nor herding behavior performs well [46]. Pure individualistic behavior means that each pedestrian finds an exit only accidentally, while pure herding behavior implies that the complete crowd is eventually moving into the same and probably congested direction, so that available emergency exits are not efficiently used. Optimal chances of survival are expected for a certain mixture of individualistic and herding behavior, where individualism allows *some* people to detect the exits and herding guarantees that successful solutions are imitated by small groups of others [46].

**“Freezing by Heating”** Another effect of getting nervous has been investigated in [55]. Let us assume the individual fluctuation strength is given by

$$\eta_\alpha = (1 - n_\alpha)\eta_0 + n_\alpha\eta_{\max}, \quad (15)$$

where  $n_\alpha$  with  $0 \leq n_\alpha \leq 1$  measures the nervousness of pedestrian  $\alpha$ . The parameter  $\eta_0$  means the normal and  $\eta_{\max}$  the maximum fluctuation strength. It turns out that, at sufficiently high pedestrian densities, lanes are destroyed by increasing the fluctuation strength (which is analogous to the temperature). However, instead of the expected transition from the “fluid” lane state to a disordered, “gaseous” state, a solid state is formed. It is characterized by an at least temporarily blocked, “frozen” situation so that one calls this paradoxical transition “*freezing by heating*” (see Fig. 6). Notably enough, the blocked state has a *higher* degree of order, although the internal energy is *increased* [55].

The preconditions for this unusual freezing-by-heating transition are the driving term  $v_\alpha^0 e_\alpha^0 / \tau_\alpha$  and the dissipative friction  $-v_\alpha / \tau_\alpha$ , while the sliding friction force is not required. Inhomogeneities in the channel diameter or other impurities which temporarily slow down pedestrians can further this transition at the respective places. Finally note that a transition from fluid to blocked pedes-



Pedestrian, Crowd and Evacuation Dynamics, Figure 6  
Result of the noise-induced formation of a “frozen” state in a (periodic) corridor used by oppositely moving pedestrians (after [55])

trian counter flows is also observed, when a critical density is exceeded [31,55].

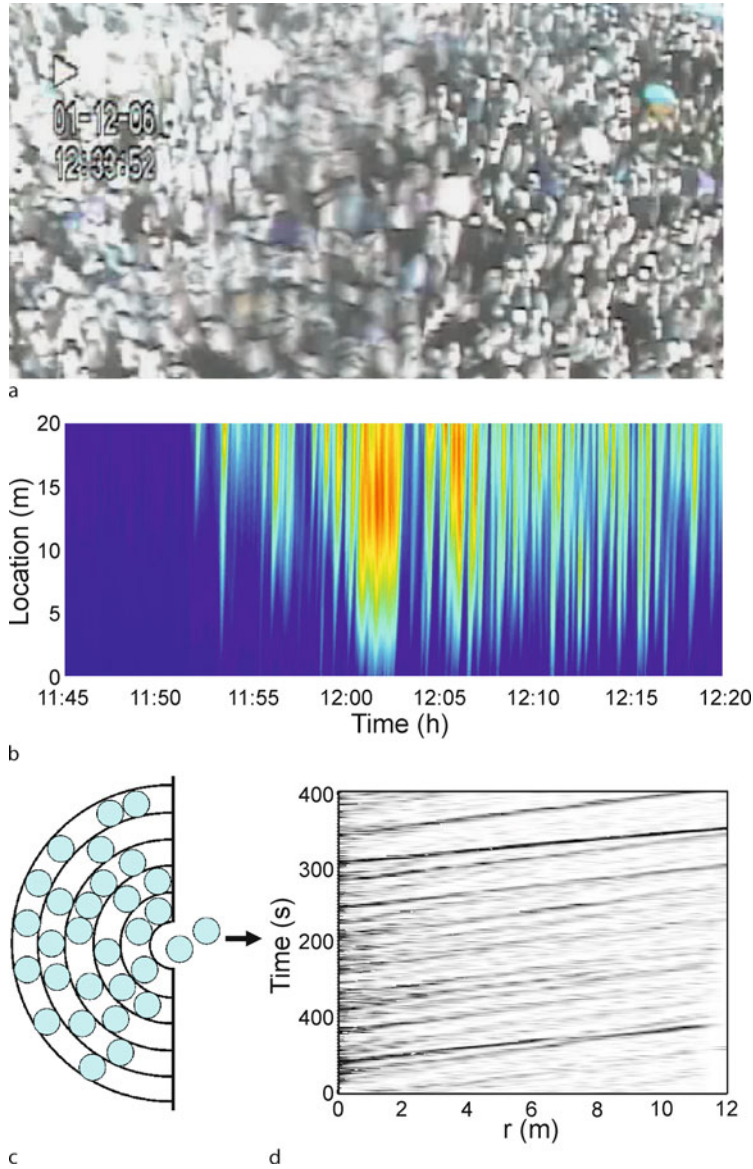
**Intermittent Flows, Faster-Is-Slower Effect, and “Phantom Panic”** If the overall flow towards a bottleneck is higher than the overall outflow from it, a pedestrian queue emerges [91]. In other words, a waiting crowd is formed upstream of the bottleneck. High densities can result, if people keep heading forward, as this eventually leads to higher and higher compressions. Particularly critical situations may occur if the arrival flow is much higher than the departure flow, especially if people are trying to get towards a strongly desired goal (“acquisitive panic”) or away from a perceived source of danger (“escape panic”) with an increased driving force  $v_\alpha^0 e_\alpha^0 / \tau$ . In such situations, the high density causes coordination problems, as several people compete for the same few gaps. This typically causes body interactions and frictional effects, which can slow down crowd motion or evacuation (“*faster is slower effect*”).

A possible consequence of these coordination problems are intermittent flows. In such cases, the outflow from the bottleneck is not constant, but it is typically interrupted. While one possible origin of the intermittent flows are clogging and arching effects as known from granular flows through funnels or hoppers [89,90], stop-and-go waves have also been observed in more than 10 meter wide streets and in the 44 meters wide entrance area to the Jamarat Bridge during the pilgrimage in January 12, 2006 [87], see Fig. 7. Therefore, it seems to be important that people do not move continuously, but have minimum strides [25]. That is, once a person is stopped, he or she will not move until some space opens up in front. However, increasing impatience will eventually reduce the minimum stride, so that people eventually start moving again, even if the outflow through the bottleneck is stopped. This will lead to a further compression of the crowd.

In the worst case, such behavior can trigger a “*phantom panic*”, i. e. a crowd disaster *without* any serious reasons (e. g., in Moscow, 1982). For example, due to the “faster-is-slower effect” panic can be triggered by small pedestrian counterflows [70], which cause delays to the crowd intending to leave. Consequently, stopped pedestrians in the back, who do not see the reason for the temporary slowdown, are getting impatient and pushy. In accordance with observations [7,25], one may model this by increasing the desired velocity, for example, by the formula

$$v_\alpha^0(t) = [1 - n_\alpha(t)]v_\alpha^0(0) + n_\alpha(t)v_\alpha^{\max}. \quad (16)$$

Herein,  $v_\alpha^{\max}$  is the maximum desired velocity and  $v_\alpha^0(0)$  the initial one, corresponding to the expected velocity of



Pedestrian, Crowd and Evacuation Dynamics, Figure 7

**a** Long-term photograph showing stop-and-go waves in a densely packed street. While stopped people appear relatively sharp, people moving from right to left have a fuzzy appearance. Note that gaps propagate from *right* to *left*. **b** Empirically observed stop-and-go waves in front of the entrance to the Jamarat Bridge on January 12, 2006 (after [87]), where pilgrims moved from *left* to *right*. *Dark areas* correspond to phases of motion, *light colors* to stop phases. **c** Illustration of the “shell model”, in particular of situations where several pedestrians compete for the same gap, which causes coordination problems. **d** Stop-and-go waves resulting from the alternation of forward pedestrian motion and backward gap propagation

leaving. The time-dependent parameter

$$n_\alpha(t) = 1 - \frac{\bar{v}_\alpha(t)}{v_\alpha^0(0)} \quad (17)$$

reflects the nervousness, where  $\bar{v}_\alpha(t)$  denotes the average speed into the desired direction of motion. Altogether,

long waiting times increase the desired speed  $v_\alpha^0$  or driving force  $v_\alpha^0(t)e_\alpha^0/\tau$ , which can produce high densities and inefficient motion. This further increases the waiting times, and so on, so that this tragic feedback can eventually trigger so high pressures that people are crushed or falling and trampled. It is, therefore, imperative, to have sufficiently

wide exits and to prevent counterflows, when big crowds want to leave [46].

**Transition to Stop-and-Go Waves** Recent empirical studies of pilgrim flows in the area of Makkah, Saudi Arabia, have shown that intermittent flows occur not only when bottlenecks are obvious. On January 12, 2006, pronounced stop-and-go waves have been even observed upstream of the 44 m wide entrance to the Jamarat Bridge [87]. While the pilgrim flows were smooth and continuous (“laminar”) over many hours, at 11:53 am stop-and-go waves suddenly appeared and propagated over distances of more than 30 m (see Fig. 7). The sudden transition was related to a significant drop of the flow, i. e. with the onset of congestion [87]. Once the stop-and-go waves set in, they persisted over more than 20 min.

This phenomenon can be reproduced by a recent model based on two continuity equations, one for forward pedestrian motion and another one for backward gap propagation [91]. The model was derived from a “shell model” (see Fig. 7) and describes very well the observed alternation between backward gap propagation and forward pedestrian motion.

**Transition to “Crowd Turbulence”** On the same day, around 12:19, the density reached even higher values and the video recordings showed a sudden transition from stop-and-go waves to *irregular* flows (see Fig. 8). These irregular flows were characterized by random, unintended displacements into all possible directions, which pushed

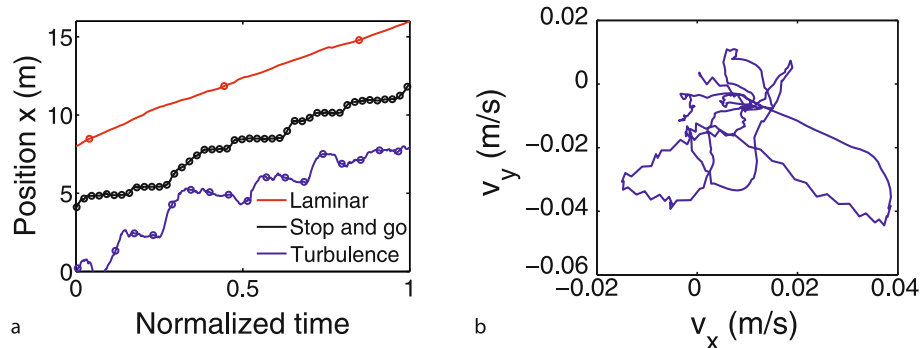
people around. With a certain likelihood, this caused them to stumble. As the people behind were moved by the crowd as well and could not stop, fallen individuals were trampled, if they did not get back on their feet quickly enough. Tragically, the area of trampled people grew more and more in the course of time, as the fallen pilgrims became obstacles for others [87]. The result was one of the biggest crowd disasters in the history of pilgrimage.

How can we understand this transition to irregular crowd motion? A closer look at video recordings of the crowd reveals that, at this time, people were so densely packed that they were moved involuntarily by the crowd. This is reflected by random displacements into all possible directions. To distinguish these irregular flows from laminar and stop-and-go flows and due to their visual appearance, we will refer to them as “*crowd turbulence*”.

As in certain kinds of fluid flows, “turbulence” in crowds results from a sequence of instabilities in the flow pattern. Additionally, one finds a sharply peaked probability density function of velocity increments

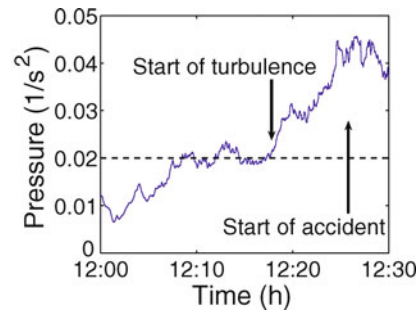
$$V_x^\tau = V_x(\mathbf{r}, t + \tau) - V_x(\mathbf{r}, t), \quad (18)$$

which is typical for turbulence [92], if the time shift  $\tau$  is small enough [87]. One also observes a power-law scaling of the displacements indicating self-similar behavior [87]. As large eddies are not detected, however, the similarity with *fluid* turbulence is limited, but there is still an analogy to turbulence at currency exchange markets [92]. Instead of vortex cascades like in turbulent fluids, one rather finds a hierarchical fragmentation dynamics: At extreme



Pedestrian, Crowd and Evacuation Dynamics, Figure 8

Pedestrian dynamics at different densities. **a** Representative trajectories (space-time plots) of pedestrians during the laminar, stop-and-go, and turbulent flow regime. Each trajectory extends over a range of 8 meters, while the time required for this stretch is normalized to 1. To indicate the different speeds, symbols are included in the curves every 5 seconds. While the laminar flow (*top line*) is fast and smooth, motion is temporarily interrupted in stop-and-go flow (*medium line*), and backward motion can occur in “turbulent” flows (*bottom line*). **b** Example of the temporal evolution of the velocity components  $v_x(t)$  into the average direction of motion and  $v_y(t)$  perpendicular to it in “turbulent flow”, which occurs when the crowd density is extreme. One can clearly see the irregular motion into all possible directions characterizing “crowd turbulence”. For details see [87]



Pedestrian, Crowd and Evacuation Dynamics, Figure 9

**Left:** Snapshot of the on-line visualization of “crowd pressure”. Red colors (see the lower ellipses) indicate areas of critical crowd conditions. In fact, the sad crowd disaster during the Muslim pilgrimage on January 12, 2006, started in this area. **Right:** The “crowd pressure” is a quantitative measure of the onset of “crowd turbulence”. The crowd disaster started when the “crowd pressure” reached particularly high values

densities, individual motion is replaced by mass motion, but there is a stick-slip instability which leads to “rupture” when the stress in the crowd becomes too large. That is, the mass splits up into clusters of different sizes with strong velocity correlations *inside* and distance-dependent correlations *between* the clusters.

“Crowd turbulence” has further specific features [87]. Due to the physical contacts among people in extremely dense crowds, we expect commonalities with granular media. In fact, dense driven granular media may form density waves, while moving forward [93], and can display turbulent-like states [94,95]. Moreover, under quasi-static conditions [94], force chains [96] are building up, causing strong variations in the strengths and directions of local forces. As in earthquakes [97,98] this can lead to events of sudden, uncontrollable stress release with power-law distributed displacements. Such a power-law has also been discovered by video-based crowd analysis [87].

### Some Warning Signs of Critical Crowd Conditions

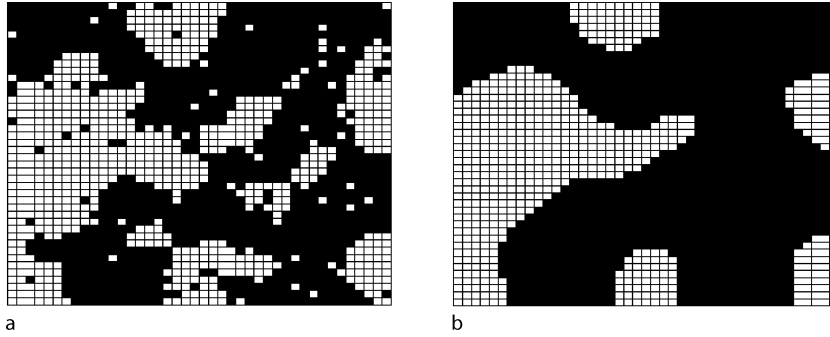
Turbulent waves are experienced in dozens of crowd-intensive events each year all over the world [88]. Therefore, it is necessary to understand why, where and when potentially critical situations occur. Viewing real-time video recordings is not very suited to identify critical crowd conditions: While the average density rarely exceeds values of 6 persons per square meter, the local densities can reach almost twice as large values [87]. It has been found, however, that even evaluating the local densities is not enough to identify the critical times and locations precisely, which also applies to an analysis of the velocity field [87]. The decisive quantity is rather the “crowd pressure”, i. e. the density, multiplied with the variance of speeds. It allows one to identify critical locations and times (see Fig. 9).

There are even advance warning signs of critical crowd conditions: The crowd accident on January 12, 2006 started about 10 minutes after “turbulent” crowd motion set in, i. e. after the “pressure” exceeded a value of  $0.02/s^2$  (see Fig. 9). Moreover, it occurred more than 30 min after stop-and-go waves set in, which can be easily detected in accelerated surveillance videos. Such advance warning signs of critical crowd conditions can be evaluated on-line by an automated video analysis system. In many cases, this can help one to gain time for corrective measures like flow control, pressure-relief strategies, or the separation of crowds into blocks to stop the propagation of shock-waves [87]. Such anticipative crowd control could increase the level of safety during future mass events.

### Evolutionary Optimization of Pedestrian Facilities

Having understood some of the main factors causing crowd disasters, it is interesting to ask how pedestrian facilities can be designed in a way that maximizes the efficiency of pedestrian flows and the level of safety. One of the major goals during mass events must be to avoid extreme densities. These often result from the onset of congestion at bottlenecks, which is a consequence of the breakdown of free flow and causes an increasing degree of compression. When a certain critical density is increased (which depends on the size distribution of people), this potentially implies high pressures in the crowd, particularly if people are impatient due to long delays or panic.

The danger of an onset of congestion can be minimized by avoiding bottlenecks. Notice, however, that jamming can also occur at widenings of escape routes [46]. This surprising fact results from disturbances due to pedestrians, who try to overtake each other and expand in the wider area because of their repulsive interactions.



Pedestrian, Crowd and Evacuation Dynamics, Figure 10

The evolutionary optimization based on Boolean grids [99] uses a two-stage algorithm. **a** In the randomization stage, obstacles are distributed over the grid with some randomness, thereby allowing for the generation of new topologies. **b** In the agglomeration stage, small nearby obstacles are clustered to form larger objects with smooth boundaries

These squeeze into the main stream again at the end of the widening, which acts like a bottleneck and leads to jamming. The corresponding drop of efficiency  $E$  is more pronounced,

1. if the corridor is narrow,
2. if the pedestrians have different or high desired velocities, and
3. if the pedestrian density in the corridor is high.

Obviously, the emerging pedestrian flows decisively depend on the geometry of the boundaries. They can be simulated on a computer already in the planning phase of pedestrian facilities. Their configuration and shape can be systematically varied, e. g. by means of evolutionary algorithms [28,100] and evaluated on the basis of particular mathematical performance measures [7]. Apart from the *efficiency*

$$E = \frac{1}{N} \sum_{\alpha} \frac{\mathbf{v}_{\alpha} \cdot \mathbf{e}_{\alpha}^0}{v_{\alpha}^0} \quad (19)$$

we can, for example, define the *measure of comfort*  $C = (1 - D)$  via the discomfort

$$D = \frac{1}{N} \sum_{\alpha} \frac{(\mathbf{v}_{\alpha} - \overline{\mathbf{v}_{\alpha}})^2}{(v_{\alpha}^0)^2} = \frac{1}{N} \sum_{\alpha} \left( 1 - \frac{\overline{v_{\alpha}^2}}{(v_{\alpha}^0)^2} \right). \quad (20)$$

The latter is again between 0 and 1 and reflects the frequency and degree of sudden velocity changes, i. e. the level of discontinuity of walking due to necessary avoidance maneuvers. Hence, the optimal configuration regarding the pedestrian requirements is the one with the highest values of efficiency and comfort.

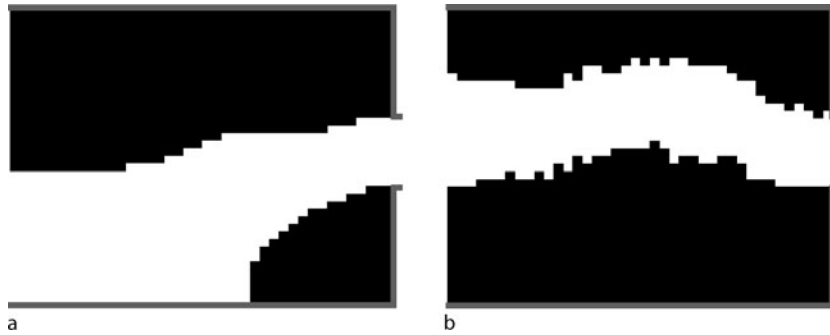
During the optimization procedure, some or all of the following can be varied:

1. the location and form of planned buildings,
2. the arrangement of walkways, entrances, exits, staircases, elevators, escalators, and corridors,
3. the shape of rooms, corridors, entrances, and exits,
4. the function and time schedule. (Recreation rooms or restaurants are often continuously frequented, rooms for conferences or special events are mainly visited and left at peak periods, exhibition rooms or rooms for festivities require additional space for people standing around, and some areas are claimed by queues or through traffic.)

In contrast to early evolutionary optimization methods, recent approaches allow to change not only the dimensions of the different elements of pedestrian facilities, but also to vary their topology. The procedure of such algorithms is illustrated in Fig. 10. Highly performing designs are illustrated in Fig. 11. It turns out that, for an emergency evacuation route, it is favorable if the crowd does not move completely straight towards a bottleneck. For example, a zigzag design of the evacuation route can reduce the pressure on the crowd upstream of a bottleneck (see Fig. 12). The proposed evolutionary optimization procedure can, of course, not only be applied to the design of new pedestrian facilities, but also to a reduction of existing bottlenecks, when suitable modifications are implemented.

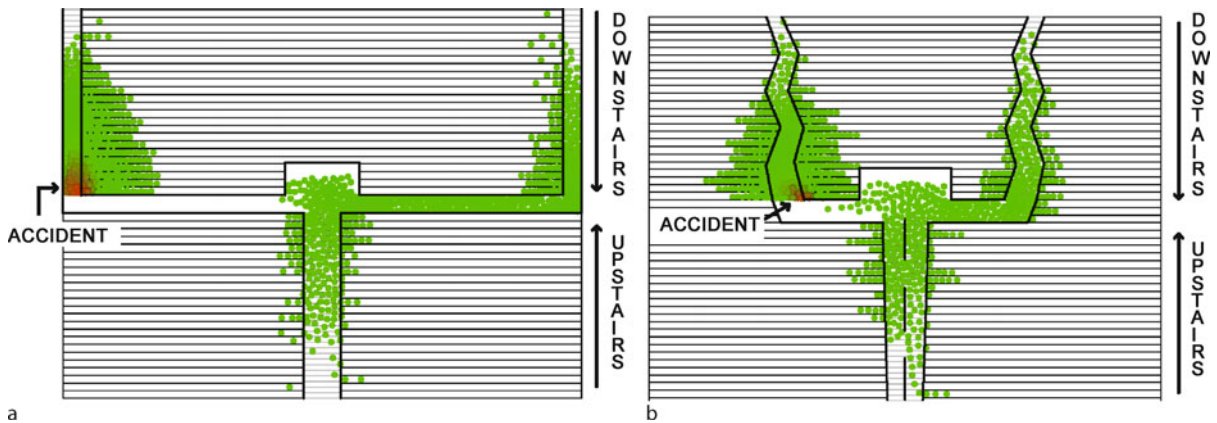
### Future Directions

In this contribution, we have presented a multi-agent approach to pedestrian and crowd dynamics. Despite the great effort required, pedestrian interactions can be well quantified by video tracking. Compared to other social interactions they turn out to be quite simple. Neverthe-



Pedestrian, Crowd and Evacuation Dynamics, Figure 11

Two examples of improved designs for cases with a bottleneck along the escape route of a large crowd, obtained with an evolutionary algorithm based on Boolean grids. People were assumed to move from left to right only. **a** Funnel-shaped escape route. **b** Zig-zag design



Pedestrian, Crowd and Evacuation Dynamics, Figure 12

**a** Conventional design of a stadium exit in an emergency scenario, where we assume that some pedestrians have fallen at the end of the downwards staircase to the left. The dark color indicates high pressures, since pedestrians are impatient and pushing from behind. **b** In the improved design, the increasing diameter of corridors can reduce waiting times and impatience (even with the same number of seats), thereby accelerating evacuation. Moreover, the zigzag design of the downwards staircases changes the pushing direction in the crowd. (After [8])

less, they cause a surprisingly large variety of self-organized patterns and short-lived social phenomena, where coordination or cooperation emerges spontaneously. For this reason, they are interesting to study, particularly as one can expect new insights into coordination mechanisms of social beings beyond the scope of classical game theory. Examples for observed self-organization phenomena in normal situations are lane formation, stripe formation, oscillations and intermittent clogging effects at bottlenecks, and the evolution of behavioral conventions (such as the preference of the right-hand side in continental Europe). Under extreme conditions (high densities or panic), however, coordination may break down, giving rise to “freezing-by-heating” or “faster-is-slower effects”, stop-and-go waves or “crowd turbulence”.

Similar observations as in pedestrian crowds are made in other social systems and settings. Therefore, we expect that realistic models of pedestrian dynamics will also promote the understanding of opinion formation and other kinds of collective behaviors. The hope is that, based on the discovered elementary mechanisms of emergence and self-organization, one can eventually also obtain a better understanding of the constituting principles of more complex social systems. At least the same underlying factors are found in many social systems: Non-linear interactions of individuals, time-dependence, heterogeneity, stochasticity, competition for scarce resources (here: Space and time), decision-making, and learning. Future work will certainly also address issues of perception, anticipation, and communication.



## Acknowledgments

The authors are grateful for partial financial support by the German Research Foundation (research projects He 2789/7-1, 8-1) and by the “Cooperative Center for Communication Networks Data Analysis”, a NAP project sponsored by the Hungarian National Office of Research and Technology under grant No. KCKHA005.

## Bibliography

### Primary Literature

- Hankin BD, Wright RA (1958) Passenger flow in subways. *Operat Res Q* 9:81–88
- Older SJ (1968) Movement of pedestrians on footways in shopping streets. *Traffic Eng Control* 10:160–163
- Weidmann U (1993) *Transporttechnik der Fußgänger*. In: Schriftenreihe des Instituts für Verkehrsplanung, Transporttechnik, Straßen- und Eisenbahnbau. Institut für Verkehrsplanung, Transporttechnik, Straßen- und Eisenbahnbau, Zürich
- Fruin JJ (1971) Designing for pedestrians: A level-of-service concept. In: Highway research record, Number 355: Pedestrians. Highway Research Board, Washington DC, pp 1–15
- Pauls J (1984) The movement of people in buildings and design solutions for means of egress. *Fire Technol* 20:27–47
- Whyte WH (1988) *City. Rediscovering the center*. Doubleday, New York
- Helbing D (1997) *Verkehrsdynamik*. Springer, Berlin
- Helbing D, Buzna L, Johansson A, Werner T (2005) Self-organized pedestrian crowd dynamics: Experiments, simulations, and design solutions. *Transport Sci* 39(1):1–24
- Predtetschenski WM, Milinski AI (1971) *Personenströme in Gebäuden – Berechnungsmethoden für die Projektierung*. Müller, Köln-Braunsfeld
- Transportation Research Board (1985) *Highway Capacity Manual, Special Report 209*. Transportation Research Board, Washington DC
- Yuhaski SJ Jr, Macgregor Smith JM (1989) Modelling circulation systems in buildings using state dependent queueing models. *Queueing Syst* 4:319–338
- Garbrecht D (1973) Describing pedestrian and car trips by transition matrices. *Traffic Q* 27:89–109
- Ashford N, O’Leary M, McGinity PD (1976) Stochastic modelling of passenger and baggage flows through an airport terminal. *Traffic Engin Control* 17:207–210
- Borgers A, Timmermans H (1986) City centre entry points, store location patterns and pedestrian route choice behaviour: A microlevel simulation model. *Socio-Econ Plan Sci* 20:25–31
- Helbing D (1993) *Stochastische Methoden, nichtlineare Dynamik und quantitative Modelle sozialer Prozesse*. Ph.D. thesis University of Stuttgart, 1992 (published by Shaker, Aachen)
- Helbing D, Isobe M, Nagatani T, Takimoto K (2003) Lattice gas simulation of experimentally studied evacuation dynamics. *Phys Rev E* 67:067101
- Daamen W, Hoogendoorn SP (2003) Experimental research on pedestrian walking behavior (CDROM). In: Proceedings of the 82nd annual meeting at the transportation research board, Washington DC
- Isobe M, Helbing D, Nagatani T (2004) Experiment, theory, and simulation of the evacuation of a room without visibility. *Phys Rev E* 69:066132
- Seyfried A, Steffen B, Klingsch W, Boltes M (2005) The fundamental diagram of pedestrian movement revisited. *J Stat Mech* P10002
- Kretz T, Wölki M, Schreckenberg M (2006) Characterizing correlations of flow oscillations at bottlenecks. *J Stat Mech* P02005
- Henderson LF (1974) On the fluid mechanics of human crowd motion. *Transp Res* 8:509–515
- Hughes RL (2002) A continuum theory for the flow of pedestrians. *Transp Res B* 36:507–535
- Helbing D (1992) A fluid-dynamic model for the movement of pedestrians. *Complex Syst* 6:391–415
- Hoogendoorn SP, Bovy PHL (2000) Gas-kinetic modelling and simulation of pedestrian flows. *Transp Res Rec* 1710:28–36
- Helbing D (1991) A mathematical model for the behavior of pedestrians. *Behav Sci* 36:298–310
- Helbing D, Molnár P (1995) Social force model for pedestrian dynamics. *Phys Rev E* 51:4282–4286
- Gipps PG, Marksjö B (1985) A micro-simulation model for pedestrian flows. *Math Comp Simul* 27:95–105
- Bolay K (1998) *Nichtlineare Phänomene in einem fluid-dynamischen Verkehrsmodell*. Master’s thesis, University of Stuttgart
- Blue VJ, Adler JL (1998) Emergent fundamental pedestrian flows from cellular automata microsimulation. *Transp Res Rec* 1644:29–36
- Fukui M, Ishibashi Y (1999) Self-organized phase transitions in cellular automaton models for pedestrians. *J Phys Soc Japan* 68:2861–2863
- Muramatsu M, Irie T, Nagatani T (1999) Jamming transition in pedestrian counter flow. *Physica A* 267:487–498
- Klüpfel H, Meyer-König M, Wahle J, Schreckenberg M (2000) Microscopic simulation of evacuation processes on passenger ships. In: Bandini S, Worsch T (eds) *Theory and practical issues on cellular automata*. Springer, London
- Burstedde C, Klauck K, Schadschneider A, Zittartz J (2001) Simulation of pedestrian dynamics using a 2-dimensional cellular automaton. *Physica A* 295:507–525
- Gopal S, Smith TR (1990) NAVIGATOR: An AI-based model of human way-finding in an urban environment. In: Fischer MM, Nijkamp P, Papageorgiou YY (eds) *Spatial choices and processes*. North-Holland, Amsterdam, pp 169–200
- Reynolds CW (1994) Evolution of corridor following behavior in a noisy world. In: Cliff D, Husbands P, Meyer J-A, Wilson S (eds) *From animals to animats 3: Proceedings of the third international conference on simulation of adaptive behavior*. MIT Press, Cambridge, pp 402–410
- Helbing D (1992) A mathematical model for attitude formation by pair interactions. *Behav Sci* 37:190–214
- Helbing D, Molnár P, Farkas I, Bolay K (2001) Self-organizing pedestrian movement. *Env Planning B* 28:361–383
- Klockgether J, Schwefel H-P (1970) Two-phase nozzle and hollow core jet experiments. In: Elliott DG (ed) *Proceedings of the eleventh symposium on engineering aspects of magnetohydrodynamics*. California Institute of Technology, Pasadena, pp 141–148

39. Helbing D (1992) A mathematical model for behavioral changes by pair interactions. In: Haag G, Mueller U, Troitzsch KG (eds) *Economic evolution and demographic change. Formal models in social sciences*. Springer, Berlin, pp 330–348
40. Miller NE (1944) *Experimental studies of conflict*. In: Mc Hunt VJ (ed) *Personality and the behavior disorders*, vol 1. Ronald, New York
41. Miller NE (1959) Liberalization of basic S-R-concepts: Extension to conflict behavior, motivation, and social learning. In: Koch S (ed) *Psychology: A study of science*, vol 2. McGraw Hill, New York
42. Lewin K (1951) *Field theory in social science*. Harper, New York
43. Helbing D (1994) A mathematical model for the behavior of individuals in a social field. *J Math Sociol* 19(3):189–219
44. Hoogendoorn S, Bovy PHL (2003) Simulation of pedestrian flows by optimal control and differential games. *Optim Control Appl Meth* 24(3):153–172
45. Johansson A, Helbing D, Shukla PK (2007) Specification of the social force pedestrian model by evolutionary adjustment to video tracking data. *Adv Complex Syst* 10:271–288
46. Helbing D, Farkas I, Vicsek T (2000) Simulating dynamical features of escape panic. *Nature* 407:487–490
47. Kerridge J, Chamberlain T (2005) Collecting pedestrian trajectory data in real-time. In: Waldau N, Gattermann P, Knoflach H, Schreckenberg M (eds) *Pedestrian and evacuation dynamics '05*. Springer, Berlin
48. Hoogendoorn SP, Daamen W, Bovy PHL (2003) Extracting microscopic pedestrian characteristics from video data (CDROM). In: *Proceedings of the 82nd annual meeting at the transportation research board*. Mira Digital, Washington DC
49. Teknomo K (2002) *Microscopic pedestrian flow characteristics: Development of an image processing data collection and simulation model*. Ph D thesis, Tohoku University Japan
50. Kadanoff LP (1985) Simulating hydrodynamics: A pedestrian model. *J Stat Phys* 39:267–283
51. Stanley HE, Ostrowsky N (eds) (1986) *On growth and form*. Nijhoff, Boston
52. Arns T (1993) *Video films of pedestrian crowds*. Stuttgart
53. Stølum H-H (1996) River meandering as a self-organization process. *Nature* 271:1710–1713
54. Rodríguez-Iturbe I, Rinaldo A (1997) *Fractal river basins: Chance and self-organization*. Cambridge University, Cambridge
55. Helbing D, Farkas I, Vicsek T (2000) Freezing by heating in a driven mesoscopic system. *Phys Rev Lett* 84:1240–1243
56. Schelling T (1971) Dynamic models of segregation. *J Math Sociol* 1:143–186
57. Helbing D, Platkowski T (2000) Self-organization in space and induced by fluctuations. *Int J Chaos Theory Appl* 5(4):47–62
58. Ando K, Oto H, Aoki T (1988) Forecasting the flow of people. *Railw Res Rev* 45(8):8–13 (in Japanese)
59. Smith RA, Dickie JF (eds) (1993) *Engineering for crowd safety*. Elsevier, Amsterdam
60. Dräger KH, Løvås G, Wiklund J, Soma H, Duong D, Violas A, Lanérés V (1992) EVACSIM – A comprehensive evacuation simulation tool. In: *The proceedings of the 1992 Emergency Management and Engineering Conference*. Society for Computer Simulation, Orlando, pp 101–108
61. Ebihara M, Ohtsuki A, Iwaki H (1992) A model for simulating human behavior during emergency evacuation based on classificatory reasoning and certainty value handling. *Microcomput Civ Engin* 7:63–71
62. Ketchell N, Cole S, Webber DM, Marriott CA, Stephens PJ, Brearley IR, Fraser J, Doheny J, Smart J (1993) The EGRESS code for human movement and behaviour in emergency evacuations. In: Smith RA, Dickie JF (eds) *Engineering for crowd safety*. Elsevier, Amsterdam, pp 361–370
63. Okazaki S, Matsushita S (1993) A study of simulation model for pedestrian movement with evacuation and queuing. In: Smith RA, Dickie JF (eds) *Engineering for crowd safety*. Elsevier, Amsterdam, pp 271–280
64. Still GK (1993) New computer system can predict human behaviour response to building fires. *Fire* 84:40–41
65. Still GK (2000) *Crowd dynamics*. Ph.D. thesis, University of Warwick
66. Thompson PA, Marchant EW (1993) Modelling techniques for evacuation. In: Smith RA, Dickie JF (eds) *Engineering for crowd safety*. Elsevier, Amsterdam, pp 259–269
67. Løvås GG (1998) On the importance of building evacuation system components. *IEEE Trans Engin Manag* 45:181–191
68. Hamacher HW, Tjandra SA (2001) Mathematical modelling of evacuation problems: A state of the art. In: Schreckenberg M, Sharma SD (eds) *Pedestrian and evacuation dynamics*. Springer, Berlin, pp 227–266
69. Keating JP (1982) The myth of panic. *Fire J* 57–61, 147
70. Elliott D, Smith D (1993) Football stadia disasters in the United Kingdom: Learning from tragedy? *Ind Env Crisis Q* 7(3):205–229
71. Jacobs BD, 't Hart P (1992) Disaster at Hillsborough Stadium: A comparative analysis. In: Parker DJ, Handmer JW (eds) *Hazard management and emergency planning*, Chapt 10. James and James Science, London
72. Canter D (ed) (1990) *Fires and human behaviour*. Fulton, London
73. Mintz A (1951) Non-adaptive group behavior. *J Abnorm Norm Soc Psychol* 46:150–159
74. Miller DL (1985) Introduction to collective behavior (Fig. 3.3 and Chap. 9). Wadsworth, Belmont
75. Coleman JS (1990) *Foundations of social theory*, Chaps. 9 and 33. Belknap, Cambridge
76. Johnson NR (1987) Panic at "The Who Concert Stampede": An empirical assessment. *Soc Probl* 34(4):362–373
77. LeBon G (1960) *The crowd*. Viking, New York
78. Quarantelli E (1957) The behavior of panic participants *Sociol Soc Res* 41:187–194
79. Smelser NJ (1963) *Theory of collective behavior*. Free Press, New York
80. Brown R (1965) *Social psychology*. Free Press, New York
81. Turner RH, Killian LM (1987) *Collective behavior*, 3rd edn. Prentice Hall, Englewood Cliffs
82. Bryan JL (1985) Convergence clusters. *Fire J* 27–30, 86–90
83. Axelrod R, Hamilton WD (1981) The evolution of cooperation. *Science* 211:1390–1396
84. Axelrod R, Dion D (1988) The further evolution of cooperation. *Science* 242:1385–1390
85. Glance NS, Huberman BA (1994) The dynamics of social dilemmas. *Scientific American* 270:76–81
86. Kelley HH, Condry JC Jr, Dahlke AE, Hill AH (1965) Collective behavior in a simulated panic situation. *J Exp Soc Psychol* 1:20–54

87. Helbing D, Johansson A, Al-Abideen HZ (2007) The dynamics of crowd disasters: An empirical study. *Phys Rev E* 75:046109
88. Fruin JJ (1993) The causes and prevention of crowd disasters. In: Smith RA, Dickie JF (eds) *Engineering for crowd safety*. Elsevier, Amsterdam, pp 99–108
89. Ristow GH, Herrmann HJ (1994) Density patterns in two-dimensional hoppers. *Phys Rev E* 50:R5–R8
90. Wolf DE, Grassberger P (eds) (1997) *Friction, arching, contact dynamics*. World Scientific, Singapore
91. Helbing D, Johansson A, Mathiesen J, Jensen HM, Hansen A (2006) Analytical approach to continuous and intermittent bottleneck flows. *Phys Rev Lett* 97:168001
92. Ghashghaie S, Breymann W, Peinke J, Talkner P, Dodge Y (1996) Turbulent cascades in foreign exchange markets. *Nature* 381:767–770
93. Peng G, Herrmann HJ (1994) Density waves of granular flow in a pipe using lattice-gas automata. *Phys Rev E* 49:R1796–R1799
94. Radjai F, Roux S (2002) Turbulentlike fluctuations in quasi-static flow of granular media. *Phys Rev Lett* 89:064302
95. Sreenivasan KR (1990) Turbulence and the tube. *Nature* 344:192–193
96. Cates ME, Wittmer JP, Bouchaud J-P, Claudin P (1998) Jamming, force chains, and fragile matter. *Phys Rev Lett* 81:1841–1844
97. Bak P, Christensen K, Danon L, Scanlon T (2002) Unified scaling law for earthquakes. *Phys Rev Lett* 88:178501
98. Johnson PA, Jia X (2005) Nonlinear dynamics, granular media and dynamic earthquake triggering. *Nature* 437:871–874
99. Johansson A, Helbing D (2007) Pedestrian flow optimization with a genetic algorithm based on Boolean grids. In: Waldau N, Gattermann P, Knoflacher H, Schreckenberg M (eds) *Pedestrian and evacuation dynamics 2005*. Springer, Berlin, pp 267–272
100. Baeck T (1996) *Evolutionary algorithms in theory and practice*. Oxford University Press, New York

### Books and Reviews

- Decicco PR (ed) (2001) *Evacuation from fires*. Baywood, Amityville
- Helbing D (2001) Traffic and related self-driven many-particle systems. *Rev Mod Phys* 73:1067–1141
- Helbing D, Molnár P, Farkas I, Bolay K (2001) Self-organizing pedestrian movement. *Environ Plan B* 28:361–383
- Helbing D, Buzna L, Johansson A, Werner T (2005) Self-organized pedestrian crowd dynamics: Experiments, simulations, and design solutions. *Transp Sci* 39(1):1–24
- Le Bon G (2002) *The Crowd*. Dover, New York (1st edn: 1895)
- Predtechenskii VM, Milinskii AI (1978) *Planning for foot traffic flow in buildings*. Amerind, New Delhi
- Schreckenberg M, Sharma SD (eds) (2002) *Pedestrian and evacuation dynamics*. Springer, Berlin
- Smith RA, Dickie JF (eds) (1993) *Engineering for crowd safety*. Elsevier, Amsterdam
- Still GK (2000) *Crowd Dynamics*. Ph.D thesis, University of Warwick
- Surowiecki J (2005) *The Wisdom of Crowds*. Anchor, New York
- Tubbs J, Meacham B (2007) *Egress design solutions: A guide to evacuation and crowd management planning*. Wiley, New York
- Waldau N, Gattermann P, Knoflacher H (eds) (2006) *Pedestrian and evacuation dynamics 2005*. Springer, Berlin
- Weidmann U (1993) *Transporttechnik der Fußgänger*. In: *Schriftenreihe des Institut für Verkehrsplanung, Transporttechnik, Straßen- und Eisenbahnbau* 90. ETH Zürich

# Percolation, and Faults and Fractures in Rock

PIERRE M. ADLER<sup>1</sup>, JEAN-FRANÇOIS THOVERT<sup>2</sup>,  
VALERI V. MOURZENKO<sup>2</sup>

<sup>1</sup> UPMC-Sisyphé, Paris, France

<sup>2</sup> CNRS-LCD, Chasseneuil du Poitou, France

## Article Outline

Glossary

Definition of the Subject

Introduction

Fracture Networks

Percolation of Fracture Networks

Determination of the Dimensionless Density  
from Experimental Data

Role of the Dimensionless Density  
in Other Geometrical Properties and Permeability

Future Directions

Bibliography

## Glossary

**Dimensionless density** The dimensionless density is the number of objects per excluded volume.

**Excluded volume** The *excluded volume*  $V_{\text{ex}}$  of an object is defined as the volume surrounding it, in which the center of another object must be in order for them to intersect.

**Fracture network** A fracture network is generally defined as a set of individual fractures which may or may not intersect.

**Percolation and percolation threshold** Percolation is defined as the existence of a spanning connected cluster in the fracture network. Percolation occurs when the number of fractures per unit volume is equal or larger than a certain value called the percolation threshold.

**Plane convex fractures** A plane fracture is convex if for any points  $A$  and  $B$  which belong to the fracture, all the points of the segment  $AB$  belong to the fracture.

## Definition of the Subject

The study of fractured porous media is of great practical and theoretical importance. It has been first generated by the fact that the presence of fractures can change completely the macroscopic properties of porous media which are present for instance in oil reservoirs, aquifers or waste repositories. The first contributions to this subject were made from very different standpoints by Barenblatt and

coworkers [7], Conrad and Jacquin [13], and Witherspoon and coworkers (see for instance [22]).

In the eighties, these studies were renewed by concepts such as percolation and fractals. Fracture networks were first addressed in the framework of continuum percolation by [12] and [4].

Since the mid nineties, important progress have been made in this field thanks to systematic numerical experiments which can be rationalized by using the concept of excluded volume.

## Introduction

Knowledge of geometrical properties of fracture networks is crucial to the understanding of flow and other transport processes in geological formations, both at small and large scales; introduction of fractures in a porous rock matrix seriously alters the macroscopic properties of the formation. Moreover, studies of fracture geometries during the last 30 years show that naturally occurring geological fractures exist on scales ranging from a few mm to several kilometers [33]. Therefore, fracture networks are likely to influence the transports on a large range of scales. Because of their importance, fracture networks are studied and applied in various areas such as oil and gas recovery, hydrology, nuclear waste storage and geothermal energy exploitation.

Geological fractures can be defined as discrete discontinuities within a rock mass; these breaks are characterized by the fact that their local aperture (defined as the local distance between the two surfaces which limit the fracture) is significantly smaller than their lateral extent; in other words, when they are viewed from far away, fractures can be assimilated to surfaces of discontinuity; in most cases, these surfaces are relatively plane. Fractures have varying degrees of aperture, and may in some cases be completely closed either because of deposition of material induced by fluid flow, or by displacements of the matrix.

An important property of these fracture sets or fracture network is their connectivity and their percolation properties. If a network percolates, fluid can circulate only through it and most likely much more rapidly than in the surrounding porous medium itself. Connectivity studies of fracture networks were initiated in 3D by Charlaix et al. [12] and Balberg [4].

The purpose of this paper is to provide a complete and updated view of the percolation properties of fracture networks in rocks. It is organized as follows. In Sect. “**Fracture Networks**”, fractures are modeled as plane convex polygons which enables the introduction of the concept of excluded volume  $V_{\text{ex}}$ . This volume is a simple function of

the surface and perimeter of the fractures, and it enables to introduce a dimensionless fracture density  $\rho'$  which is defined as the number of fractures per excluded volume. The tools necessary for the numerical study of the percolation thresholds are detailed and applied to mono- and poly-disperse fracture networks. It is shown that when expressed in terms of  $\rho'$ , the percolation threshold does not depend anymore on the fracture shapes. This crucial property is presented and discussed.

Section “**Determination of the Dimensionless Density from Experimental Data**” is devoted to the determination of the dimensionless density from experimental data. In most cases, these data are based on 1D and 2D measurements of fracture traces along boreholes or on exposed outcrops. These measurements necessitate extrapolation by stereological techniques to three dimensions. Significant progress can be made for plane convex fractures. Some recent applications of the methodology are given.

Finally, the independence of the dimensionless percolation threshold on the fracture shape can be extended to other properties such as other geometric properties and the macroscopic permeability of fractured rocks. These extensions are summarized in Sect. “**Role of the Dimensionless Density in Other Geometrical Properties and Permeability**”.

## Fracture Networks

A fracture network is generally defined as a set of individual fractures which may or may not intersect.

On a scale large with respect to the fracture aperture, fractures are usually modeled as convex, finite polygons possibly based on an embedding disk as shown in Fig. 1a.

This is only a simplifying assumption which however provides a standard starting point for studying fracture networks. Convex polygons can be used to analyze shape and area dependencies of geometrical and topological features in the fracture systems in a systematic way.

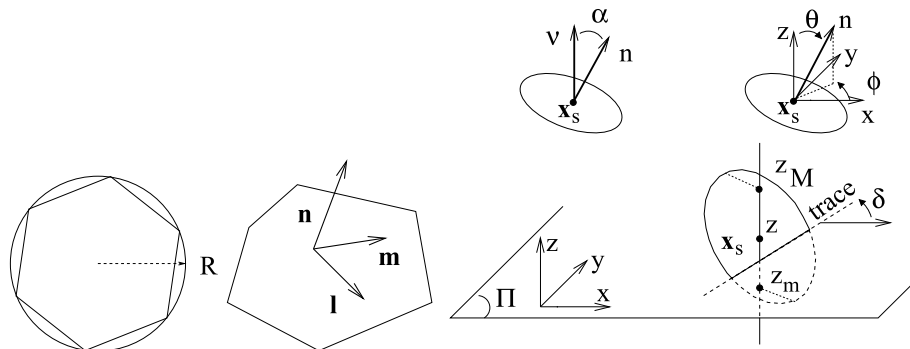
The individual fractures are characterized by their orientation. This orientation is usually given by two unit vectors  $\mathbf{n}$  and  $\mathbf{m}$  (cf. Fig. 1).  $\mathbf{n}$  is the normal to the fracture plane;  $\mathbf{m}$  gives the orientation of the polygon in the fracture plane.

The simplest model consists of a network in which all fractures have the same shape and are inscribed in a circle with a given radius  $R$ . The normal vectors  $\mathbf{n}$  are uniformly distributed on the unit sphere. The density  $\rho$  of this isotropic monodisperse network is defined as the number of fractures per unit volume. An illustration of such a fracture system is shown in Fig. 2a.

Next consider three-dimensional networks made up of polydisperse fractures with plane polygonal shapes. These polygons may be regular or not, but all their vertices are supposed to lie on their circumscribed circle, whose radius  $R$  provides a measure of their size. In agreement with many observations of fractured rocks [2], the statistical distribution of the fracture sizes is supposed to be given by a power law

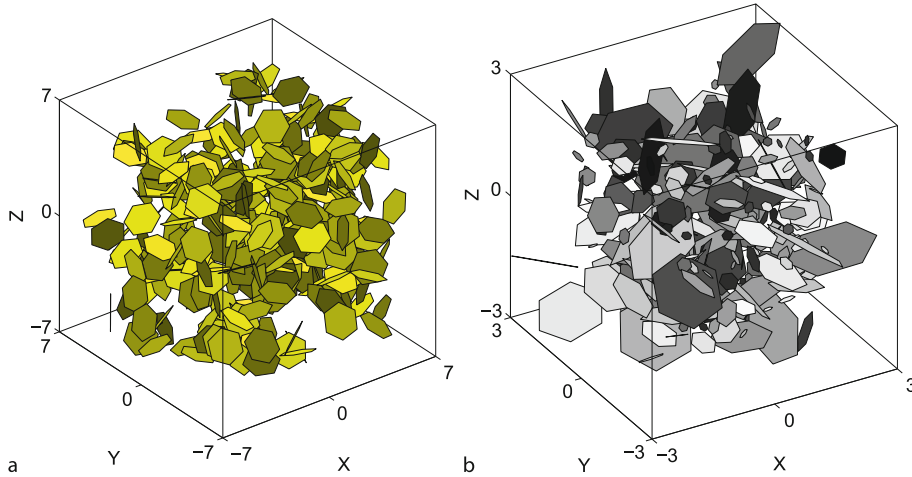
$$n(R) = \alpha R^{-a} \quad (1)$$

where  $n(R)dR$  is the probability of fracture radii in the range  $[R, R + dR]$ ;  $\alpha$  is a normalization coefficient, and the exponent  $a$  ranges between 1 and 5. In practice,  $R$  may vary over a large interval which can span five orders of magnitude, from the size  $R_m$  of the microcracks to the size  $R_M$  of the largest fractures in the system. The normalization



Percolation, and Faults and Fractures in Rock, Figure 1

**Notations.** Convex polygons such as hexagons are created within a circle of radius  $R$  (a). This polygon requires two unit vectors  $\mathbf{n}$  and  $\mathbf{m}$  to be oriented in space (b);  $\mathbf{l}$  is a unit vector perpendicular to  $\mathbf{n}$  and  $\mathbf{m}$ . c illustrates the notations which are mostly used in Sect. “**Determination of the Dimensionless Density from Experimental Data**”



Percolation, and Faults and Fractures in Rock, Figure 2  
 Examples of three-dimensional fracture networks. **a** Monodisperse network made of identical polygons. The volume of size  $L^3$  contains 495 hexagons;  $L = 12R$  where  $R$  is the radius of the circle in which the hexagon is inscribed. **b** Polydisperse network of hexagonal fractures, with  $L' = 4$ ,  $a = 1.5$ ,  $R'_m = 0.1$ , which contains  $N_{fr} = 300$  fractures ( $\rho'_{21} = 1.25$ ,  $\rho'_3 = 2.44$ ). The unit for the coordinates is  $R_M$

condition implies that  $\alpha$  verifies

$$\alpha = \frac{a - 1}{R_m^{1-a} - R_M^{1-a}} \quad (a \neq 1); \tag{2}$$

$$\alpha = \frac{1}{\ln R_M - \ln R_m} \quad (a = 1).$$

The definition of the network density  $\rho$  for polydisperse networks should be modified. To this end, we introduce the volumetric number density of fracture per fracture size  $F(R)$ ,

$$F(R) = \rho n(R) \tag{3}$$

where  $F(R)dR$  is the number of fractures with radius in the range  $[R, R + dR]$  per unit volume.

An example of such polydisperse networks is given in Fig. 2b.

### Percolation of Fracture Networks

#### General Considerations on Continuum Percolation

**Continuum Percolation** Percolation, i. e., the existence of a spanning connected cluster in the fracture network, is a crucial topological property which conditions many other geometrical or transport properties of the network.

Percolation of discrete sites or bonds lattices has been closely studied (see, e. g. [32,37]). In these lattices, the sites or bonds are occupied with a probability  $p$ , which can be interpreted as a concentration. In large systems, percolation occurs when  $p$  exceeds a critical value  $p_c$ , known as

the percolation threshold, which depends on the underlying lattice structure. For  $p$  close to  $p_c$ , however, many geometrical or transport coefficients are known to scale as power laws of the difference  $p - p_c$ , according to the standard form

$$X \propto (p - p_c)^\alpha. \tag{4}$$

The quantity  $X$  may represent the correlation length, the fraction of sites connected to the infinite cluster, or the conductivity of the system. Different exponents are associated with the various quantities, but each is generally believed to be universal, i. e., insensitive to the details of the underlying lattice.

It is, of course, tempting to try to transpose this theoretical framework to the problem at hand. It is intuitively obvious that a fracture network will start percolating if some critical concentration is reached. The main difficulty is to define an equivalent of the probability  $p$  in discrete lattices. As shown below, this can be done by using the concept of excluded volume, introduced by [6] in the context of fracture networks.

Fracture networks belong to the general class of continuum percolation systems. Applications of continuum percolation concepts to geophysical problems have been reviewed by [9]. Continuum percolation differs from lattice percolation in several respects. First, the occupancy probability  $p$  in a discrete lattice ranges between 0 and 1, which means that there is a maximal concentration; the filling of the system can be defined relative to this upper

bound. In continuum percolation, there is generally no such upper bound. For instance, there is ideally no upper limit to the degree of fracturation of a piece of rock. Consequently, the relative concentration  $p$  has to be replaced by a volumetric density. Second, any site or bond in lattice percolation cannot have more than a maximum number  $z$  of neighbors, called the lattice coordination number, whereas there is no limitation to the number of intersections for a fracture in a network. Other differences result from the variable lengths and orientations of the bonds, in contrast with the discrete set of values imposed by a lattice, which may be significant for transport properties (see [5]).

Note that in this section we only consider “large” systems, i. e., the size of the objects in the percolation system may have a broad distribution, but the overall domain extension is supposed to widely exceed the size of the largest objects it contains. This condition may sometimes be difficult to fulfill; natural fracture networks often involve large-scale faults, which may in themselves ensure percolation if they cross the domain of investigation. [11] and later [24] considered such broad size-distributions where the probability of a spanning single fracture is non-zero.

In view of the previous considerations, two definitions of the system concentration appear possible. One is volumetric, quantified by the average number of objects in a reference volume; the other is topological, defined as the average number of connections with surrounding objects. These two definitions are nicely reconciled by the introduction of the concept of excluded volume.

The *excluded volume*  $V_{\text{ex}}$  of an object was defined by [6] as the volume surrounding it, in which the center of another object must be in order for them to intersect. We first discuss the simplest case of populations of identical objects, with volume  $V$ . For example, the excluded volumes of a sphere with volume  $V$  in 3D and of disks with area  $A$  in 2D are

$$\begin{aligned} V_{\text{ex}} &= 8V \quad \text{for spheres;} \\ A_{\text{ex}} &= 4A \quad \text{for disks in the plane.} \end{aligned} \quad (5)$$

These equations are also valid for any object with convex shape, if all the objects in the population have identical orientations.

If the objects are anisotropic and have distributed orientations, the excluded volume has to be averaged over all possible relative orientations of the intersecting objects.

Now suppose that the volumetric density of objects per unit volume is  $\rho$ . It is natural to use  $V_{\text{ex}}$  as a reference volume, and we may define the dimensionless density  $\rho'$  as the number of objects per volume  $V_{\text{ex}}$

$$\rho' = \rho V_{\text{ex}}. \quad (6)$$

On the other hand, the definition of  $V_{\text{ex}}$  implies that  $\rho'$  is also the average number of intersections per object, if they are randomly located according to a Poisson process. Therefore, given the shape of the object and its orientation distribution (and thus  $V_{\text{ex}}$ ), the definition (6) incorporates both the volumetric and topologic aspects mentioned above.

It should be emphasized however, that the definition of the excluded volume is meaningful only if the object locations are uniformly distributed in space. If there are spatial correlations, they should be replaced by a spatial integral of the pair separation distribution function (see for instance [14] for applications to the physics of liquids).

### Calculation of the Excluded Volume for Plane Convex Fractures

A general expression for the excluded volume was established very early in the context of statistical mechanics by [19], for isotropically oriented objects. For two three-dimensional convex objects  $A$  and  $B$  with volumes  $V_A$  and  $V_B$ , areas  $A_A$  and  $A_B$  and surface averaged mean radius of curvature  $R_A$  and  $R_B$ , [19] obtained the mutual exclusion volume

$$V_{\text{ex},AB} = V_A + V_B + (A_A R_B + A_B R_A). \quad (7)$$

This expression can then be averaged over the distributions of object shapes and sizes. For equal spheres, Eq. (5) is obtained. For flat convex objects randomly oriented in space with perimeters  $P_A$  and  $P_B$ , it is reduced to [12]

$$V_{\text{ex},AB} = \frac{1}{4} (A_A P_B + A_B P_A). \quad (8)$$

On averaging (8) over the size distribution of objects with identical shapes, one obtains

$$V_{\text{ex}} = \frac{1}{2} \langle A \rangle \langle P \rangle \quad (9)$$

where  $\langle \cdot \rangle$  is the statistical average. If  $A$  and  $B$  are identical, (9) yields

$$V_{\text{ex}}^{\text{iso}} = \frac{1}{2} A P. \quad (10)$$

If the population of polygons is not isotropic and has a probability distribution  $n(f)$ , which may involve the shape or the size of the polygons, the average of (8) yields

$$\begin{aligned} V_{\text{ex}}^{\text{iso}} &= \frac{1}{4} \iint n(F_1) n(F_2) (A_1 P_2 + A_2 P_1) dF_1 dF_2 \\ &= \frac{1}{2} \langle A \rangle \langle P \rangle \end{aligned} \quad (11)$$

where  $\langle A \rangle$  and  $\langle P \rangle$  are the average area and perimeter. Alternatively, the polygon orientation may be incorporated into  $n(f)$  and a general expression of  $V_{\text{ex}}$  can be obtained.

### Determinations of Continuum Percolation Thresholds

The percolation thresholds of various simple continuous systems have been determined, since the pioneering papers of [35] and [27]. These early works were reviewed by [6] and [5]. A few examples are given in Table 1, for monodisperse objects in a  $d$ -dimensional space. The critical concentration is described in terms of the average number  $\rho'_c$  of connections per object.

The influence of the orientation distribution was investigated by [29,30], and [6]. For sticks with constant length in the plane, [29] has shown that  $\rho'_c$  is identical for uniform orientation distributions in any angular sector and equal to the value 3.6 for an isotropic distribution. On the other hand, the value 3.2 for orthogonal sticks is also valid for any bimodal orientation distribution. By considering three-dimensional systems [6], also conclude that the total excluded volume at percolation is independent of the degree of anisotropy. [4] proposed a set of bounds which correspond to orthogonal and parallel object systems

$$3.2 \leq \rho'_c \leq 4.5 \quad d = 2; \quad 0.7 \leq \rho'_c \leq 2.8 \quad d = 3. \quad (12)$$

All these results were obtained by numerical simulations. One should also mention the heuristic criterion developed by [3]. They define the average “bonding distance”  $l$  as the mean distance between connected objects, which is essentially the gyration radius of the excluded volume

$$l^2 = \frac{1}{V_{\text{ex}}} \int_{V_{\text{ex}}} r^2 d^3 \mathbf{r}. \quad (13)$$

Note that  $l$  does not depend on the density of objects. They then postulate that percolation occurs when the average distance  $L_d$  between objects with at least two neighbors is smaller than or equal to  $2l$ . To evaluate  $L_d$ , they note that

Percolation, and Faults and Fractures in Rock, Table 1  
Thresholds  $\rho'_c$  for various continuum percolation systems in  $d$  dimensions

$d$	Object type	$\rho'_c$	$d$	Object type	$\rho'_c$
2	Orthogonal sticks	3.2	3	Orthogonal elongated rods	0.7
2	Randomly oriented sticks	3.6	3	Randomly oriented elongated rods	1.4
2	Disks or parallel objects	4.5	3	Orthogonal squares	2.0
			3	Randomly oriented squares	2.46
			3	Spheres or parallel objects	2.80

the number  $k$  of connections to a given object corresponds to a Poisson distribution

$$\text{Pr}(k) = \frac{\rho'^k}{k!} e^{-\rho'}. \quad (14)$$

Therefore, the density  $\rho_2$  of objects with at least two neighbors is

$$\rho_2 = \rho \left[ 1 - (1 + \rho') e^{-\rho'} \right]. \quad (15)$$

Thus, an estimate of  $L_d$  follows from

$$\frac{4}{3} \pi \left( \frac{L_d}{2} \right)^3 = \frac{1}{\rho_2}. \quad (16)$$

An equation for the critical concentration  $\rho'_c$  can be directly deduced from the statement that  $L_d = 2l$ . Although the argument is not substantiated, it is quite successful. It yields directly  $\rho'_c = 2.80$  for spheres. In two dimensions,  $L_d$  is replaced by the average distance between objects with at least 5 neighbors.

An interesting feature of this argument is that it can be easily generalized to account for spatial correlations. If  $\rho g(r)$  denotes the probability density of finding an object center at a distance  $r$  from an object located at the origin, the bonding distance is defined by the weighted average

$$l^2 = \frac{\int_{V_{\text{ex}}} r^2 g(r) d^3 \mathbf{r}}{\int_{V_{\text{ex}}} g(r) d^3 \mathbf{r}}. \quad (17)$$

Similarly, the average number of bonds per object appears as

$$\rho' = \rho \int_{V_{\text{ex}}} g(r) d^3 \mathbf{r}. \quad (18)$$

Using these two definitions, an equation for  $\rho'_c$  can be obtained. Its predictions were successfully compared by [3] to numerical simulations for systems of hard-core spheres with or without interaction potentials.

Only monodisperse objects have been addressed so far in this subsection. For polydisperse populations, there seems to be some confusion in the literature. For flat objects, the statistical derivation of the excluded volume in Sect. “Calculation of the Excluded Volume for Plane Convex Fractures” quite naturally yielded the averages (9) or (11), which account for the sizes of the two intersecting objects. For isotropic populations of segments with length  $l$  in the plane or disks with radius  $R$  in space, for instance, the averages can be expressed as

$$V_{\text{ex}} = \frac{2}{\pi} \langle l \rangle^2 \text{ segments}, \quad d = 2 \quad (19a)$$



$$V_{\text{ex}} = \frac{\pi^2}{8} \langle R^2 \rangle \langle R \rangle \text{ disks, } d = 3. \quad (19b)$$

However, another type of average has been proposed by [6], namely,

$$W_{\text{ex}} = \frac{2}{\pi} \langle l^2 \rangle \text{ segments, } d = 2 \quad (20a)$$

$$W_{\text{ex}} = \frac{\pi^2}{8} \langle R^3 \rangle \text{ disks, } d = 3. \quad (20b)$$

On the basis of the numerical simulations of [29,30], [6] and others claim that the average bond number for polydisperse objects is not given by Eq. (6) but instead by

$$\rho'' = \rho W_{\text{ex}}. \quad (21)$$

However, a careful examination of [29]'s data shows that they correspond very accurately to Eq. (6) with (19a). Finally, the derivations of [8] are based on (19b) and yield consistent results.

Actually [29], and [30] showed that  $\rho'_c$  is not invariant for similar systems of segments in the plane with various degrees of polydispersity, while  $\rho''_c$  is. [11] also observed that  $\rho''_c$  is roughly constant for very broad power-law segment size distributions. The profound meaning of this observation is that continuum percolation is not determined only by the average coordination, when connections over various ranges may coexist. As suggested by [28], this is probably because contacts between objects too close to each other are redundant to percolation.

To summarize, the density  $\rho'$  based on  $V_{\text{ex}}$  resulting from the averages (9), (11), or (19a) is always equal to the mean number of intersections per object, but it cannot be used to relate the percolation thresholds of mono- and polydisperse systems. The alternative definition  $\rho''$  in (21) is very successful in this respect and is going to be generalized as  $\rho'_3$  in (29).

## Methods

The three main tools necessary for the numerical study of the percolation properties of the fracture network models are summarized in this section.

First, the medium is assumed to be spatially periodic on a large scale. A detailed description of spatially periodic media is given by [1], and only the main characteristics of these models are briefly repeated here. The geometrical and physical properties of the system under investigation are invariant under the translations

$$\mathbf{R}_i = i_1 \mathbf{l}_1 + i_2 \mathbf{l}_2 + i_3 \mathbf{l}_3 \quad (22)$$

where  $\mathbf{i} = (i_1, i_2, i_3) \in \mathbb{Z}^3$ , and where  $\mathbf{l}_1, \mathbf{l}_2$  and  $\mathbf{l}_3$  define a unit cell where the system is studied. The entire space is tiled by replicas of this unit cell, translated by  $\mathbf{R}_i$ . All the studies presented in this chapter are performed in cubic unit cells where  $|\mathbf{l}_1| = |\mathbf{l}_2| = |\mathbf{l}_3| = L$ .

Spatial periodicity implies that fractures may cross the imaginary unit cell boundaries, and reach the neighboring cells of the periodic medium. Therefore, for polydisperse fractures,  $R_M$  should be at least smaller than  $L/2$ . Moreover, in order to represent a homogeneous medium by a periodic model, one has to set the unit cell size much larger than any finite characteristic length scale in the system. Practically speaking, because of the finite size effects which will be discussed in Sect. "Methods",  $R_M$  is at least smaller than  $L/4$ .

Second, the networks are characterized by a graph which provides all the necessary relations and information. This graph, denoted by  $\Gamma_1$ , consists of vertices which correspond to the polygons, and edges which correspond to the intersection between polygons.  $\Gamma_1$  will be used to study network percolation as a function of fracture shape, distribution and density, as well as to characterize the topological features of the percolating components of the networks.

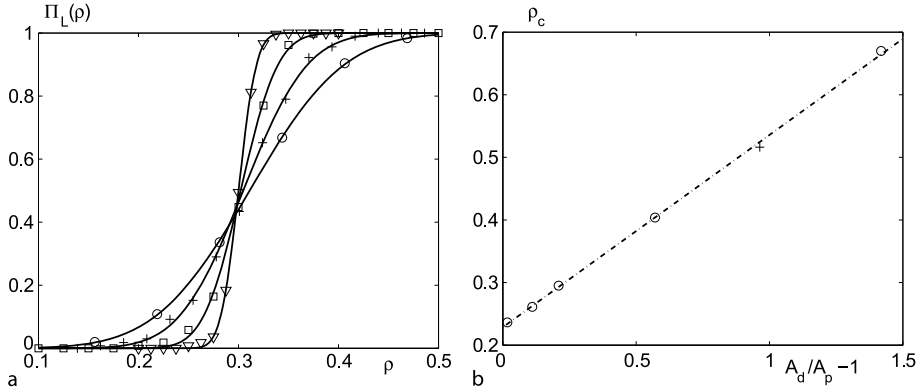
The information relative to the intersections is stored in the graph  $\Gamma_1$ . Since the networks considered here are spatially periodic, intersections of a polygon  $P_1$  with the periodic replicas of a polygon  $P_2$  in the 26 neighboring unit cells have to be checked as well. Once the intersections have been identified, the edges of  $\Gamma_1$  are known, and  $\Gamma_1$  can be set up.

Third, in order to estimate the percolation thresholds, the classical finite-size scaling method described in [37] is used. The percolating system is studied for various cell sizes  $L$ . For given values of  $L$  and  $\rho$ , the probability  $\Pi_L(\rho)$  of having a percolating cluster is derived from numerous realizations of the system. Then, the numerical data are used to estimate  $\rho_{Lc}$  (the value for which  $\Pi_L(\rho) = 1/2$ ) and an estimate of the width  $\Delta_L$  of the transition region of  $\Pi_L(\rho)$ .  $\Pi_L(\rho)$  was fitted with an error function of the form

$$\Pi_L(\rho) = \frac{1}{\sqrt{2\pi}\Delta_L} \int_{-\infty}^{\rho} \exp\left\{-\frac{(\xi - \rho_{Lc})^2}{2(\Delta_L)^2}\right\} d\xi \quad (23)$$

where  $\rho_{Lc}$  and  $\Delta_L$  are fit parameters. Once they have been evaluated for several values of  $L$ , the asymptotic value  $\rho_{Lc}$  for infinite systems  $\rho_c$  can be derived from the two scaling relations

$$\rho_{Lc} - \rho_c \propto L^{-1/\nu} \quad \Delta_L \propto L^{-1/\nu}. \quad (24)$$



Percolation, and Faults and Fractures in Rock, Figure 3  
**a** The probability of percolation  $\Pi_L(\rho)$  vs the density  $\rho$  of fractures in fracture networks created by equal sized, regular hexagons. Data are for sample sizes  $L/R = 4(\circ), 6(+), 10(\square), 20(\nabla)$ . The solid lines are the fitted error functions. **b** The percolation thresholds  $\rho_c$  for regular polygons ( $\circ$ ) and rectangles with  $a/b = 0.5(+)$  vs  $(A_d/A_p - 1)$ . The linear fit ( $-\cdot-\cdot-$ ) yields  $\rho_c = 0.231 \pm 0.002$  for disks

**Monodisperse Fractures**

This case was addressed by [18]. Since the computer time increases proportionally to the square of the number  $N$  of objects,  $L$  (measured in units of the disk radius  $R$ ) was kept below 16 in this early contribution. Despite the small cell sizes, the scaling laws (24) are well verified, which justifies the extrapolations of  $\rho_c$  at  $L \rightarrow \infty$ . The polygons were created, and intersections identified. Percolation was checked in all possible directions  $x, y$  and  $z$ . Periodic boundary conditions were applied to the 3d graph during this search; this means that a cluster must touch two opposite faces of the unit cell, and in addition contain fractures intersecting one another across the faces.

An example of the plots of the estimated  $\Pi_L(\rho)$  data points is given in Fig. 3a, together with the fitted error functions. Plots of  $\ln(\Delta_L)$  vs  $\ln(L/R)$  were used to obtain the critical exponent  $\nu$ . The various polygons are expected to belong to the same universality class, and  $\nu$  was expected to be the same in all cases. Values were in the range  $\nu = 1.011 \pm 0.044$ . The plots of  $\rho_{Lc}$  vs  $\Delta_L$  were extrapolated for  $\Delta_L \rightarrow 0$  to find  $\rho_c$  and these extrapolations are shown in Fig. 3b as functions of the shape factor  $A_d/A_p - 1$  where  $A_d$  is the area  $\pi R^2$  of the circumscribed disk.

These results can be analyzed in terms of the average number of intersections per fracture  $\rho'$ . (10) can be applied to networks made of identical polygons

$$\frac{V_{ex}}{R^3} = \pi^2 \left(\frac{N_v}{\pi}\right)^2 \cos\left(\frac{\pi}{N_v}\right) \sin^2\left(\frac{\pi}{N_v}\right),$$

(regular  $N_v$ -polygons) (25a)

$$\frac{V_{ex}}{R^3} = \frac{8a(a+1)}{(a^2+1)^{3/2}}, \quad (\text{rectangles with aspect ratio } a)$$

(25b)

The resulting values of  $\rho'_c$  are remarkably constant (cf. [18]). For all the fracture networks, including the cases of anisotropic (rectangular) polygons,  $\rho'_c$  is within the range

$$\rho'_c = 2.26 \pm 0.04.$$

(26)

Note that (26) concords with the limits (12) set up by [4] for 3d systems.

To summarize, this set of numerical results suggests that the percolation threshold of a network of identical Poissonian polygons has a universal value, expressed as Eq. (26).

**Polydisperse Fractures**

Since natural fracture networks are likely to have more complex size and shape distributions, the extension of (26) to these cases is of great interest. The key for this extension is the definition of a proper averaging procedure for the excluded volume.

The fracture size  $R$  is always supposed to follow the power law (1). Moreover, fractures of various shapes  $S$  are considered as well as mixtures of shapes. The three length scales  $R_m, R_M$  and  $L$  define two dimensionless ratios

$$R'_m = \frac{R_m}{R_M}, \quad L' = \frac{L}{R_M}.$$

(27)

Moreover, it will be shown below that global connectivity (percolation) is no longer controlled solely by the lo-

cal one (mean coordination), in the case of size polydispersity, and the definition of the percolation parameter has to be generalized. Since shape effects are well accounted for by  $\langle v_{ex} \rangle$ , it is useful to define the dimensionless shape factor  $\langle v_{ex} \rangle$ , for a set of fractures with identical shapes, but possibly different sizes

$$\langle v_{ex} \rangle = \frac{\langle V_{ex} \rangle}{\langle R \rangle \langle R^2 \rangle} \quad (28)$$

It can then be used to define two dimensionless densities, with different weightings of the fracture sizes

$$\begin{aligned} \rho'_{21} &= \rho \langle v_{ex} \rangle \langle R^2 \rangle \langle R \rangle = \rho \langle V_{ex} \rangle ; \\ \rho'_3 &= \rho \langle v_{ex} \rangle \langle R^3 \rangle . \end{aligned} \quad (29)$$

The subscripts are reminders of the statistical moments of  $R$  involved in each definition.  $\rho'_{21}$  is the generalization of  $\rho'$  for monodisperse networks, since it can be shown that it is still equal to the mean number of intersections per fracture [2]. Both  $\rho'_{21}$  and  $\rho'_3$  reduce of course to  $\rho'$  in case of equal-sized fractures.

The main tools required to study the percolation of polydisperse networks model are similar to the ones described in Sect. "Methods".

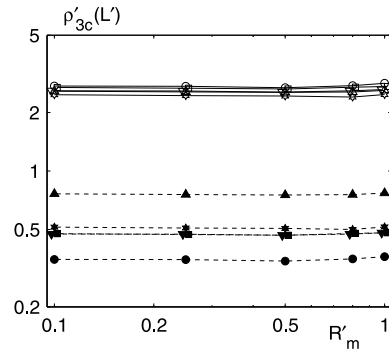
For given values of the parameters, the probability  $\Pi$  of having a percolating cluster which spans the cell along the  $x$ -direction is derived from  $N_r$  random realizations of the system; then, the value  $\rho'_c$  for which  $\Pi = 0.5$  is estimated.  $\Pi$  and  $\rho'_c$  depend on several parameters as summarized by the formulae

$$\Pi(R'_m, L', a, S, \rho'), \quad \rho'_c(R'_m, L', a, S) \quad (30)$$

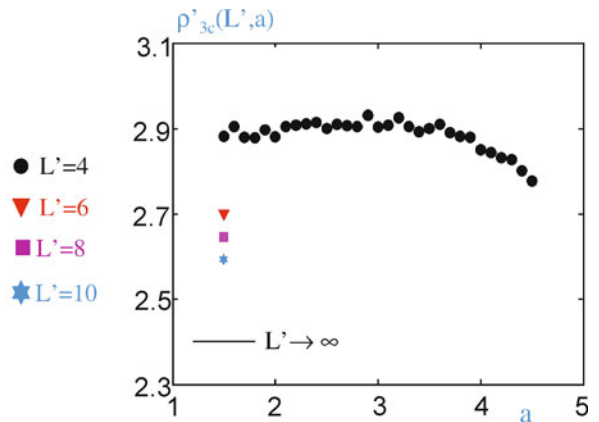
where  $\rho'$  denotes any one of the dimensionless densities defined in (29). For brevity, they will be often written as  $\Pi(L', \rho')$  and  $\rho'_c(L')$ .

In practice,  $\Pi(L', \rho')$  was evaluated from sets of 500 realizations, for about 10 values of the network density, evenly distributed in a range where  $\Pi$  varies from 0.05 to 0.95. Since there is a correspondance between  $\rho'_{21}$  and  $\rho'_3$ , for given values of  $S$ ,  $a$  and  $R_m$ , the same data sets can be used to determine  $\rho'_{21c}(L)$  and  $\rho'_{3c}(L)$ . The 95% confidence interval is estimated to be about  $\pm 0.04$  in terms of  $\rho'_{3c}(L)$ .

The influence on  $\rho'_c$  of the four parameters in Eq. (30) was systematically studied in [24]. We only state here the main result, which is that in the range  $1.5 \leq a \leq 4$ ,  $R_m \ll L$  and for (almost) any fracture shape,  $\rho'_c$  depends only on the domain size, and that in the limit of infinite domains, a unique value of  $\rho'_c(\infty)$  applies in all cases. The independence on the various parameters is illustrated in the following examples.



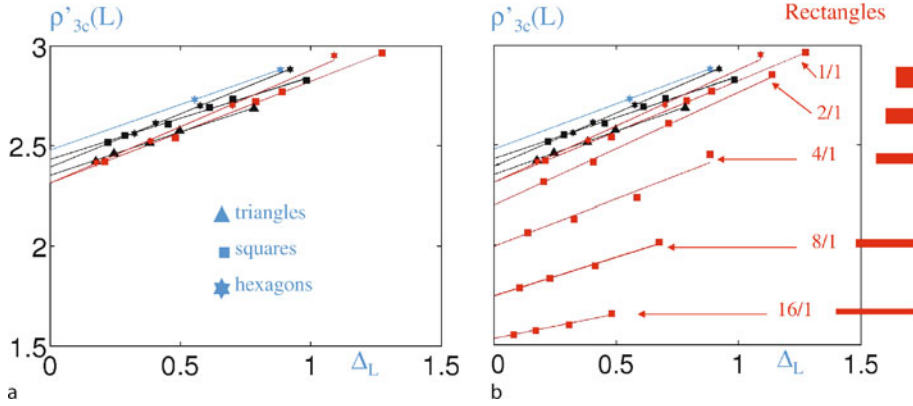
Percolation, and Faults and Fractures in Rock, Figure 4  
The percolation thresholds  $\rho'_{3c}$  (open symbols, solid lines) and  $\rho_c \langle R^3 \rangle$  (black symbols, broken lines) for networks with  $L' = 6$  and  $a = 1.5$  for regular hexagons (o), squares (□), triangles (△), mixture of hexagons and triangles, 50%–50% (▽), and mixture of hexagons and rectangles with aspect ratio 4, 50%–50% (☆)



Percolation, and Faults and Fractures in Rock, Figure 5  
The percolation threshold  $\rho'_{3c}(L', a)$  for networks of hexagonal fractures with  $R'_m = 0.1$ , versus the exponent  $a$ , for various domain sizes  $L$ . The lower line is the extrapolation of the data for  $a = 1.5$  when  $L'$  tends to infinity

In the example of Fig. 4,  $L'$  and  $a$  are kept constant, but the range of size and the fracture shapes varied. The networks contain hexagons, squares or triangles, or mixtures of hexagons with triangles or rectangles with a four to one aspect ratio. The upper set of curves shows that  $\rho'_c$  is indeed independent of  $R_m$  and  $S$ . Note that the rightmost points are actually monodisperse networks. For comparison, the thresholds  $\rho_c \langle R^3 \rangle$ , which do not include the shape factor  $\langle v_{ex} \rangle$  (see Eq. 29), are also shown in the same figure and they are clearly much more scattered. It is the incorporation of  $\langle v_{ex} \rangle$  in the definition of  $\rho'_3$  which unifies the results for the different shapes.

Conversely, the fracture shape (hexagonal) and the range of size ( $R'_m = 0.1$ ) are kept constant in the example



Percolation, and Faults and Fractures in Rock, Figure 6

The percolation threshold  $\rho'_{3c}(L')$  for mono- or polydisperse networks of fractures with various shapes, versus the width  $\Delta_L$  of the percolation transition. In a, the fractures are hexagons, squares or triangles.  $\rho'_{3c}(\infty)$  is the extrapolation for  $\Delta_L \rightarrow 0$ , which falls in the range of Eq. 31. Data for monodisperse networks of rectangles with aspect ratios from 1 to 16 are added in b

of Fig. 5, whereas the exponent  $a$  and the domain size  $L$  are varied. It is seen that  $\rho'_c$  does not vary when  $a$  ranges from 1.5 to 4. However, a definite dependence on the domain size is observed, which corresponds to the well known finite size effects.

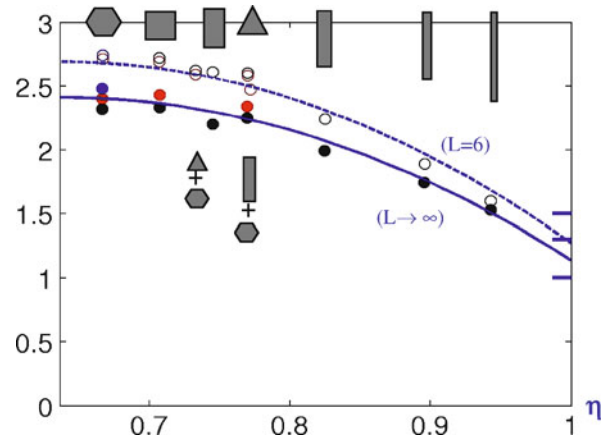
The data for increasing  $L'$  can be extrapolated for infinite systems by use of a classical technique. The combination of (23) and (24) shows that  $\rho'_c(L) - \rho'_c(\infty)$  is proportional to the width  $\Delta_L$  of the percolation transition zone. Hence,  $\rho'_c(\infty)$  can be read on the vertical axis of the plot of  $\rho'_c(L)$  versus  $\Delta_L$ , which is shown in Fig. 6. The data for many cases, including various fracture shapes in monodisperse and polydisperse networks, are gathered in Fig. 6a. In all cases, the extrapolated values  $\rho'_{3c}(\infty)$  fall in the narrow range

$$\rho'_{3c}(R'_m, a, S, L' \rightarrow \infty) = \rho'_{3c}(\infty) \approx 2.4 \pm 0.1 . \quad (31)$$

This applies for a variety of shapes, as well as for mixtures of fractures with different shapes (see Fig. 4).

However, when the polygons become elongated,  $\rho'_{3c}(L')$  varies with the aspect ratio. Data for rectangles with aspect ratios  $A_r$  up to 16 are shown in Fig. 6b. It appears that  $\rho'_{3c}(L')$  decreases significantly when  $A_r$  increases.

This can be taken into account by using the shape factor  $\eta = 4R/P$  of the fractures. This ratio is minimum for disks, with  $\eta = 2/\pi \approx 0.637$ , and it increases up to one when the shape deviates from circularity. It turns out that a quadratic correction in terms of  $\eta$  is very successful for the representation of the data for very different and irregular fracture shapes.



Percolation, and Faults and Fractures in Rock, Figure 7

The percolation thresholds  $\rho'_{3c}(L' = 6)$  and  $\rho'_{3c}(\infty)$  for a variety of fracture shapes and size distributions, in comparison with the expressions (32), (33). The marks on the right are the predictions of [15,16,31] for infinitely elongated objects. The fracture shapes are indicated by the icons above or below the data points

All the thresholds obtained in cells with  $L' = 6$  and mono- or polydisperse size distributions with  $a = 1.5$  or 2 and  $R_m = 0.1$  are plotted in Fig. 7 as functions of  $\eta$ . This includes networks of hexagons, squares, triangles, mixtures of hexagons with rectangles or triangles, and rectangles with  $h/w$  up to 16. The data are well fitted by the expression

$$\rho'_{3c}(L') = 2.69 \left[ 1 - 4 \left( \eta - \frac{2}{\pi} \right)^2 \right] \quad (L' = 6) . \quad (32)$$

The extrapolated data for infinite systems are also presented in Fig. 7, in comparison with the corrected version

of Eq. (31),

$$\rho'_{3c}(\infty) = 2.41 \left[ 1 - 4 \left( \eta - \frac{2}{\pi} \right)^2 \right]. \quad (33)$$

In both cases, the deviations never exceed  $\pm 0.1$ . The corrective term becomes significant, i. e., larger than the error bar in (31), when  $\eta > 3/4$ , which corresponds for rectangles to aspect ratios larger than 2.

It can be noted that Eq. (33) predicts a threshold value 1.14 when  $h/w$  tends to infinity (i. e., when  $\eta \rightarrow 1$ ), which is in the range of the predictions 1.5 for prolate ellipsoids [16], 1 for capped cylinders [15] and 1.3 for elongated prisms [31], in the limit of infinite slenderness.

### Determination of the Dimensionless Density from Experimental Data

Since percolation properties are controlled by the dimensionless density  $\rho'$ , it is theoretically and practically important to derive estimations of  $\rho'$  from field data. In most cases, these data are based on 1D and 2D measurements of fracture traces along boreholes or on exposed outcrops which necessitate extrapolation by stereological techniques to 3D. Such extrapolations have already been made for specific fracture shapes by Warburton [39,40], Piggott [26], Berkowitz and Adler [8] and Sisavath et al. [36] (see also the references therein).

Our general methodology which is detailed in [38] can be illustrated by the intersection of a family of convex fractures with a line of length  $L$  which is parallel to the unit vector  $\mathbf{p}$ . Consider a fracture of surface  $A$ , of normal  $\mathbf{n}$  and of in-plane orientation  $\omega$ ; this object does not intersect the line when its center is located out of a surface of area  $A$ . Since this is valid for any in-plane orientation, the excluded volume of the line and of the surface is equal to  $AL|\mathbf{p}\cdot\mathbf{n}|$ . Hence, the average number of intersections  $\langle n_I \rangle$  per unit length between such a line and an isotropic network of a monodisperse family of fractures is

$$\langle n_I \rangle = \frac{1}{2} A \rho. \quad (34)$$

Of course, the major interests of this formula are that it does not depend on the precise shape  $S$  of the fractures and that  $\rho$  can be deduced from  $n_I$  and  $A$ . However, it depends in a crucial way on the convexity of the fractures.

### Isotropic Networks

In order to derive the average number of intersections  $\Sigma_t$  of a family of convex fractures  $\mathcal{F}(R)$  with a plane  $\Pi$  per unit area of the plane, define in  $\Pi$  a large convex region

$\mathcal{R}$  of area  $\mathcal{A}$  and perimeter  $\mathcal{P}$ . The excluded volume of  $\mathcal{F}(R)$  and  $\mathcal{R}$  is thus given by (8). The number of intersections  $d\Sigma_t$  of the fractures of size ranging from  $R$  to  $R + dR$  is proportional to the volumetric density of such fractures multiplied by the excluded volume of  $\mathcal{F}(R)$  and  $\mathcal{R}$  as expressed by (8); when  $\mathcal{A} \rightarrow \infty$ ,  $\mathcal{A} \gg \mathcal{P}$ ; therefore,

$$d\Sigma_t(R) \rightarrow \frac{1}{4} \rho P(R) n(R) dR \quad \text{when } \mathcal{A} \rightarrow \infty. \quad (35)$$

This relation can be averaged over the sizes  $R$

$$\Sigma_t = \int d\Sigma_t(R) = \frac{1}{4} \rho \langle P \rangle. \quad (36)$$

The intersections of the fractures with a plane are called *traces* or *chords*. Let  $c$  be the length of a trace as illustrated in Fig. 1c. Such an intersection of length  $c(z, \mathbf{n}, \omega)$  exists if the vertical coordinate  $z$  of the center verifies

$$z_m(\mathbf{n}, \omega) \leq z \leq z_M(\mathbf{n}, \omega). \quad (37)$$

For a given fracture of size  $R$ , the average trace length  $\langle c \rangle_R$  when the intersection exists, can be expressed as

$$\langle c \rangle_R = \frac{\int d\omega \int d\mathbf{n} \int_{z_m}^{z_M} c dz}{\int d\omega \int d\mathbf{n} \int_{z_m}^{z_M} dz}. \quad (38)$$

Surprisingly, the numerator  $N_R$  of this fraction is easier to evaluate than its denominator  $D_R$ . The most internal integral  $\int_{z_m}^{z_M} c dz$  is equal to the area  $A$  of the fracture projected onto the plane perpendicular to  $\Pi$  which contains the trace, i. e.,  $A \sin \theta$ . Therefore,  $N_R$  is equal to  $\pi^3 A$ . The derivation of  $D_R$  is slightly more involved Santalo [34]; it is proportional to the integral of the Feret (or caliper) diameter over  $\omega$ . Frenet formulae are used to express this integral. Finally,

$$\langle c \rangle_R = \pi \frac{A(R)}{P(R)}. \quad (39)$$

For polydisperse fractures, the overall average  $\langle c \rangle$  is given by

$$\langle c \rangle = \frac{\int dR \Sigma_t(R) \langle c \rangle_R}{\int dR \Sigma_t(R)} = \pi \frac{\langle A \rangle}{\langle P \rangle} \quad (40)$$

a formula which is again an obvious generalization of the disk formula (cf. (24a) of Berkowitz and Adler [8]).

The density of trace intersections  $\Sigma_p$  is defined as the number per unit surface in the observation plane of the points which are intersections of traces. Since the fractures are randomly oriented and distributed in space, the same properties are valid for the traces. Moreover, as a trivial extension of the concept of excluded volume, the excluded

surface  $S_{\text{ex}}$  of two traces of random orientations and of lengths  $c_1$  and  $c_2$  is equal to (cf. [2])

$$S_{\text{ex}} = \frac{2}{\pi} c_1 c_2. \tag{41}$$

Let  $\sigma_t(R, c)dc dR$  be the surface density of traces of length  $c$  ranging from  $c$  to  $c + dc$ , for the fractures of size  $R$  ranging from  $R$  to  $R + dR$ . Hence, the surface density of intersections of traces  $c_1$  corresponding to fractures of size  $R_1$  and of traces  $c_2$  corresponding to fractures of size  $R_2$  is

$$\sigma = \frac{1}{2} \sigma_t(R_1, c_1) \sigma_t(R_2, c_2) \frac{2}{\pi} c_1 c_2. \tag{42}$$

As a direct consequence

$$\Sigma_p = \iiint \sigma \, dc_1 \, dR_1 \, dc_2 \, dR_2. \tag{43}$$

This last expression can be split into a product of integrals since the populations 1 and 2 are independent. According to (35), (36), (43),  $\Sigma_p$  can be expressed as

$$\Sigma_p = \frac{1}{\pi} \frac{\pi^2}{16} \rho^2 \langle A \rangle^2 = \frac{\pi}{16} \rho^2 \langle A \rangle^2. \tag{44}$$

**Extensions**

Let us now examine various possible extensions of the previous formulae.

The precise shape  $S$  of the fractures is never taken into account. Therefore, all the previous formulae are valid whatever the mixture of shapes  $S$ .

For anisotropic networks, the normal vector  $\mathbf{n}$  is not uniformly distributed over the unit sphere. Let  $\theta$  and  $\varphi$  be the two polar angles of  $\mathbf{n}$  (cf. Fig. 1c); the probability that the end of  $\mathbf{n}$  for fractures of sizes in the interval  $[R, R + dR]$  belongs to the interval  $[\theta, \theta + d\theta] \times [\varphi, \varphi + d\varphi]$  is given by  $\rho n(R, \mathbf{n}) d\theta d\varphi dR$ . The statistical average  $\langle \cdot \rangle$  can be calculated with this differential element.

The first quantity which can be easily generalized is  $\langle n_I \rangle$  (cf. (34))

$$\langle n_I \rangle = \rho \iiint n(R, \mathbf{n}) A \cos \theta \, d\theta \, d\varphi \, dR = \rho \langle A | \mathbf{p} \cdot \mathbf{n} | \rangle. \tag{45}$$

The other generalized formulae are summarized in Table 2.  $\alpha$  is the angle between the normal  $\mathbf{v}$  to the plane  $\Pi$  and  $\mathbf{n}$ ; in most cases, by choosing the  $z$ -axis perpendicular to  $\Pi$ ,  $\alpha$  is equal to  $\theta$ ;  $\beta_{12}$  is the angle between the normals  $\mathbf{n}_1$  and  $\mathbf{n}_2$  to the two fractures 1 and 2.

These formulae can be specialized to networks of subvertical fractures with a horizontal observation plane  $\Pi$ .

Percolation, and Faults and Fractures in Rock, Table 2  
The major relations for the various kinds of networks.  $\mathcal{B}_{12} = \langle A_1 A_2 | \sin \beta_{12} | \rangle$

	Isotropic 3D	Anisotropic 3D	Subvertical isotropic	Subvertical anisotropic
$\langle n_I \rangle$	$\frac{1}{2} \rho \langle A \rangle$	$\rho \langle A   \mathbf{p} \cdot \mathbf{n}   \rangle$	$\frac{2}{\pi} \rho \langle A \rangle$	$\rho \langle A   \mathbf{p} \cdot \mathbf{n}   \rangle$
$\Sigma_t$	$\frac{1}{4} \rho \langle P \rangle$	$\frac{\rho}{\pi} \langle   \sin \alpha   P \rangle$	$\frac{\rho}{\pi} \langle P \rangle$	$\frac{\rho}{\pi} \langle P \rangle$
$\langle c \rangle$	$\pi \frac{\langle A \rangle}{\langle P \rangle}$	$\pi \frac{\langle A   \sin \alpha   \rangle}{\langle P   \sin \alpha   \rangle}$	$\pi \frac{\langle A \rangle}{\langle P \rangle}$	$\pi \frac{\langle A \rangle}{\langle P \rangle}$
$\Sigma_p$	$\frac{\pi}{16} \rho^2 \langle A \rangle^2$	$\frac{1}{2} \rho^2 \mathcal{A}_{12}$	$\frac{1}{\pi} \rho^2 \langle A \rangle^2$	$\frac{\rho^2}{2} \mathcal{B}_{12}$

Then,  $\alpha$  is equal to  $\pi/2$  and  $\beta_{12}$  is equal to the angles between the two traces in  $\Pi$ . Such networks can be either isotropic (i. e., the directions of the traces in  $\Pi$  are isotropic), or anisotropic. The corresponding results are detailed in Table 2.

**Discussion**

**Discrete Families of Fractures** In many practical cases, the fractures are perpendicular to a finite set of normals  $\{\mathbf{n}_i; i = 1, \dots, m\}$  with probabilities  $\{n(R, \mathbf{n}_i); i = 1, \dots, m\}$ . The integrals over  $d\theta d\varphi$  are thus replaced by the following summation for a function  $f(R, \mathbf{n}_i)$

$$\rho \sum_{i=1}^m n(R, \mathbf{n}_i) f(R, \mathbf{n}_i). \tag{46}$$

**Practical Use of the Formulae** The major interest of the formulae summarized in Table 2 is to try to use them to derive the macroscopic quantities  $\rho$ ,  $\langle A \rangle$  and  $\langle P \rangle$ . It is easy (and frustrating) to realize that only two of these quantities can be obtained. For instance, (40) implies that  $\langle A \rangle = \pi^{-1} \langle P \rangle \langle c \rangle$ ; from (36),  $\langle P \rangle = 4\rho^{-1} \Sigma_t$ ; therefore  $\langle A \rangle = 4\pi^{-1} \rho^{-1} \Sigma_t \langle c \rangle$ . When these expressions are introduced into (34) or (44), one obtains that the three following ratios should be equal to one

$$\kappa_1 = \frac{\pi}{2} \frac{\langle n_I \rangle}{\Sigma_t \langle c \rangle}, \quad \kappa_2 = \frac{\pi \Sigma_p}{\Sigma_t^2 \langle c \rangle^2}, \quad \kappa_3 = \frac{\pi}{4} \frac{\langle n_I \rangle^2}{\Sigma_p}. \tag{47}$$

The third relation is derived by eliminating  $\Sigma_t \langle c \rangle$  between  $\kappa_1$  and  $\kappa_2$ . These relations provide consistency relations between the data, but not  $\rho$ .

In other words, only two of the three quantities  $\rho$ ,  $\langle A \rangle$  and  $\langle P \rangle$  can be simultaneously derived from the average measured data. Note also that  $\kappa_1$  is insensitive to the spatial organization, and that this is not true for  $\kappa_2$  and  $\kappa_3$  which depend on trace intersections.

One can go further if some geometrical information is available which could be  $\langle V_{\text{ex}} \rangle$ . Here, we shall use a shape

factor  $\eta$  which is defined as  $\langle A \rangle \langle P \rangle^{-2}$ . For 3D isotropic networks, this expression can be combined to (40) and to (36) to yield  $\langle A \rangle$ ,  $\langle P \rangle$  and  $\rho$

$$\langle P \rangle = \frac{\langle c \rangle}{\pi \eta}, \quad \langle A \rangle = \frac{\langle c \rangle^2}{\pi^2 \eta}, \quad \rho = 4\pi \eta \frac{\Sigma_t}{\langle c \rangle} \quad (48)$$

or for a set of fractures normal to  $\mathbf{n}_i$

$$\rho_i = \frac{\pi^2 \eta_i}{|\sin \alpha_i|} \frac{\Sigma_{t_i}}{\langle c \rangle_i}. \quad (49)$$

There are many equivalent ways to derive  $\rho$ . The choice of the adequate formula depends on the available data. Note that formulae which contains  $\Sigma_p$  cannot be applied to families of parallel fractures.

When  $\rho$  and therefore  $\langle V_{ex} \rangle$  (cf. (8)) are known by one way or another, one can derive the dimensionless density  $\rho' = \rho \langle V_{ex} \rangle$  for isotropic and anisotropic networks

$$\rho' = \frac{\rho}{\pi} \langle (A_1 P_2 + A_2 P_1) |\sin \beta_{12}| \rangle \quad (50)$$

$$\rho' = \frac{\rho}{2} \langle A \rangle \langle P \rangle \quad (3d); \quad \rho' = \frac{4\rho}{\pi^2} \langle A \rangle \langle P \rangle \quad (2d). \quad (51)$$

Then, if the fracture network is not too polydisperse, one can use a classical mean field argument and approximate its properties by the properties of a monodisperse network of density  $\rho'$ .

### Applications

Several applications have already been made of the previous methodology and they can be summarized as follows. [36] showed that when data relative to fractures are collected along a line (e.g. a road or a well), estimations can be given to the major geometrical properties of the corresponding fracture networks, such as the volumetric density of fractures and their percolation character. [38] used the two dimensional maps obtained by [25] for subvertical fractures. Among other results, some of the consistency relationships (47) are well verified by these data. As previously,  $\rho'$  is estimated.

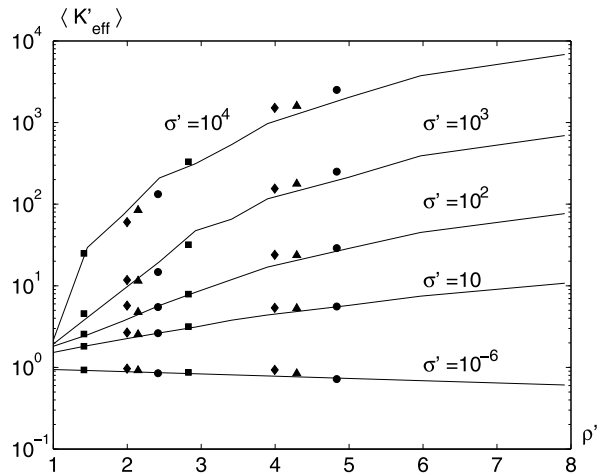
Finally [17], reconstructed a three-dimensional fracture network in a granite block from a series of experimental serial sections provided by [21]. It was visualized and its most important geometrical characteristics were studied. Though the network mostly consists of two families of fractures, it is interesting to note that a simple model of randomly oriented, monodisperse hexagons often yields a good order of magnitude for the various geometrical properties, which have been measured on the real block.

### Role of the Dimensionless Density in Other Geometrical Properties and Permeability

Though this article is focused on percolation properties, it is important to notice that the dimensionless densities which were introduced, play a crucial role in other properties as well. [18] studied two main other geometrical properties for monodisperse networks. Fracture networks partition the solid space into blocks; the block density is denoted by  $\rho_b$ . One can introduce the cyclomatic number of the graph  $\Gamma_1$  which is the number of independent cycles of this graph, and more precisely the number of cycles  $\bar{\beta}_1$  per unit volume. [18] showed that  $\rho_b$  and  $\bar{\beta}_1$  when made dimensionless by the excluded volume are independent of the fracture shapes.

Similar properties are found for the macroscopic permeability of fracture networks [20] and of fractured porous media whether they are monodisperse [10] or polydisperse [23]. In a series of contributions, the corresponding dimensionless quantities were shown to depend only on the dimensionless density  $\rho'$ . This is illustrated in Fig. 8. The porous medium has a local permeability  $K_m$  and the monodisperse fractures a conductivity  $\sigma$ . The macroscopic permeability of this medium is denoted by  $K_{eff}$ . Dimensionless quantities denoted by primes can be defined as

$$\sigma = R K_m \sigma', \quad K_{eff} = K_m K'_{eff}. \quad (52)$$



Percolation, and Faults and Fractures in Rock, Figure 8 Statistical averages of the permeability  $\langle K'_{eff} \rangle$  for samples containing  $N_{fr}=16$  or 32 fractures, with 4-, 6- or 20-gonal shapes, as functions of the network density  $\rho'$  and of the fracture conductivity  $\sigma'$ . The cell size is  $L = 4R$ . Data are for squares ( $\square$ ), rectangles with aspect ratios two to one ( $\Delta$ ) or four to one ( $\diamond$ ), hexagons (lines) and icosagons ( $\circ$ )

Figure 8 shows that the average macroscopic permeability ( $\langle K'_{\text{eff}} \rangle$ ) does not depend significantly on the fracture shape. The two major parameters are  $\rho'$  and  $\sigma'$ .

### Future Directions

The percolation properties of networks of random and convex plane fractures are successfully addressed by means of the excluded volume. Many important dimensionless properties of isotropic fracture networks only depend on the dimensionless density of fractures and not on the fracture shapes and sizes which represents a significant simplification.

These properties are not specific of fractures present in rocks and the same methodology can be applied for any other fracture system whatever the characteristic sizes and the nature of the material where it occurs.

These results should be generalized in several directions. In most cases, real fractures are not isotropically oriented and this feature should be incorporated in the next studies on this subject. The same is true for the homogeneous character of the network.

### Bibliography

#### Primary Literature

- Adler PM (1992) Porous Media: Geometry and Transports. Butterworth/Heinemann, Stoneham
- Adler PM, Thovert J-F (1999) Fractures and fracture networks. Kluwer Academic Publishers, Dordrecht
- Alon U, Balberg I, Drory A (1991) New, heuristic, percolation criterion for continuum systems. *Phys Rev Lett* 66:2879–2882
- Balberg I (1985) Universal percolation threshold limits in the continuum. *Phys Rev B* 31:4053–4055
- Balberg I (1987) Recent developments in continuum percolation. *Phil Mag* B56:991–1003
- Balberg I, Anderson CH, Alexander S, Wagner N (1984) Excluded volume and its relation to the onset of percolation. *Phys Rev B* 30:3933–3943
- Barenblatt GI, Zheltov IP, Kochina IN (1960) Basic concepts in the theory of seepage of homogeneous liquids in fissured rocks. *Soviet Appl Math Mech (PMM)* 24:852–864
- Berkowitz B, Adler PM (1998) Stereological analysis of fracture network structure in geological formations. *J Geophys Res* B103:15339–15360
- Berkowitz B, Ewing RP (1998) Percolation theory and network modeling applications in soil physics. *Survey Geophys* 19:23–72
- Bogdanov I, Mourzenko VV, Thovert J-F, Adler PM (2003) Effective permeability of fractured porous media in steady state flow. *Water Resour Res* 39. doi:10.1029/2001WR000756
- Bour O, Davy P (1997) Connectivity of random fault networks following a power law fault length distribution. *Water Resour Res* 33:1567–1583
- Charlaix E, Guyon E, Rivier N (1984) A criterion for percolation threshold in a random array of plates. *Solid State Commun* 50:999–1002
- Conrad F, Jacquin C (1973) Représentation d'un réseau bidimensionnel de fractures par un modèle probabiliste. Application au calcul des grandeurs géométriques des blocs matriciels. *Rev IFP* 28:843–890
- Drory A, Berkowitz B, Parisi G, Balberg I (1997) Theory of continuum percolation. III. Low-density expansion. *Phys Rev E* 56:1379–1395
- Florian R, Neda Z (2001) Improved percolation thresholds for rods in three-dimensional boxes. oai:arXiv.org:cond-mat/01110067
- Garboczi EJ, Snyder KA, Douglas JF, Thorpe MF (1995) Geometrical percolation threshold of overlapping ellipsoids. *Phys Rev E* 52:819–828
- Gonzalez Garcia R, Huseby O, Thovert J-F, Ledésert B, Adler PM (2000) Three-dimensional characterization of fractured granite and transport properties. *J Geophys Res* 105(B)21387–21401
- Huseby O, Thovert J-F, Adler PM (1997) Geometry and topology of fracture systems. *J Phys A* 30:1415–1444
- Isihara A (1950) Determination of molecular shape by osmotic measurement. *J Chem Phys* 18:1446–1449
- Koudina N, Gonzalez Garcia R, Thovert J-F, Adler PM (1998) Permeability of three-dimensional fracture networks. *Phys Rev E* 57:4466–4479
- Ledésert B, Dubois J, Velde B, Meunier A, Genter A, Badri A (1993) Geometrical and fractal analysis of a three-dimensional hydrothermal vein network in a fractured granite. *J Volcanol Geotherm Res* 56:267–280
- Long JCS, Remer JS, Wilson CR, Witherspoon PA (1982) Porous media equivalents for networks of discontinuous fractures. *Water Resour Res* 18:645–658
- Mourzenko V, Thovert J-F, Adler PM (2004) Macroscopic permeability of three dimensional fracture networks with power law size distribution. *Phys Rev E* 69:066307
- Mourzenko V, Thovert J-F, Adler PM (2004) Percolation of three-dimensional fracture networks with power-law size distribution. *Phys Rev E* 72:036103
- Odling NE (1997) Scaling and connectivity of joint systems in sandstones from western Norway. *J Struct Geol* 19:1257–1271
- Piggott AR (1997) Fractal relations for the diameter and trace length of disc-shaped fractures. *J Geophys Res* 102(B):18121–18125
- Pike GE, Seager CH (1974) Percolation and conductivity: A computer study. I. *Phys Rev B* 10:1421–1434
- Rivier N, Guyon E, Charlaix E (1985) A geometrical approach to percolation through random fractured rocks. *Geol Mag* 122:157–162
- Robinson PC (1983) Connectivity of fracture systems - A percolation theory approach. *J Phys A* 16:605–614
- Robinson PC (1984) Numerical calculations of critical densities for lines and planes. *J Phys A* 17:2823–2830
- Saar MO, Manga M (2002) Continuum percolation for randomly oriented soft-core prisms. *Phys Rev E* 65:056131
- Sahimi M (1995) Flow and transport in porous media and fractured rocks. VCH, Weinheim
- Sahimi M, Yortsos TL (1990) Applications of Fractal Geometry to Porous Media: A review. Society of Petroleum Engineers. Paper 20476
- Santalo LA (1943) Sobre la distribución probable de corpuscu-



- los en un cuerpo, deducida de la distribución en sus secciones y problema analogos. *Rev Unión Mat Argent* 9:145–164
35. Sher H, Zallen R (1970) Critical density in percolation processes. *J Chem Phys* 53:3759–3761
  36. Sisavath S, Mourzenko V, Genthon P, Thovert J-F, Adler PM (2004) Geometry, percolation and transport properties of fracture networks derived from line data. *Geophys J Int* 157:917–934
  37. Stauffer D, Aharony A (1994) *Introduction to Percolation Theory*, 2nd edn. Taylor and Francis, Bristol
  38. Thovert J-F, Adler PM (2005) Trace analysis for fracture networks of any convex shape. *Geophys Res Lett* 31:L22502
  39. Warburton PM (1980a) A stereological interpretation of joint trace data. *Int J Rock Mech Min Sci Geomech Abstr* 17:181–190
  40. Warburton PM (1980b) Stereological interpretation of joint trace data: Influence of joint shape and implication for geological surveys. *Int J Rock Mech Min Sci Geomech Abstr* 17:305–316

### Books and Reviews

- Bear J, Tsang C-F, de Marsily G (1993) *Flow and contaminant transport in fractured rock*. Academic Press, San Diego
- Myer LR, Tsang CF, Cook NGW, Goodman RE (1995) *Fractured and jointed rock masses*. Balkema, Rotterdam
- van Golf-Racht TD (1982) *Fundamentals of fractured reservoir engineering*. Developments in Petroleum Science, vol 12. Elsevier, Amsterdam

# Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation

MIE ICHIHARA<sup>1</sup>, TAKESHI NISHIMURA<sup>2</sup>

<sup>1</sup> Earthquake Research Institute, University of Tokyo, Tokyo, Japan

<sup>2</sup> Tohoku University, Sendai, Japan

## Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Elementary Processes in a Single-Bubble Dynamics](#)

[Bubbly Magma in an Elastic Rock as a Pressure Source](#)

[Acoustic Bubbles in Hydrothermal Systems](#)

[Future Directions](#)

[Acknowledgments](#)

[Bibliography](#)

## Glossary

**Impulse** The word *impulse* is used in many areas in different ways. In classical mechanics, the impulse is the integral of force with respect to time. It is also used to refer to a fast-acting force, which is often idealized by a step function or a delta function. In this text, it is used to represent any functional form of pressure increase, either static or transient, which can generate observable signals.

**Magma, melt, liquid** Magma is a general name for molten rock. It is fluid but contains solid and gas inclusions in liquid matrix. The matrix in magma is silicate melt (which is often called just melt), and that in a hydrothermal system is water.

**Volatile** Volatile is compound in silicate melt. The major component is H<sub>2</sub>O, of which concentration is 1–5 wt% depending mainly on pressure and composition of the melt. It exsolves from melt and forms gas bubbles at relatively low pressure (ca. 100 MPa corresponding to the litho-static pressure around several kilo-meters). The second major component is CO<sub>2</sub>. Although its concentration is usually several ppm, some kinds of melts may dissolve 3–30 wt% of CO<sub>2</sub> at several GPa.

**Long period seismic events** Long-period (or very long-period) seismic events are dominant in the period from about 1 s to more than a few tens of seconds. These signals at volcanoes are considered to be generated by interaction or resonance between volcanic fluid and the surrounding medium.

**Ground deformation** Ground deformation is often observed at volcanoes when magma chambers inflate or deflate. Such ground deformation is detected by geodetic measurements such as GPS, tilt or strain meters, and the deformations often continue for a few tens of minutes to days or even months.

**Magma chamber** A magma chamber is a storage system of molten magma. It is generally hard to detect, but is probably located at from a shallow depth (ca. 1 km) to a few tens of km beneath the volcanoes. The shape and size have not been confirmed yet, but it is usually assumed to be rather round and hundreds to thousands of meters in scale. A magma storage system which has a horizontal extent is called a sill, and one which has a vertical extent is called a dike.

**Rectified diffusion and rectified heat transfer** Rectified diffusion is a mechanism which can push dissolved volatiles into bubbles in a sound field. Bubbles take in more volatiles during expansion than they discharge during contraction, mainly because of the following two non-linear effects. Firstly, during expansion the bubble radius becomes larger so that the bubble surface is also larger than the surface during contraction. Secondly, radial bubble expansion tangentially stretches the diffusion layer and sharpens the radial gradient of the volatile concentration in the diffusion layer, so that the volatile flux into the bubble. The mechanism also works to push heat into bubbles and enhances evaporation in a liquid-vapor system. The rectified diffusion and heat transfer have been known and studied in mechanical and chemical engineering.

**Bubble collapse** When the bubble is compressed, oscillates, or loses its mass by diffusion or phase change, it contracts to a very small size and sometimes disappear. Bubble collapse indicates the contraction of a bubble and does not necessarily indicate its disappearance.

## Definition of the Subject

A volcano consists of solids, liquids, gases, and intermediate materials of any two of these phases. Mechanical and thermo-dynamical interactions of these phases are essential in generating a variety of volcanic activities. In particular, the gas phase is mechanically distinct from the other phases and plays important roles in the dynamic phenomena of volcanoes. When we work on volcanic activities, we are almost certainly confronted with physics problems associated with bubbles.

The roles of bubbles in volcanic activities have been investigated mainly in three aspects. Firstly, the nucleation, growth, and expansion of bubbles is considered to be the

Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Table 1

List of important variables and constants

Notation	Unit	Definition
$C_{eq}$	–	Equilibrium volatile concentration (weight fraction) in the liquid
$c_l$	$m\ s^{-1}$	Sound speed in the liquid
$c_{pg}$	$J\ kg^{-1}\ K^{-1}$	Heat capacity of the gas at constant pressure
$c_{pl}$	$J\ kg^{-1}\ K^{-1}$	Heat capacity of the liquid at constant pressure
$K_g$	Pa	Effective bulk modulus of the bubble
$K_l$	Pa	Bulk modulus of the liquid
$L$	$J\ kg^{-1}$	Latent heat
$p_g$	Pa	Pressure in the bubble
$p_g$	Pa	Pressure in the liquid far from the bubble
$R$	m	Bubble radius
$S$	m	Outer radius of the cellular bubble
$T$	K	Temperature
$U$	$m\ s^{-1}$	Translational velocity of the bubble
$\gamma$	–	Specific heat ratio
$\eta_l$	Pa s	Liquid viscosity
$\kappa_{gl}$	$m^2\ s^{-1}$	Diffusivity of the volatile in the liquid
$\kappa_T$	$m^2\ s^{-1}$	Thermal diffusivity in the bubble
$\kappa_{Tl}$	$m^2\ s^{-1}$	Thermal diffusivity in the liquid
$\tilde{\mu}$	Pa	Effective stiffness of the magma chamber
$\mu_l$	Pa	Shear elasticity of the liquid
$\rho_l$	$kg\ m^{-3}$	Liquid density
$\Sigma$	$m\ s^{-1}$	Surface tension
$\sigma_\infty$	Pa	Ambient stress change given to the magma chamber
$\tau$	s	Maxwell relaxation time
$\omega$	$rad\ s^{-1}$	Angular frequency

main force that brings the magma to the surface [62,88]. Secondly, a single bubble, if it is sufficiently large, may generate seismic waves when it rapidly expands or accelerates in the volcanic conduit [11,35,81,95], and may generate acoustic waves in the air when it oscillates or bursts at the magma surface [36,82,95,96]. Thirdly, the existence of bubbles can significantly reduce the sound velocity [16,42] and increase the attenuation and dispersion of the waves [15,30,44]. This effect is considered to be relevant to many spectral features of seismic waves and air-waves associated with volcanic activities [3,10,22,45].

Studies on bubble dynamics relevant to the volcanology are spread over many research fields and cannot be covered by a single paper. Good review papers and textbooks have already been published on bubble phenomena in sound fields [61,70,73,74] and on the nucleation and

growth of bubbles in magma [62]. In this paper, we discuss several bubble dynamics phenomena selected from a particular point of view that the bubbly fluid works as an impulse generator. Here the term impulse means a pressure increase, either static or transient, which can generate any observable signal (e. g. earthquakes, ground deformations, airwaves, and an eruption itself). Especially, we focus on the processes that the impulse is excited by nonlinear coupling between the internal processes of a bubbly fluid and an external perturbation. The importance of these processes have recently become noticed as a possible triggering mechanism of eruptions, earthquakes, and inflation of a volcano [57,64]. Although it is generally considered that stress perturbation caused by preceding events is important, exact mechanisms to generate a pressure increase, which is required to trigger the subsequent events, are yet under discussion. In the first place, factors controlling single bubble dynamics are summarized as the elementary processes in the bubbly fluid. Then two distinct liquid-bubble systems are considered, both of which are included in a volcano. The one is a body of bubbly magma confined in an elastic chamber, where elasticity of the chamber, melt viscosity, and gas diffusion are important. The other is a hydrothermal system, where bubble oscillation, evaporation, and heat transfer are important.

## Introduction

### Elementary Processes in Single-Bubble Dynamics

Radial motion of a single bubble interacting with ambient pressure perturbation is the elementary process controlling behaviors of the liquid-bubble mixtures. Although it appears quite simple, it contains various mechanisms in plenty. The great variety of behaviors of a single bubble has attracted many scientists, among whom is Leonardo da Vinci [76]. Nowadays, knowledge of the single bubble dynamics is used and studies are continued in many academic and industrial areas such as mechanical engineering, chemical engineering, medical science, and earth science.

Factors which may control the radial motion of a bubble are the pressure difference inside and outside the bubble, inertia and stress associated with the deformation of the surrounding liquid, propagation of pressure waves, heat and mass transport, phase transition at the bubble wall, chemical reactions, relative translational motion between the bubble and the liquid, and so on. Because including all these mechanisms at the same time in order to calculate the behavior of a bubble is unrealistic, we need to make adequate simplification and assumptions. Each mechanism has its own characteristic time scale in

## Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Table 2

### List of characteristic times

Notation	Equation	Definition
$\tau_c$	(2)	Inertia-controlled bubble collapse
$\tau_g$	(25)	Mass diffusion in the liquid around the bubble
$\tau_T$	(24)	Thermal diffusion in the bubble
$\tau_v <$	(11)	Viscosity-controlled bubble expansion
$\omega_o$	(9)	Natural frequency of a bubble

which the effect is dominant (Table 2), and its own effect on the bubble dynamics. Knowing the individual time scales and features is important when we want to understand and simulate a certain phenomenon correctly and efficiently. A brief review of some representative mechanisms with linearized analyses are presented in Sect. “**Elementary Processes in a Single-Bubble Dynamics**” for this purpose. Based on the results, geophysical phenomena and proposed models are discussed in the latter sections.

### Bubbly Magma in an Elastic Rock as a Pressure Source

We consider a body of bubbly magma confined in an elastic rock. Pressure perturbations to the system are caused by a change of tectonic stress due to local earthquakes, surface unloading by dome collapse, passing seismic waves from a near or distant source, or depressurization of the chamber by degassing or magma leakage. Dynamic response of the system may be relevant to subsequent activities of the volcano as follows.

Nishimura [64] investigated pressure re-equilibration between the bubbles, the melt, and the surrounding elastic medium. It is assumed that the pressure of the system is suddenly decreased. After re-equilibration, the original magma pressure is partially or completely recovered or even exceeded, depending on the size of the bubbles, stiffness of the elastic container, and the confining pressure. His model is used to explain rapid pressurization of a magma chamber triggered by the lava-dome collapse at Soufriere Hills Volcano [97], and pressure recovery in magma filling the chamber after explosive degassing to continue activities at Popocatepetl Volcano [12]. Shimomura et al. [87] extended the formulation of [64] to calculate the time profile of the pressure recovery after sudden decompression. They showed that the time scale of the pressure recovery is strongly controlled by the system parameters, which include stiffness of the elastic container, bubble number density, diffusivity of the volatile in the melt, ambient pressure, and properties of the melt.

Chouet et al. [13] also calculated the time profile assuming the system parameters for Popocatepetl Volcano, and compared the results with a particular source time function of a very-long-period seismic signal. Furthermore, in the same year, Lensky et al. [53] independently developed a mathematically equivalent model considering magma with CO<sub>2</sub> bubbles in mantle rock. They interpreted the results as a possible pressurization mechanism to initiate dikes in mantle which allow the fast transport of magma. There are in fact quite a few documented cases in which eruptions were triggered by local tectonic earthquakes (e.g. [47,68]) and wave propagation from a distant earthquake (e.g. [6,54]). Recently, Manga and Brodsky [57] have given a comprehensive review on the phenomena and possible mechanisms. Brodsky et al. [6] investigated the possibility that a strain wave from a distant earthquake can increase the pressure in bubbly magma by rectified diffusion, which is the mechanism by which volatiles are pumped into a bubble by cyclic expansion and contraction. However, it has turned out that the mechanism by itself can cause a negligibly small pressure increase [6,29]. Several other mechanisms for long-range triggering have been proposed, which include pressure increase from rising bubbles [55], sub-critical crack growth [7], and fracture unclipping [8].

### Acoustic Bubbles in Hydrothermal Systems

A hydrothermal system is another major source of pressure increase, long-period volcano seismic events [46], and triggered seismicity [57,89]. Behaviors of a single bubble and liquid-bubble mixtures in a hydrothermal system are quite different from those in a magmatic system, mainly because of the water viscosity which is less than that of magma by several orders of magnitude. We introduce several phenomena which are particular to the hydrothermal systems in Sect. “**Acoustic Bubbles in Hydrothermal Systems**”.

Geysers activity is well known for intermittent activity of hot-water effusion. The effusion process looks quite similar to volcanic eruptions, and some geysers are characterized by regular intervals of time and duration, which are also recognized in particular types of eruptions and seismic activities. Consequently, the geysers have been widely studied using seismological and geophysical techniques, as well as field observations, not only for clarifying the mechanism of the geysers but also for understanding the volcanic activities (e.g., [39,40,43,65]). Kedar et al. [39,40] conducted a unique experiment at Old Faithful Geyser, Yellowstone. They measured pressure within the geyser’s water column simultaneously with seismic measurements

on the surface. The data demonstrated that the tremor observed at Old Faithful results from impulsive events in the geyser. The impulsive events were modeled by a collapse of a spherical bubble by cooling that occurred when the water column reached a critical temperature. Their data are reviewed in Sect. “Acoustic Bubbles in Hydrothermal Systems” in relation to other studies on the dynamics of gas and vapor bubbles.

## Elementary Processes in a Single-Bubble Dynamics

### Equation of Motion for the Bubble Radius

Motion of a bubble is in fact a fluid dynamical problem for the liquid surrounding the bubble. The simplest model describing the behavior of a bubble is based on three assumptions:

- (1) The bubble is spherical,
- (2) The liquid is incompressible, and
- (3) The motion is radial.

Using the basic equations of fluid mechanics, which are the continuity equation and the momentum equation, and the force balance at the bubble surface, the first equation of motion for the bubble radius was obtained by Rayleigh [80]:

$$\rho_l \left( R\ddot{R} + \frac{3}{2}\dot{R}^2 \right) = p(R) - p_1, \quad (1)$$

where  $R$  is the bubble radius,  $\rho_l$  is the liquid density,  $p(R)$  is the pressure in the liquid at the bubble surface, and  $p_1$  is the pressure in the liquid at a large distance from the bubble. Using Eq. (1), Rayleigh [80] solved the problem of the collapse of an empty cavity in a large body of liquid at a constant  $p_1$  and showed the characteristic collapse time is

$$\tau_c = R_o \sqrt{\rho_l / p_1}. \quad (2)$$

The time  $\tau_c$  is called the Rayleigh collapse time and is one of the most important time scales in the bubble dynamics.

Plesset [69] extended Eq. (1) including the effects of surface tension and time-dependent pressure field, and Proitsky [71] included the effect of viscosity. The generalized Rayleigh equation for bubble dynamics is known as the Rayleigh–Plesset equation. liquid viscosity and surface tension. The generalized Rayleigh equation for bubble dynamics is called the Rayleigh–Plesset equation [70]:

$$\rho_l \left( R\ddot{R} + \frac{3}{2}\dot{R}^2 \right) = p_g - p_1 - 4\eta_l \frac{\dot{R}}{R} - \frac{2\Sigma}{R}, \quad (3)$$

where  $\eta_l$  is the liquid viscosity, and  $\Sigma$  is the surface tension. Equation (3) is valid for a Newtonian liquid under

conditions of negligible mass exchange at the bubble surface. A further generalized equation to which these two restrictions do not apply is [72]:

$$\begin{aligned} \rho_l \left( R\dot{u}_l + \frac{3}{2}u_l^2 \right) - J \left[ 2u_l + J \left( \frac{1}{\rho_g} - \frac{1}{\rho_l} \right) \right] \\ = p_g - p_1 + \int_R^\infty \frac{3\tau_{rr}}{r} dr - \frac{2\Sigma}{R}, \quad (4) \end{aligned}$$

where  $u_l$  is the radial liquid velocity at the bubble surface,  $J = \rho_l(u_l - \dot{R})$  is the outgoing mass flux through the bubble wall,  $\rho_g$  is the density of the gas in the bubble, and  $\tau_{rr}$  is the normal radial stress. When the interfacial mass flux  $J$  vanishes,  $u_l = \dot{R}$  as in the left-hand side of the original Eq. (3).

While the above equations consider a single bubble in an infinite melt, magmatic systems often contain bubbles with some finite spacing. Cellular models of packing which include a finite volume of melt in interaction with each bubble have been employed for closely spaced bubbles [79]. When the elementary cell is represented by a sphere with an outer radius of  $S$ , the equation corresponding to (3) is [76]:

$$\begin{aligned} \rho_l \left[ R\ddot{R} \left( 1 - \frac{R}{S} \right) + \frac{3}{2}\dot{R}^2 \left( 1 - \frac{4R}{3S} + \frac{1}{3}\frac{R^4}{S^4} \right) \right] \\ = p_g - p_1 - 4\eta_l \frac{\dot{R}}{R} \left( 1 - \frac{R^3}{S^3} \right) - \frac{2\Sigma}{R}. \quad (5) \end{aligned}$$

Equation (5) agrees with Eq. (3) for  $S \rightarrow \infty$ .

When there is no transport of heat or mass between the liquid and the bubble, the pressure in the bubble is determined by the instantaneous bubble radius alone. Using the ideal gas approximation, we have

$$p_g R^{3\gamma} = p_{g_o} R_o^{3\gamma}, \quad (6)$$

where  $\gamma$  is the specific heat ratio, and the subscript  $o$  indicates the equilibrium value of the variable. Substituting Eq. (6) into Eq. (3) for  $p_g$  and linearizing the equation, we obtain a damped oscillator equation:

$$\ddot{X} + 2b_v \dot{X} + \omega_o^2 X = -\frac{p'_1}{\rho_l R_o^2}, \quad (7)$$

$$b_v = \frac{2\eta_l}{\rho_l R_o^2}, \quad (8)$$

$$\omega_o = \frac{1}{R_o} \sqrt{\frac{3\gamma p_{g_o} - 2\Sigma/R_o}{\rho_l}}, \quad (9)$$

where  $X$  and  $p'_1$  are defined by  $R = R_o(1 + X)$  and  $p_1 = p_{g_o} - 2\Sigma/R_o + p'_1$ , respectively. Equation (7) is useful to see the characteristic behaviors of a bubble and their time scales. The resonant frequency of the bubble is

$\omega_o$  ( $\text{rad s}^{-1}$ ). When the second term in the left-hand side of Eq. (7) dominates the first one in the time scale of the resonant oscillation, namely when  $\omega_o < b_v$ , the resonant oscillation is damped. In the case of a gas bubble with a radius of  $10^{-3}$  m in magma ( $\rho_l = 2500 \text{ kg m}^{-3}$ ) at 10 MPa ( $p_{go} - 2\Sigma/R_o = 10^7 \text{ Pa}$ ), the frequency ( $\omega_o/(2\pi)$ ) is about 20 kHz. The oscillation is damped when the viscosity is larger than 160 Pa s. This viscosity is relatively small for magma. According to these estimations, we see that the free oscillation of a bubble in magma is possible in the limited cases that the viscosity is small and the bubble is large. We also see that the bubble oscillation is easily excited in water which has a viscosity about  $10^{-3}$  Pa s.

### Liquid Rheology

The shear rheology of the liquid surrounding the bubble is one of the controlling factors for the bubble dynamics. According to experimental results, magma has viscoelastic nature, which is the most simply represented by a linear Maxwell model [99]. Then the normal radial stress  $\tau_{rr}$  in Eq. (4) is related to the corresponding strain rate  $\dot{e}_{rr}$  by

$$\tau_{rr} = \mu_1 \int_0^t \exp\left(-\frac{t-t'}{\tau}\right) \dot{e}_{rr} dt', \quad (10)$$

where  $\mu_1$  is the shear elasticity and  $\tau$  is the relaxation time. In the limit of  $t \ll \tau$ , the Maxwell relation (10) is reduced to a linear elastic stress-strain relation as  $\tau_{rr} = \mu_1 e_{rr}$ . While in the limit of  $t \gg \tau$ , it is reduced to a Newtonian viscous relation, that is a linear stress-strain rate relation as  $\tau_{rr} = \mu_1 \tau \dot{e}_{rr}$ , where  $\mu_1 \tau$  corresponds to the Newtonian viscosity  $\eta_1$ .

Fogler and Goddard [21] first used the viscoelastic relation (10) in the generalized Rayleigh–Plesset Eq. (4) without mass flux, and demonstrated that the influence of the viscoelastic effects on the radial motion of a bubble is characterized by a dimensionless parameter called the Deborah number  $De = \tau/\tau_c$ , which compares the relaxation time  $\tau$  and Rayleigh collapse time  $\tau_c$  defined in Eq. (2): the influence is large when  $\tau \gg \tau_c$ . Extending the formulation by Fogler and Goddard [21] to a cellular bubble, Ichihara [28] investigated its characteristic behaviors in magmatic conditions. It is shown that the elastic oscillation of a bubble, which occurs in the case of  $\tau \gg \tau_c$ , is in a frequency of order of MHz and with very small displacement of the bubble wall because of the large shear modulus of the magma.

Change of the bubble radius in magma is mainly controlled by the viscosity except in magma with very small viscosity in which the bubble oscillation is possible. Therefore, in most of the studies for bubble growth in magma,

effects of liquid inertia and viscoelasticity are not considered, and Eq. (3) or (5) for a Newtonian fluid is used, neglecting the left-hand side terms representing the inertia [2,62,79,88]. Barclay et al. [2] analytically solved the problem of the viscosity-controlled bubble expansion for instantaneous decompression, and showed the characteristic expansion time is

$$\tau_v = \frac{4\eta_1}{3p_l}, \quad (11)$$

where  $p_l$  is the pressure in the liquid. The time  $\tau_v$  is one of the most important time scales of the bubble dynamics in magma [2,30], while the Rayleigh collapse time  $\tau_c$ , which is controlled by the inertia, is important in low-viscosity fluids including hydrothermal systems.

Definition of the viscous expansion time corresponding to Eq. (11) is different depending on which problems and literature are being referenced. The time scale of the entire expansion of a bubble for instantaneous decompression is represented by Eq. (11) using the reduced pressure for  $p_l$  [2]. Volumetric oscillation of a bubble in an acoustic field is prevented by the viscous resistance if the period is shorter than  $\tau_v$ , in this case with the initial static pressure for  $p_l$  [30]. When the bubble expansion is driven by a constant gas pressure, which occurs at the initial stage of diffusion-drive gas expansion when the gas is efficiently supplied from the liquid, the bubble grows approximately as  $R \sim R_o \exp[t\Delta p/(4\eta_1)]$  [62,93], where  $\Delta p$  is the pressure difference. In this case,  $\tau_v = 4\eta_1/\Delta p$ . The last case is discussed again later in the section of mass transport.

### Liquid Compressibility

The effect of liquid compressibility on radial motion of a bubble was first considered in connection with underwater explosions [14,41]. Liquid compressibility allows energy transport as a pressure wave so that it causes radiation damping. Noting that the effect is considerable in the case of a violent oscillation or collapse of a bubble, several mathematical approaches were proposed to include the effect in the equation of bubble radius. According to mathematical and numerical studies by Prosperetti and Lezzi [78], which compared the proposed equations, the following Keller's equation [41] is widely accepted as the most adequate form.

$$\begin{aligned} \rho_l \left[ \left(1 - \frac{\dot{R}}{c_l}\right) R \ddot{R} + \frac{3}{2} \left(1 - \frac{\dot{R}}{3c_l}\right) \dot{R}^2 \right] \\ = \left(1 + \frac{\dot{R}}{c_l} + \frac{R}{c_l} \frac{d}{dt}\right) \left(p_g - p_l - 4\eta_1 \frac{\dot{R}}{R} - \frac{2\Sigma}{R}\right), \end{aligned} \quad (12)$$

where  $c_l$  is the sound speed in the liquid. Although Prosperetti and Lezzi [78] further proposed to use the liquid enthalpy at the bubble wall instead of the pressure for the best accuracy, Eq. (12) is generally used in the literature.

Comparing Eqs. (12) and (3), we can see that the correction terms due to the liquid compressibility have the order of  $\dot{R}/c_l$ . It means that the correction is considerable only when the bubble wall velocity becomes as large as the sound speed of the liquid. By applying  $[1 + (\dot{R}/c_l) + (R/c_l)d/dt]^{-1}$  to both sides of Eq. (12) and linearizing the equation, Prosperetti [75] derived the acoustic damping coefficient, which corresponds to  $b_v$  in Eq. (8) for the viscous damping, as  $b_{ac}$ :

$$b_{ac} = \frac{\omega^2 R_o}{2c_l}, \quad (13)$$

where  $\omega$  is the angular frequency of the oscillation. The acoustic damping is more significant when the bubble is larger and the oscillation frequency is higher.

Effects of liquid compressibility on bubble dynamics in a highly viscous liquid are not understood comprehensively. Derivation of Eq. (12) and related studies were performed thinking of liquids with ordinary viscosities like water. Therefore, the Reynolds number  $Re = \rho_l R_o c_l / \eta_l$  was presupposed to be large. On the other hand, the viscosity of magma can be large enough to make  $Re$  very small. In this case, the same mathematical approximation is not necessarily applicable. Yamada et al. [100] pointed out this problem and solved the equations including the viscous force associated with the volumetric strain rate. Although the equation of motion for the bubble radius appears not to be affected by the compressibility when the system is initially hydrostatic, the velocity field around the bubble is different from the incompressible solution, even if the wall velocity is much smaller than the acoustic velocity.

There is argument whether the equation of radial motion of a bubble surrounded by a finite volume of liquid needs correction terms for the compressibility. However, it seems to be negligible in magma, which is evaluated as follows [28]. In the case that the bubble is surrounded by an elastic shell, contribution of the compressibility to the bubble expansion is  $\delta_c R = R_o (p_g - p_l) (1 - R_o^3/S_o^3)^{-1} (3K_l)^{-1}$ , where  $K_l$  is the bulk modulus of the liquid, which is the reciprocal of the compressibility [48]. We can evaluate  $\delta_c R/R_o < 10^{-3}$ , because  $K_l \sim 10^{10} - 10^{11}$  Pa for magma [99], the realistic pressure difference,  $p_g - p_l$  is not much larger than  $10^7$  Pa, and the volume fraction of the bubbles,  $R_o^3/S_o^3$ , is reasonably assumed to be smaller than the close-packing limit ( $\sim 0.74$ ). The change of  $p_g$  due to  $\delta_c R$  is  $\delta_c p_g/p_{go} \sim -3\delta_c R/R_o$ , which is also in the same or-

der. When the shell deforms viscously, displacement due to non-volumetric deformation grows, while that from the volumetric deformation remains in the same order.

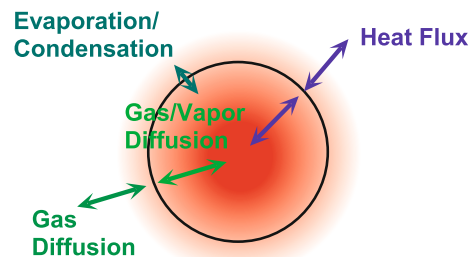
### Heat and Mass Transport

Each equation of motion for the bubble radius includes  $p_g$ , which is the pressure of the gas in the bubble, as we see in Eqs. (3), (5), and (12). Equation (6) is available to calculate  $p_g$  only for an adiabatic process. In fact, when a bubble expands, the pressure and temperature in the bubble decreases. Then heat and volatile components flow into the bubble from the surrounding liquid. The opposite occurs when a bubble shrinks. The internal processes which control  $p_g$  are schematically shown in Fig. 1. These transport effects are essential in most of actual systems including magmatic and hydrothermal systems.

Growth of a bubble by the mass diffusion in an over-saturated liquid is one of the fundamental problems. Based on purely dimensional considerations, an approximate growth law is given by

$$R\dot{R} = \frac{\kappa_{gl}\rho_l(C_o - C_{eq})}{\rho_g}, \quad (14)$$

where  $\kappa_{gl}$  is the diffusivity of the volatile in the liquid,  $C_o$  is the dissolved volatile concentration at a large distance from the bubble, and  $C_{eq}$  is the equilibrium concentration at the given pressure [62,70]. From Eq. (14) we find that, asymptotically,  $R \sim \sqrt{2\kappa_{gl}\rho_l(C_o - C_{eq})t/\rho_g}$ . This expression is not valid at  $t \rightarrow 0$  making  $\dot{R} \rightarrow \infty$ . In the initial stage, the diffusion is very efficient and the bubble growth is controlled by viscous resistance [62,93]. In this stage,  $R \sim R_o \exp[t\Delta p/(4\eta_l)]$  as discussed in the section of liquid rheology. The approximate time of the transition from the viscosity-controlled exponential solution to the diffusion-controlled square-root solution is found by



Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 1

The internal heat and mass transport processes which control pressure change and energy loss associated with the bubble oscillation

Navon and Lyakhovskiy [62] to be

$$\tau_{\text{vd}} \sim [-15 - 10 \log(Pe)] \eta_l / \Delta p, \quad (15)$$

where  $Pe = \Delta p R_o^2 \eta_l^{-1} \kappa_{\text{gl}}^{-1}$  is the Pecret number that compares the time scales of viscous expansion and diffusion. It is noted that Eq. (15) is validated for  $Pe < 10^{-2}$ , that is for relatively large viscosity and small bubble radius [62]. If we consider  $\Delta p \sim 10^6$  Pa and  $\kappa_{\text{gl}} \sim 10^{-11} \text{ m}^2 \text{ s}^{-1}$ , this condition is satisfied when  $R_o^2 \eta_l^{-1} < 10^{-19}$ , that is  $\eta_l > 10^7$  Pa s when  $R_o \sim 10^{-6}$  m, and  $\eta_l > 10^9$  Pa s when  $R_o \sim 10^{-5}$  m. Then the corresponding times are  $\tau_{\text{vd}} > 50$  s and  $\tau_{\text{vd}} > 5000$  s, respectively. Lensky et al. [51] have suggested that the change of the characteristic behavior of the bubble expansion over the time scale  $\tau_{\text{vd}}$  generates a non-linear response of the liquid-bubble mixture to the pressure perturbation, which may cause amplification of a pressure wave. Coupling of effects of viscosity and diffusion on the bubble expansion also occurs through the material properties, because magma viscosity and water diffusivity are strongly influenced by the amount of dissolved water, which is the major volatile component in magma [4,50].

Matsumoto and Takemura [59] numerically solved a complete set of equations for the radial dynamics of a bubble including the conservation equation for mass, momentum, and energy in the bubble, heat and mass diffusion in the liquid, and heat and mass exchange between the gas and the liquid by diffusion and evaporation/condensation. Except in extremely rapid phenomena as the cases they treated, approximation of a spatially uniform pressure in the bubble is adequate [75]. With this approximation, the computational load is considerably reduced [9,38,63].

It is necessary to consider non-uniform temperature distribution and compositions, in order to quantify the amounts of energy exchange between the bubble and liquid and energy loss associated with the non-equilibrium process. Time scales required to recover uniform temperature and composition in the bubble are controlled by diffusion processes and are much longer than that to attain uniform pressure, which is controlled by the pressure wave propagation in the bubble. Assuming representative values of the thermal diffusivity,  $\kappa_T \sim 10^{-5} \text{ (m}^2 \text{ s}^{-1}\text{)}$ , and the inter-diffusivity of the components in the gas phase,  $\kappa_{\text{gi}} \sim 10^{-7} \text{ (m}^2 \text{ s}^{-1}\text{)}$  [38], development of thermal and material diffusion layers all over the bubble with radius of  $10^{-3}$  m takes  $\sim 0.1$  s and  $\sim 10$  s, respectively. The time range of 0.1–10 s is exactly what studies on seismoacoustic phenomena in volcanology are mainly concerned with. It takes an even longer amount of time to recover uniform concentration of volatile components in the liquid around

the bubble. Therefore approximation of uniform temperature and compositions are not always adequate. Again we introduce results from the linearized theory for a periodic acoustic field. The bulk modulus of a bubble,  $K_g$ , is defined as:

$$K_g = -\frac{R}{3} \frac{\partial p_g}{\partial R}, \quad (16)$$

which is generally a complex number. The elasticity and the energy loss associated with volumetric change of a bubble are represented by the real and imaginary parts of  $K_g$ , respectively. Then the damping factor and the resonant frequency for the bubble oscillation, which correspond to Eqs. (8) and (9), respectively, are [75]:

$$b_t = \frac{3\text{Im}(K_g)}{2\rho_l \omega R_o^2} \quad (17)$$

$$\omega_o = \frac{1}{R_o} \sqrt{\frac{3\text{Re}(K_g) - 2\Sigma/R_o}{\rho_l}}. \quad (18)$$

In the case of an adiabatic process for an ideal gas, where Eq. (6) holds,  $K_g = \gamma p_{g0}$  and Eq. (18) agree with Eq. (9). While in the case of an isothermal process,  $K_g = p_{g0}$ . In these two extreme conditions,  $\text{Im}(K_g) = 0$  and there is no thermal damping.

In order to include the effect of non-uniform temperature distribution in the bubble, we have to solve the energy equation with the continuity of temperature at the bubble surface. Assuming that the pressure in the bubble is uniform, the temperature at the bubble wall is constant, which is supported by the large heat capacity of the liquid compared with that of the gas, and the pressure perturbation is periodic ( $\propto e^{i\omega t}$ ), the effective bulk modulus,  $K_g$ , is represented as [75]:

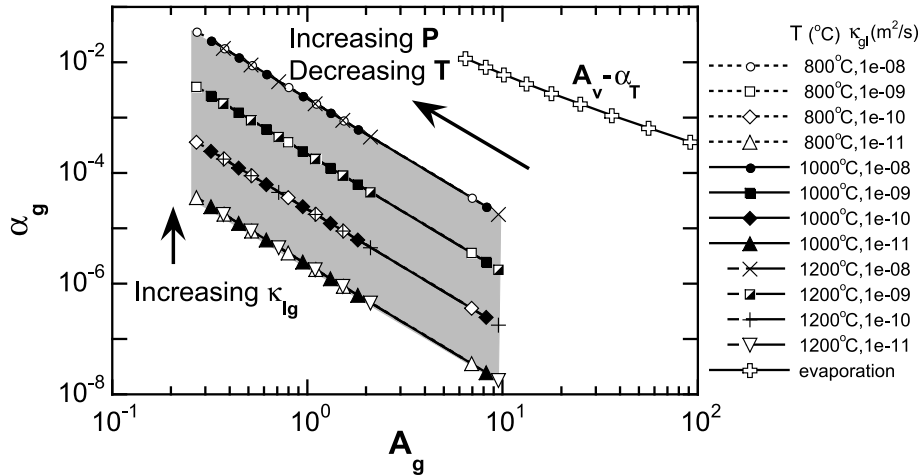
$$\frac{p_{g0}}{K_g} = \frac{1}{\gamma} - \frac{3(\gamma-1)}{\gamma} i\chi \left[ \sqrt{\frac{i}{\chi}} \coth \left( \sqrt{\frac{i}{\chi}} \right) - 1 \right], \quad (19)$$

$$\chi = \frac{\kappa_T}{\omega R^2}. \quad (20)$$

When the mass transfer is controlled by the diffusion of the volatile component in the liquid phase, the diffusion equation in the liquid and the equilibrium condition at the bubble surface are added. Then the effective bulk modulus which includes both the heat and mass transport is

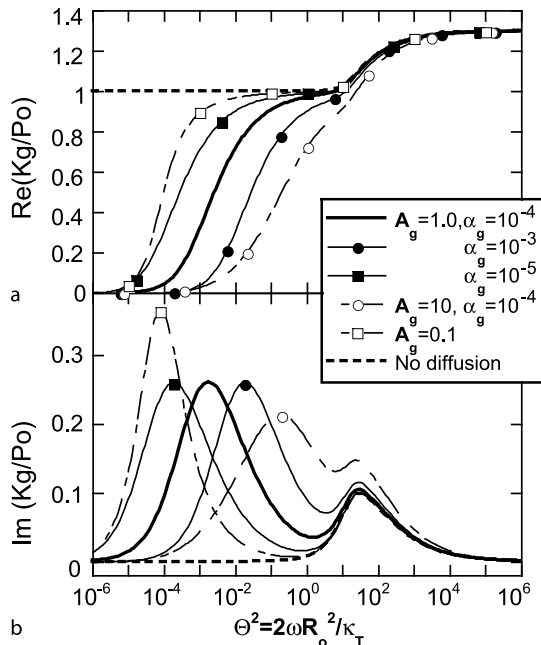
$$\frac{p_{g0}}{K_g} = \frac{1}{\gamma} - \frac{3(\gamma-1)}{\gamma} i\chi \left[ \sqrt{\frac{i}{\chi}} \coth \left( \sqrt{\frac{i}{\chi}} \right) - 1 \right] - 3A_g \sqrt{\alpha_g} i\chi \left( \sqrt{\frac{i}{\chi}} + \sqrt{\alpha_g} \right), \quad (21)$$





Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 2

The relevant range of the dimensionless parameters representing the effect of the volatile transfer in the magmatic system is shown as the gray area. The parameter  $A_g$  and  $\alpha_g$  are defined in Eqs. (22) and (23), respectively. The temperature ( $T$ ) and the volatile diffusivity ( $\kappa_{gl}$ ) are assumed as shown in the legend, and the pressure is varied from 0.1 to 100 MPa. The open crosses are the corresponding parameters for a vapor-bubble system,  $A_v$  and  $\alpha_T$  given in Eqs. (27) and (28), respectively. The temperature is varied from 380 K to 500 K and the pressure is the saturation pressure at each temperature. (Modified from Fig. 1 in [30])



Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 3

The effective bulk modulus of a gas bubble with heat and mass diffusion (calculated by Eq. (21)) as a function of the dimensionless frequency. The real and the imaginary parts are presented in a and b, respectively. The thick broken lines indicate no diffusion and include only thermal effects (which is calculated by Eq. (19)). (Fig. 2 in [30])

$$A_g = \frac{\rho_l p_{go}}{\rho_{go}} \frac{\partial C_{eq}}{\partial p}, \tag{22}$$

$$\alpha_g = \frac{\kappa_{gl}}{\kappa_T}, \tag{23}$$

where  $C_{eq}$  is the saturation concentration at  $p_{go}$  [30]. Equation (21) has the last term in addition to Eq. (19), which represents the effect of the mass transfer. The dimensionless parameter,  $A_g$ , represents the ratio of the volatile mass going into the gas phase from a unit volume of the liquid phase by decompression to the mass change in a unit volume of the gas phase due to expansion.

Figure 2 shows the relevant range of the dimensionless parameters,  $A_g$  and  $\alpha_g$ , for an  $H_2O$  bubble in magma [30]. As temperature decreases or pressure increases,  $A_g$  decreases (Fig. 2) because of the following two reasons. With decreasing temperature,  $\rho_{go}^{-1}$  decreases. In the case of magma,  $C_{eq}(p)$  is approximately proportional to  $\sqrt{p}$  [26] so that  $\partial C_{eq}/\partial p$  decreases with increasing pressure.

The effective bulk modulus of a bubble for some selected values of the parameters in the range is presented in Fig. 3 [30]. The thick broken lines in the figure are obtained by Eq. (19), which includes only the heat transport. In this case, the real part approaches the isothermal bulk modulus and the adiabatic one in the low and high frequencies, respectively. The mass transport makes the bubble stiffness ( $Re(K_g)$ ) smaller, which is the more significant in the lower frequency regime. It is because the bubble has

more time to take in and out the volatile from the liquid in a cycle of the pressure perturbation. The imaginary part for each parameter set has a local peak around

$$\omega \sim \tau_T^{-1} = 15\kappa_T R_o^{-2}, \tag{24}$$

which is the characteristic frequency of the energy loss due to heat transfer. In the case of  $R_o = 10^{-3}$  (m) and  $\kappa_T = 4 \times 10^{-6}$  (m<sup>2</sup> s<sup>-1</sup>), which is the value for H<sub>2</sub>O at 10 MPa and 1273 K [5], the corresponding frequency ( $\omega/(2\pi)$ ) is 9.5 Hz. The imaginary part of  $K_g$  including the diffusion effect has another peak at the characteristic frequency of the energy loss due to the mass transport. The frequency is approximately represented by

$$\omega \sim \tau_g^{-1} = 9\alpha_g \kappa_T A_g^2 R_o^{-2}, \tag{25}$$

which is usually smaller than  $\tau_T^{-1}$  [30]. It is noted that the above model assumes a single bubble in an infinite liquid. When the oscillation period is very long, interaction of the diffusion layers of the adjacent bubbles has to be considered [15].

When the mechanism of the mass exchange between the liquid and the bubble is the evaporation/condensation at the bubble wall, the latent heat plays an important role. Then the thermal diffusion equation in the liquid and the balance between the heat flux through the bubble surface and generation of the latent heat should be added [19,24]. The corresponding bulk modulus of the bubble is represented as [24]:

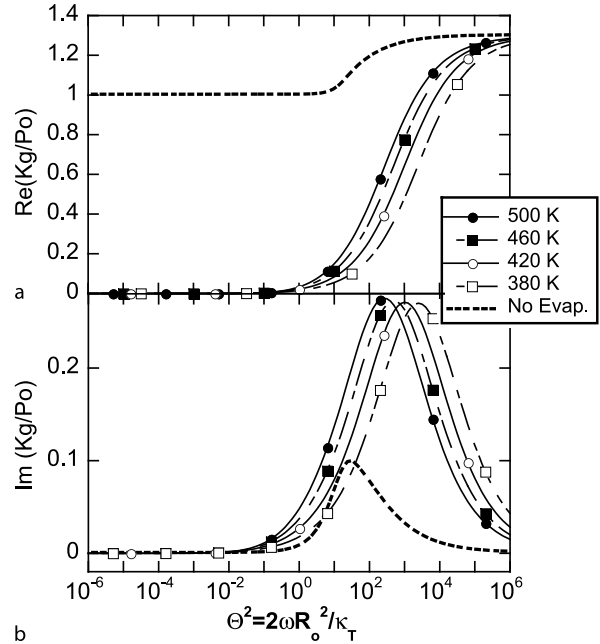
$$\frac{p_{go}}{K_g} = \frac{1}{\gamma} - \frac{3(\gamma - 1)}{\gamma} \left(1 - \frac{c_{pg} T_o}{L}\right)^2 i\chi \left[ \sqrt{\frac{i}{\chi}} \coth \left( \sqrt{\frac{i}{\chi}} - 1 \right) - 3A_v \sqrt{\alpha_T} i\chi \left( \sqrt{\frac{i}{\chi}} + \sqrt{\alpha_T} \right) \right], \tag{26}$$

$$A_v = \frac{\rho_l c_{pl} T_o p_{go}}{(\rho_{go} L)^2}, \tag{27}$$

$$\alpha_T = \frac{\kappa_{Tl}}{\kappa_T}, \tag{28}$$

where  $c_{pg}$  and  $c_{pl}$  are the heat capacity at constant pressure in the gas and the liquid phases, respectively,  $L$  is the latent heat, and  $\kappa_{Tl}$  is the thermal diffusivity in the liquid. In obtaining Eq. (26), the Clausius-Clapeyron relation:  $(dp/dT)_{sat} = L\rho_g/T$ , and thermodynamic relations for an ideal gas are used.

Equation (26) has the same form as Eq. (21) except  $c_{pg} T_o/L$ . This difference is due to the temperature change at the bubble wall. The dimensionless parameter,  $A_v$ , corresponds to  $A_g$  and has a similar physical meaning, which



Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 4

The effective bulk modulus of a vapor bubble with thermal and evaporation effects (calculated by Eq. (26)) as a function of the dimensionless frequency. The real and the imaginary parts are presented in a and b, respectively. The thick broken lines indicate no evaporation and include only thermal effects (which is calculated by Eq. (19))

represents the ratio of the mass going through a phase change in a unit volume of the liquid phase to the mass change in a unit volume of the gas phase due to expansion. Although the equation is similar, the possible range of the parameter is different (Fig. 2). As a result, the frequency dependence of  $K_g$  is also different as is shown in Fig. 4. Comparing the figure with Fig. 3, we can see that the energy loss of the vapor bubble due to the phase change is significant in a frequency range higher than that due to diffusion. The frequency range is comparable to that of the heat transfer, but the amount of energy loss is much larger. The vapor bubble loses its elasticity, which is represented by  $Re(K_g)$ , in the lower frequency as well.

Under the action of the sound field, there is a net transport of heat into the bubble by a non-linear process called the rectified heat transfer [98]. In the evaporation/condensation system, the order of the non-linear effect is so large that it affects the amplitude and damping of the oscillation in the linear regime [24]. Equation (26) does not include the effect. Some works investigating the effect of rectified diffusion process in triggering an eruption or an earthquake are introduced in the later sections.

### Translational Motion

So far we have neglected the translational motion of a bubble relative to the liquid. The translational motion is considered to be negligible when the translational displacement is smaller than the diffusion layer in the liquid surrounding the bubble. In an acoustic field with frequency  $\omega$ , this condition is represented by  $U/\omega < \sqrt{\kappa/\omega}$ , where  $U$  is the translational speed and  $\kappa$  is the relevant diffusivity. The translational velocity of a spherical bubble driven by buoyancy is estimated by  $U = k\rho_l R_o^2 g \eta_l^{-1}$ , where  $g$  is the gravitational acceleration. Although  $k = 1/3$  for a pure liquid,  $k = 2/9$  is used for most of actual liquid, which is not perfectly pure, because the pro-surface components concentrate on the bubble surface to make the surface less mobile [49]. These approximations hold for relatively slow velocity, which satisfies  $Re_t = 2\rho_l R_o U \eta_l^{-1} \leq 1$ . Then the condition in which the translational motion has a minor effect on the heat and mass transfer is

$$\omega > \frac{R^4}{\kappa} \left( \frac{k\rho_l g}{\eta_l} \right)^2. \quad (29)$$

Assuming  $k = 2/9$ ,  $R = 10^{-3}$  (m),  $\kappa = \kappa_{gl} = 10^{-11}$  ( $\text{m}^2 \text{s}^{-1}$ ),  $\rho_l = 2500$  ( $\text{kg m}^{-3}$ ),  $\eta_l = 10^5$  (Pa s) for a magma-H<sub>2</sub>O system,  $\omega > 3 \times 10^{-4}$  ( $\text{rad s}^{-1}$ ) and  $U = 6 \times 10^{-6}$  ( $\text{m s}^{-1}$ ). For a water-vapor system, on the other hand, we assume  $k = 1/3$ ,  $\kappa = \kappa_{Tl} = 10^{-7}$ ,  $\rho_l = 1000$ ,  $\eta_l = 10^{-3}$ . Then, if  $R = 10^{-5}$ ,  $\omega > 1$  and  $U = 3 \times 10^{-4}$ , and if  $R = 10^{-4}$ ,  $\omega > 10^4$  and  $U = 3 \times 10^{-2}$ . According to these estimations, we can see that the translational motion is negligible for most cases with magma except basalt, which has relatively small viscosity ( $\eta_l < 10^2$ ) and large diffusivity ( $\kappa_{gl} \sim 10^{-9}$ ), while it is considerable in hydrothermal systems, except for very small bubbles and the time scale is very short. As an example, the effect on the thermal collapse of a vapor bubble is introduced later.

Another effect of the translational motion is its mechanical coupling with the radial motion. Because a bubble has to move the surrounding liquid in order to make itself move, it is subject to the inertial force of the liquid, which depends on its volume [49]. Therefore when the bubble volume changes, the force also changes. By this consideration, an equation of the translational motion of the bubble is approximately represented as [101]:

$$\dot{U} = -\frac{3}{R} \dot{R}U + 2g - \frac{3}{4} \frac{C_D}{R} |U|U, \quad (30)$$

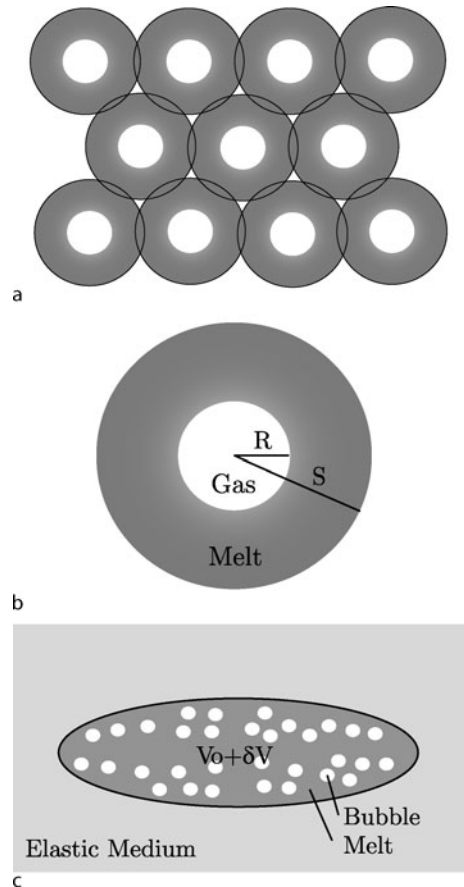
where  $C_D$  is the non-dimensional drag coefficient, which is given as a function of  $Re_t$ . From the first term in the right-hand side, we see that the translational motion is decelerated by the bubble expansion. Although it had been

theoretically recognized for long time, it was quantitatively justified by experiments recently [66]. On the other hand, the effect of the translational motion on the radial motion is represented by the term,  $\rho U^2/2$ . This term can usually be neglected [37,66] except in cases with very strong oscillation [17].

### Bubbly Magma in an Elastic Rock as a Pressure Source

#### Model Overview

Here we consider a magma chamber filled with compressible viscous melt and numerous tiny H<sub>2</sub>O gas bubbles



Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 5

Schematic illustrations of a cell model [79], b an elementary cell, and c a chamber surrounded by an elastic medium. The chamber is filled with compressible viscous melt and numerous tiny spherical gas bubbles. Magma is represented by a combination of many elementary cells.  $R$  and  $S$  is the radius of the elementary cell and gas bubble, respectively, and  $V_0 + \delta V$  is the volume of the chamber. (Modified from Fig. 1 in [87])

(Fig. 5) [79,87]. The magma chamber is confined in an elastic rock. When perturbation is given to the system, pressure may increase by interaction of the elastic deformation of the chamber, expansion of the bubbles, and gas diffusion from the melt to the bubble. Recently, the process has been discussed in the literature in relation to the observed volcanic phenomena [6,13,29,53,64,87,97], which are presented in the introduction.

The melt and bubbles are expressed by the cell model [79], in which multiple spherical bubbles of a constant radius are uniformly packed. Each bubble is surrounded by a finite volume of the melt, represented by an elementary cell. The elementary cell is spherical, in which a single gas bubble is located at the center. It is assumed there is no interaction between neighboring elementary cells such that all gas bubbles grow in the same manner. This simplification enables us to examine bubble growth processes in the entire chamber by studying the growth of just a single bubble, which is represented by Eq. (5).

The main mechanism for increasing the pressure is diffusion of the volatile. It is the slowest process of the bubble dynamics as is described in the previous section. It is certainly longer than the period of resonant oscillation of the individual bubbles so that the inertia terms in Eq. (5) are neglected. It is also longer than the time scale of the heat transport within the bubble as is shown in Fig. 3 so that we may assume uniform and constant temperature within the bubble. Then the mathematical model for the elementary cell consists of three equations, which represent the radial motion of the bubble, volatile diffusion in the melt, and ideal gas approximation, respectively, and three boundary conditions, which are phase equilibrium and mass flux at the bubble surface and no mass flux at the external boundary of the cell element.

**Interaction Between Melt and Elastic Medium**

The volumetric change of the bubbles and the melt due to the pressure change is compensated by the elastic deformation of the chamber. Here we consider the initial pres-

sure and stress conditions in relation to the physical process which brings about the condition, since the relations have not always been mentioned clearly in previous literature. We assume quasi-static deformation of the chamber, where the pressure of the melt is balanced by the elastic stress applied by the wall of the chamber. The volumetric change can be caused by (a) pressure change within the chamber and (b) stress change in the surrounding rock (Fig. 6). Each process is individually represented by

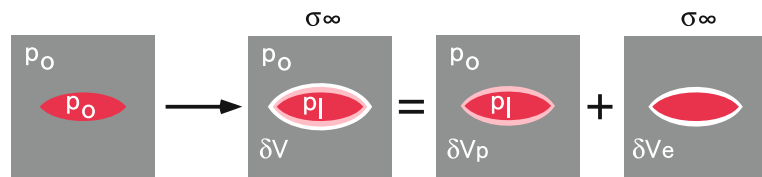
$$p_1 - p_o = \bar{\mu}_p \delta V_p / V_o, \tag{31}$$

$$-\sigma_\infty = \bar{\mu}_e \delta V_e / V_o, \tag{32}$$

where  $V_o$  is the initial equilibrium volume of the chamber,  $\sigma_\infty$  is the ambient stress change,  $\delta V_p$  and  $\delta V_e$  are the volumetric change due to (a) and (b), respectively, and  $\bar{\mu}_p$  and  $\bar{\mu}_e$  are the corresponding effective stiffness of the wall. Each effective stiffness depends on the elasticity of the rock, the geometry of the chamber, and the applied stress field. For the simplicity, we approximate  $\bar{\mu}_p = \bar{\mu}_e = \bar{\mu}$ . Then the total volumetric change,  $\delta V$ , is given by

$$p_1 - p_o - \sigma_\infty = \bar{\mu} \delta V / V_o. \tag{33}$$

The two perturbations which cause the volumetric change have not been clearly distinguished in previous literature. The mathematical treatment by [87] assumed that  $p_1 - p_o = -\Delta p$  is given at  $t = 0$ . They considered that this pressure drop is caused by a decrease of the ambient stress field by a certain amount, say  $\sigma_\infty = -\Delta\sigma$ . On the other hand, the assumption of [13] is that the pressures in all of the bubbles, the melt, and the rock are lower by  $\Delta p$  than the saturation pressure for the dissolved volatile concentration at  $t = 0$ . Strictly speaking, the consequent processes are different depending on what causes the pressure perturbation. If the pressure drop of  $\Delta p$  occurred first within the chamber, that is  $p_1 - p_o = -\Delta p$  and  $\sigma_\infty = 0$ , the chamber would initially shrink according to Eq. (33). If it is caused by an ambient stress drop first, that is  $p_1 - p_o = 0$  and  $\sigma_\infty < 0$ . Then the chamber would ini-



Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 6  
 Mathematical representation for the interaction of melt pressure  $p_1$ , stress in the ambient rock  $\sigma_\infty$ , and volume change of the chamber,  $\delta V$ . Volume change due to internal overpressure  $p_1 - p_o$  and that due to external stress is considered separately

tially expand. In either case, the initial response is almost instantaneous, which is controlled by elasticity of the rock and compressibility of the melt. The major deformation occurs later and is controlled by volumetric change of the bubbles.

**Response to Sudden Decompression and Characteristic Time for Pressure Recovery**

There are three important parameters to characterize the response of the system to the pressure drop which are useful for comparing the model and the field observations. The first one is the re-equilibrated pressure  $p_f$ . The second is the final bubble radius,  $R_f$ . The third is the characteristic time of the recovery process,  $T_{\text{growth}}$ .

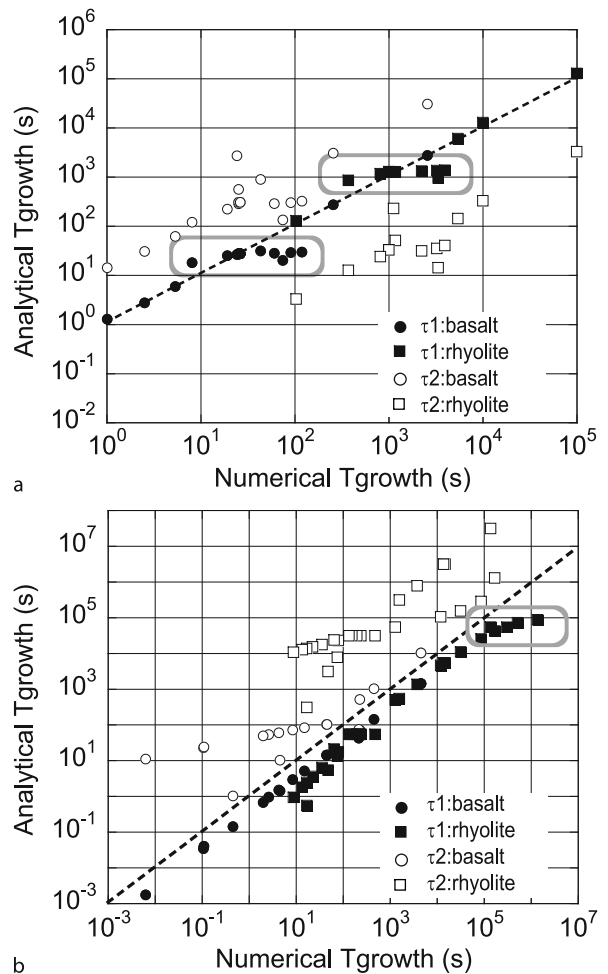
The first and the second are calculated by consideration of the equilibrium condition alone and can be calculated semi-analytically [64]. On the other hand,  $T_{\text{growth}}$  is determined by numerical calculation of the set of equations described above. Shimomura et al. [87] investigated the recovery processes and presented that  $T_{\text{growth}}$  depends on the stiffness of the chamber ( $\bar{\mu}$ ), initial bubble radius ( $R_o$ ), number density of the bubbles ( $N$ ), volatile diffusivity in the melt ( $\kappa_{\text{gl}}$ ), initial pressure ( $p_o$ ), the pressure drop ( $\Delta p$ ), and the melt properties in a complicated manner.

The corresponding study for the bubble growth in an open space, where the pressure is constant regardless of the bubble expansion, was presented by Prousevitch et al. [79]. They assumed an initially supersaturated condition, in which both  $p_{\text{lo}}$  and  $p_{\text{go}}$  are lower than the saturation pressure of the volatile dissolved in the melt. They investigated the final bubble radius and the time to reach it, which correspond to  $R_f$  and  $T_{\text{growth}}$ , respectively. They also presented effects of initial bubble radius ( $R_o$ ), number density of the bubbles ( $N$ ), volatile diffusivity in the melt ( $\kappa_{\text{gl}}$ ), initial pressure ( $p_o$ ), and initial super-saturation.

A simple theory to estimate the time scale of re-equilibration is useful to compare the model with observations, but has not been determined yet. Here we test two hypotheses.

1. The recovery time is comparable with the time scale in which the diffusion layer develops over the entire shell, that is  $\tau_1 = (S_f - R_f)^2 / \kappa_{\text{gl}}$ .
2. Based on a dimensional analysis of the simplified diffusion Eq. (14), Lensky et al. [53] proposed  $\tau_2 = (R_f^2 / \kappa_{\text{gl}})(\rho_{\text{gf}} / \rho_l)(C_o - C_f)^{-1}$ , where  $\rho_{\text{gf}}$  and  $C_f$  are the final gas density in the bubble and the volatile concentration remained in the melt, both of which are functions of  $p_f$ . They obtained this equation based on the approximation that the quasi-static mass flux through the interface is  $(C_o - C_{\text{eq}}(p_g)) / R$ .

The re-equilibration times,  $T_{\text{growth}}$ , obtained by Shimomura et al. [87] and Prousevitch et al. [79] are compared with the above models in Fig. 7. Comparing  $T_{\text{growth}} - \tau_1$  (black symbols) and  $T_{\text{growth}} - \tau_2$  (white symbols), we see that the general trend of  $T_{\text{growth}}$  is better estimated by  $\tau_1$  than by  $\tau_2$  in both confined and open systems. However, it should also be noted that  $\tau_1$  still has systematic errors which are indicated by gray frames. The errors are more dominant in the confined system (Fig. 7a). It is indicated



Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 7  
 Numerical results of the pressure recovery time for magma in an elastic rock **a** and the bubble growth time in an open system **b** for various system parameters are compared with analytical approximations:  $\tau_1$  is the time scale of mass diffusion across the final shell thickness,  $\tau_2$  is approximation by [53]. Agreement between the numerical results and  $\tau_1$  is better, but some systematic discrepancy remains, as indicated by gray frames. The numerical results for **a** are from [87], and those for **b** are from [79]

that the simple estimation does not include all the factors relevant to the re-equilibration time and it is not necessarily applicable to the wider range of the parameters.

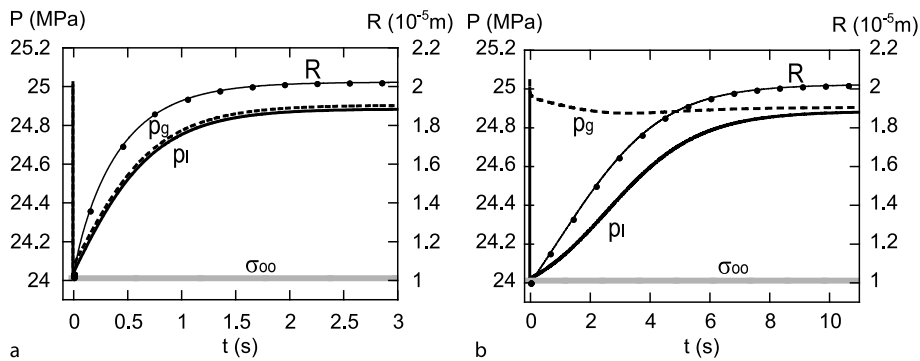
**Re-equilibration Processes**

Here we discuss different re-equilibration processes depending on the cause of the pressure drop and the relevant initial conditions. Three representative solutions are presented in Figs. 8–10. They are obtained for the standard basaltic system [87], but only viscosity is varied from 50 Pa s for (a) to 10<sup>6</sup> Pa s for (b).

Figure 8 is the case in which the stress drop occurs in the ambient rock first. It is generated by, for example, surface unloading by dome collapse [97] and stress change after a local earthquake. The condition is represented by  $\sigma_\infty = -\Delta\sigma$  at  $t \geq 0$  while  $p_l = p_g - 2\Sigma/R_o = p_o$  at  $t = 0$ . According to Eq. (33), the chamber expands in-

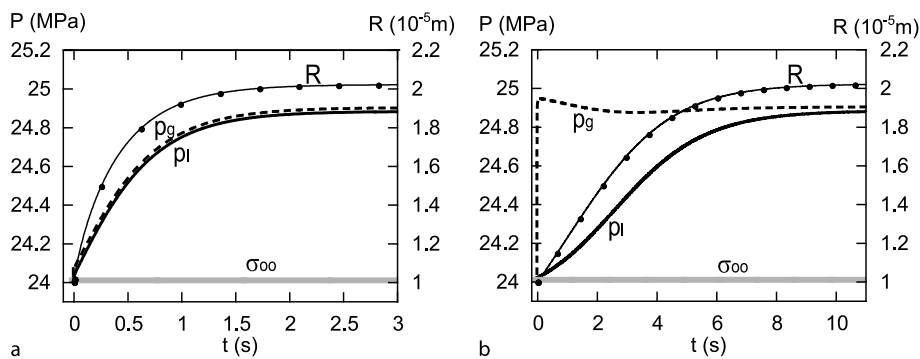
stantaneously, and  $p_l$  drops by  $\Delta p$ . Then the initial condition assumed by [64,87] is attained. Response of  $p_g$  is not instantaneous [62]. Due to the difference between  $p_g$  and  $p_l$ , the bubble expands according to Eq. (5) to decrease  $p_g$ . Then the difference between  $p_g$  and the equilibrium pressure for the volatile concentration in the melt occurs to make the volatile flow into the bubble to re-increase  $p_g$ . As the bubbles grow, the entire volume of the magma ( $\delta V$ ) increases to enlarge the chamber elastically. Then the elastic stress  $\bar{\mu}\delta V/V_o$  increases  $p_l$  according to Eq. (33). The re-equilibration proceeds in this way [87].

Figure 9 is the case in which the pressure in the bubble as well as those in the melt and the ambient rock is lower than the saturation pressure for the initial volatile concentration in the melt at  $t = 0$ . This condition occurs if bubbles are mixed with the supersaturated melt instantaneously, or if the bubbles are kept in the supersaturated mixture without interaction and suddenly allows the dif-



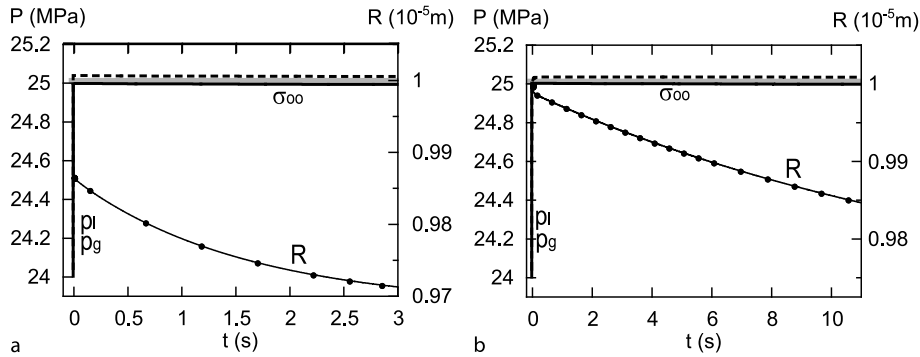
Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 8

Pressure recovery in a bubbly magma in an elastic chamber after sudden unloading  $\sigma_\infty = -1$  MPa. The initial condition is  $p_l = p_g - 2\Sigma/R = 25$  MPa and  $R = 10^{-5}$  m. The bubble radius on the right axis is plotted with a line and points. The stress and pressures on the left axis are plotted with a solid line for  $p_l$ , a dotted line for  $p_g$ , and gray line for  $\sigma_\infty$ . The system parameters are  $\kappa_{gl} = 10^{-8} \text{ m}^2 \text{ s}^{-1}$ , the bubble number density is  $10^{11} \text{ m}^{-3}$ , and  $\eta_l = 50 \text{ Pa s}$  for a and  $10^6 \text{ Pa s}$  for b. The others are the same as those for the basaltic system by [87]



Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 9

Similar to Fig. 8, but the initial condition is  $\sigma_\infty = p_l - p_o = p_g - 2\Sigma/R - p_o = -1$  MPa, with  $p_o = 25$  MPa



Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 10  
 Similar to Fig. 8, but the initial condition is  $\sigma_\infty = 0$  and  $p_l - p_o = p_g - 2\Sigma/R - p_o = -1$  MPa, with  $p_o = 25$  MPa

fusion. Mathematically, the initial condition is equivalent to those assumed by [13] and [79]. The condition is represented by  $p_l - p_o = p_g - 2\Sigma/R_o - p_o = \sigma_\infty = -\Delta p$  at  $t \geq 0$ . Diffusion of the volatile into the bubble starts, which increases  $p_g$  first. Then  $p_g - p_l$  expands the bubble and the chamber to increase  $p_l$  in the same way as the previous case. Practically, the difference between Fig. 8 and Fig. 9 occurs only during a very short period in the beginning, and the subsequent increase of the pressure and volume of the chamber may look the same from outside.

Figure 10 is the case in which the pressure drop occurs in the melt and in the bubble, while the stress in the ambient rock is unchanged. The condition is represented by  $p_l - p_o = p_g - 2\Sigma/R_o - p_o = -\Delta p$  and  $\sigma_\infty = 0$ . Although the assumed initial condition is rather imaginary, this case is presented in order to demonstrate how the response can be different depending on the way the system is decompressed. In fact, it is more realistic that  $p_l$  drops first, while  $p_g$  remains at the initial value. This situation may occur by small leakage of the melt from the system. In this case, the melt pressure just recovers almost instantaneously, because the container compresses the melt according to Eq. (33) and bubbles also compress the melt. No other significant change is expected. On the other hand, if both  $p_l$  and  $p_g$  drop, as in Fig. 10, the pressure still recovers rapidly, but bubbles are compressed. Because the mechanical balance of the bubble and the melt is attained with  $p_g - 2\Sigma/R = p_l$  according to Eq. (5),  $p_g$  has to become larger when  $R$  decreases. Then  $p_g$  exceeds the equilibrium pressure for the volatile in the melt to make the volatile dissolve into the melt.

**Rectified Diffusion**

So far, we discussed responses of the system to a stepwise pressure drop. When the perturbation is caused by a seis-

mic wave from an external source, the system is subject to a cyclic disturbance. Rectified diffusion is a mechanism which can push dissolved volatiles into bubbles in a sound field. Bubbles take in more volatiles during expansion than they discharge during contraction, mainly because of the following two non-linear effects [18,27]. Firstly, the interface is larger during expansion than during contraction. Secondly, radial bubble expansion tangentially stretches the diffusion layer and sharpens the radial gradient of the volatile concentration in the diffusion layer, so that the volatile flux into the bubble is enhanced.

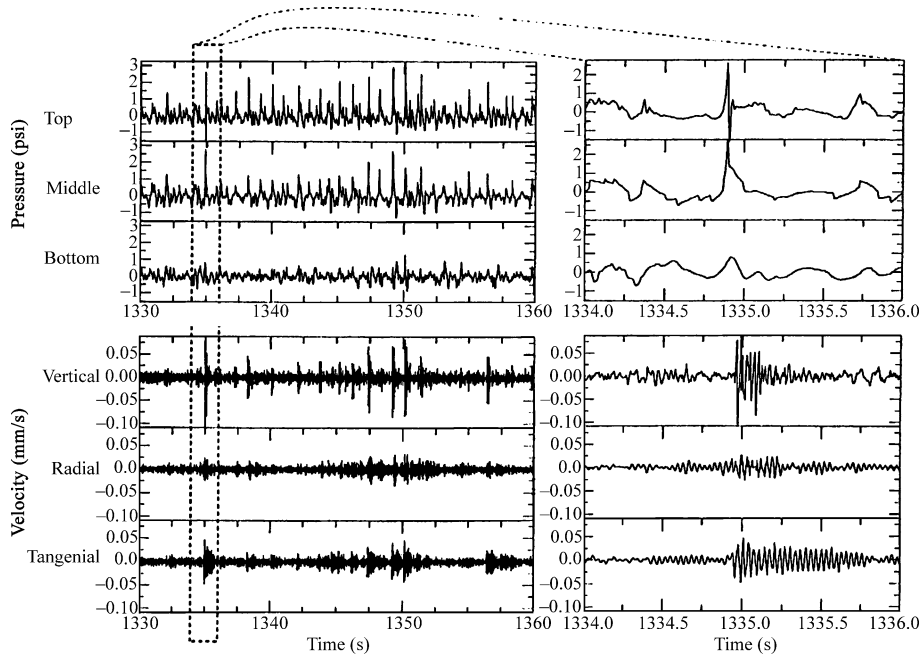
Brodsky et al. [6] discussed the possible pressure increase of a bubbly magma confined in an elastic rock by this mechanism. Using the solution by Hsieh and Plesset [27] for a periodic system, they considered that, even though the net pressure changes are determined by the pre-existing oversaturation, the rectified diffusion accelerates the pressure increase and may break the balance which had been stabilized the system prior to the oscillation. Ichihara and Brodsky [29] improved the solution by including resorption of gas as the pressure increase and development of the diffusion layer around the bubble in a self-consistent way. It is then shown that rectified diffusion is not faster than the ordinary diffusion and its contribution to the net pressure change is at the most  $2 \times 10^{-9}$  of the initial pressure regardless of the pre-existing oversaturation.

**Acoustic Bubbles in Hydrothermal Systems**

**Pressure Impulses Generated in a Geyser**

Here we consider a mixture of water and vapor bubbles, in which we expect effects of bubble oscillations and evaporation.

Kedar et al. [39,40] conducted field experiments at Old Faithful Geyser, Yellowstone. They measured pressure



**Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 11**  
 Simultaneous pressure records (through a high-pass filter at 1 Hz) and seismic traces at Old Faithful geyser, Yellowstone. The geyser's eruptions are 2–5 min long with the interval between them ranging from 30–100 min. The figure shows a 30 s data about 27 min after the previous eruption and about 52 min before the next eruption. The conduit of the geyser is 22 m deep, where the bottom sensor was located. The bottom, middle and top sensors were connected 3 m apart. The seismic station was located at  $\sim 25$  m from the geyser. The data show a direct correspondence between the pressure pulses and the seismic signals that follow them. (Fig. 3 in [39])

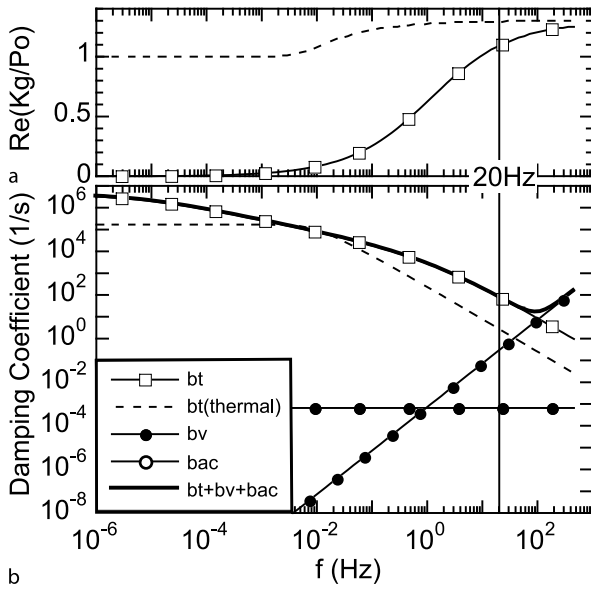
within the geyser's water column simultaneously with seismic measurements on the surface. The data show a distinct cause-and-effect relationship between the impulsive pressure source and the impulse response of the rock surrounding the water column. In addition, the pressure pulse, which is strongest at the top transducer, strongly attenuates downward. Considering that the pulse is generated by the oscillation of a single bubble, they compared one selected signal with a solution of the equation of motion for the bubble radius (Eq. (3)). In order to fit the measured oscillation with a reasonable bubble radius, they had to assume a very small ambient pressure to lower the frequency, and a very large viscosity to increase the damping. For  $R_o = 0.055$  m, for example, they used  $p_{go} = 0.02$  MPa, with which Eq. (9) gives the resonant frequency close to the observation:  $\sim 20$  Hz. The viscosity was assumed as  $\eta_l = 40$  Pa s, which is larger than the actual value by more than four orders. They compared the damping coefficient with those from radiation and heat transfer, though these effects were not included in the calculation of Eq. (3), and concluded that mechanisms other than acoustic, thermal, or viscous damping are required to explain the strong damping observed.

We have already introduced the damping coefficient  $b_t$  with the evaporation effect in Eq. (17) with Eq. (26). Then, assuming the similar bubble radius and frequency ranges as [40], let us see the damping coefficient by evaporation in comparison with the other coefficients, which are for viscous, acoustic and thermal damping, represented by Eqs. (8), (13), and (17) with Eq. (19), respectively. Their values are compared in Fig. 12b, assuming  $p_{go} = 0.13$  MPa (the saturation pressure at 380 K). We can see that the evaporation effect significantly increases the damping and dominates the other damping mechanisms in the frequency range of the geyser oscillation. The evaporation effect also decreases  $\text{Re}(K_g)$  (Fig. 12a) in the range. It is thus suggested that the evaporation effect is significant for the bubble dynamics in the hydrothermal system.

### Inertial and Thermal Collapse of a Bubble

When we heat water in a kettle, we hear strong intermittent pulses before boiling starts. The phenomenon is explained in terms of the bubble dynamics as follows [1]. In the first regime (which initiates above approximately





Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 12  
 The bubble elasticity (a) and damping factors (b) for a single vapor bubble with radius 0.055 m at 380 K, 0.13 MPa (saturation pressure)

40 °C), small bubbles form slowly out of dissolved air in the liquid, rising silently to the surface as they break off the side of the vessel. At higher temperature (~ 70 °C), vapor bubbles start to nucleate at various sites at the heated bottom surface of the container. Vapor bubbles are different from the air bubbles in that their formation and collapse (at the bottom of the vessel) occurs explosively, producing pressure impulses that traverse the liquid and cause much of the sound we hear. In the third stage (between 90 °C and 100 °C), vapor bubbles grow, coalesce, and survive their ascent through the liquid. Bursting of vapor bubbles at the top surface is considered to be the sound source in this regime. Finally, the transition to full boil is characterized by large bubble formation throughout the bulk of the liquid.

We consider whether and how the impulse generation by the bubble collapse occurs in a geyser, where water that is already boiling is injected and cooled from the surface [43]. The collapse (or growth) of a bubble is classified into two modes: the inertia mode, which is controlled by the liquid inertia and driven by the pressure difference between the liquid and the bubble, and the thermal mode, which is controlled by the heat transfer and driven by the temperature difference [20,103]. The former is more violent than the latter and is responsible for the impulse generation.

Based on theoretical and experimental studies, Florschuetz and Chao [20] proposed that the relative importance of the inertia and the heat transfer is evaluated by a dimensionless parameter,  $B$ , defined by

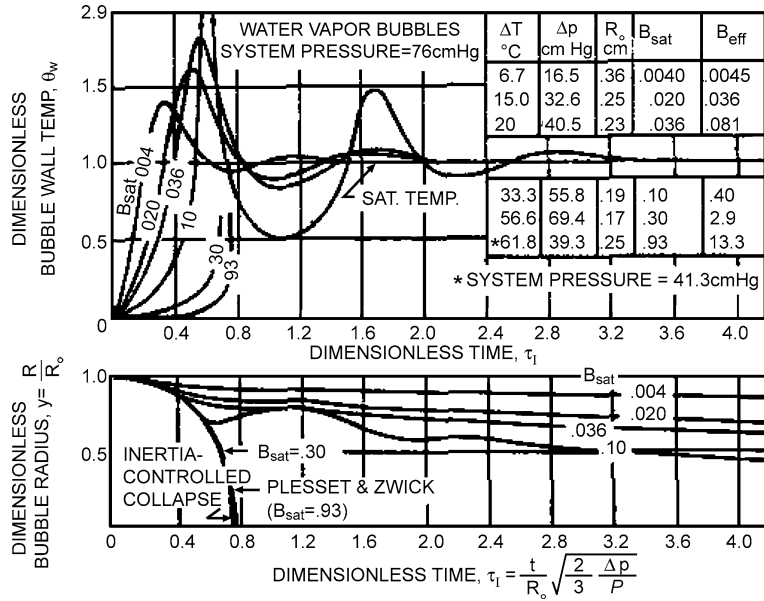
$$Ja = \frac{\rho_l c_{pl}(T_{sat}(p_{l0}) - T_o)}{\rho_{go} L}, \tag{34}$$

$$B = Ja^2 \frac{K_{Tl}}{R_o} \sqrt{\frac{\rho_l}{p_{l0} - p_{sat}(T_o)}}, \tag{35}$$

where  $T_{sat}(p_{l0})$  is the saturation temperature for the ambient pressure ( $p_{l0}$ ) and  $p_{sat}(T_o)$  is the saturation pressure for the system temperature ( $T_o$ ). The dimensionless parameter  $Ja$  is called the Jacob number, which represents the degree of subcooling. Figure 13 displays their calculation results, which clearly shows that for  $B \geq 0.3$ , the collapse rate is dominated by liquid inertia effect, while for  $B \leq 0.03$  it is much slower and is recognized as the thermal mode. For an intermediate value of  $B$ , oscillation is observed.

In these works [20,77], it is often assumed that the pressure in the bubble is initially equal to the saturation pressure of the subcooled liquid, that is  $p_{go} = p_{sat}(T_o)$ , which is less than the ambient pressure ( $p_{l0}$ ) [20,77]. Experimentally, it is achieved by preparing for a thermally equilibrium water-vapor system at a low pressure and suddenly increasing the system pressure to  $p_{l0}$  [20]. Then the initial collapse is relatively violent and continues by inertia until the vapor heating at the bubble wall increases the vapor pressure above the ambient pressure to such an extent that the liquid is momentarily brought to rest and its motion actually reverses before the vapor pressure again falls below the system pressure [20,77]. Although the oscillation is difficult to see on the radius change curves in Fig. 13 for small  $B$ , the beginning inertia controlled stage is evidenced by that all the curves start along the inertia curve.

On the other hand, in case that a bubble suddenly enters cold water,  $p_{go} = p_{l0} > p_{sat}(T_o)$  while temperature in the bubble  $T_{go}$  is larger than  $T_o$  and is close to  $T_{sat}(p_{go})$ . Then the collapse begins in the gentle mode controlled by the heat transfer. It can turn into the inertia mode only if the rate of heat transport and condensation to decrease the vapor pressure is so large that inward motion of the surrounding liquid cannot follow. Prosperetti and Hao [77] presented that relative translational motion between the bubble and the liquid significantly increases the rate of heat transport and accelerates the bubble collapse. Furthermore, they pointed out the coupling effect between the translational and the radial motions. As Eq. (30) suggests, the decreasing bubble radius ( $\dot{R} < 0$ ) works to accelerate the translational motion. Although the drag force



Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 13  
 Variation of wall temperature and radius during collapse of water vapor bubbles.  $B_{sat}$  is the dimensionless parameter given by Eq. (35), which determines relative importance of the liquid inertia and the heat transfer (Fig. 3 in [20])

( $\propto C_D R^{-1} |U|U$ ) increases as  $R$  decreases and  $U$  increases, there are cases in which the contribution of the first term is so large as to make  $\dot{U} > 0$ . Then the collapse and the translational motion of the bubble accelerate each other [77].

**Rectified Heat Transfer**

In the same way as the rectified diffusion discussed in the previous section, rectified heat transfer works for a vapor bubble in an acoustic field [24,98]. When the bubble is compressed, some vapor condenses, the surface temperature rises, and heat is conducted away into the adjacent liquid. When the bubble expands during the following half cycle, evaporation causes a temperature drop of the bubble surface, with a consequent heat flux from the liquid. The imbalance of the heat flux and the interface area between the compression phase and the expansion phase causes the net energy flux into the bubble.

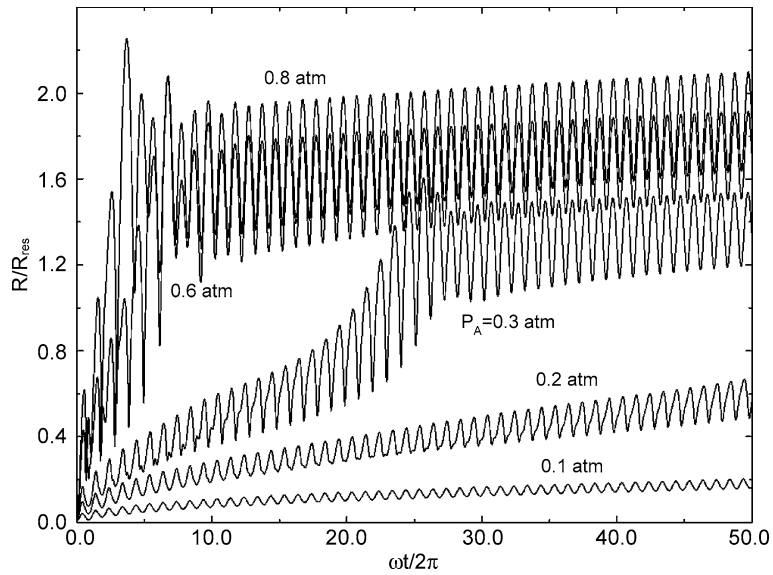
Sturtevant et al. [89] investigated the effect of rectified diffusion on pressure increase in a hydrothermal system, as a possible mechanism for triggered seismicity by a distant earthquake. They modeled the system as a two-component H<sub>2</sub>O-CO<sub>2</sub> system, and considered rectified mass diffusion. As is mentioned in the previous section, the net pressure change due to rectified mass diffusion is very small, if it is evaluated in a self-consistent way [29].

Although rectified heat transfer has a similar mechanism as the rectified mass transfer, it is much more in-

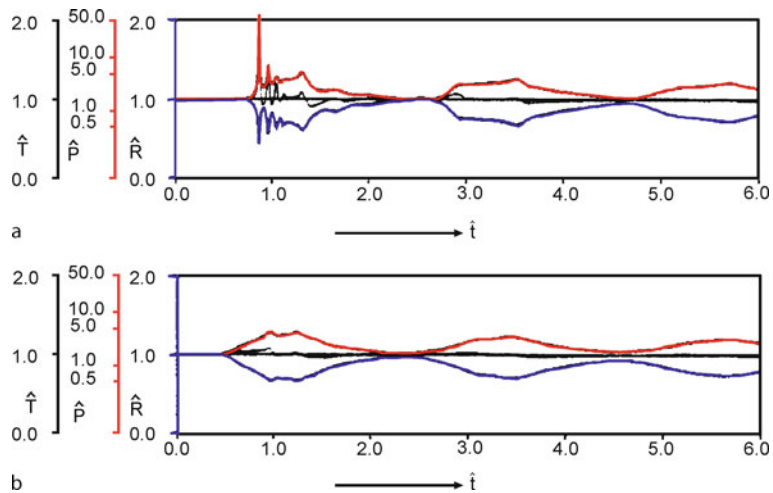
tense since the thermal diffusivity of liquids typically exceeds the mass diffusivity by two orders of magnitude [77]. It can grow a vapor bubble quickly within several cycles of oscillation at the beginning (Fig. 14), and thus can be effective even with low-frequency pressure waves. Moreover the effect is reinforced by the coupling effect of the bubble growth and translational motion [25], and coupling of evaporation and diffusion of another gas component [38]. The net energy flux into the bubble increases the temperature in the bubble, which may change the liquid static pressure [24]. It is noted that numerical results in the literature cannot be directly used to estimate the pressure increase, because they were obtained for an open system in which the bubble continues to grow instead of increasing system pressure. However, it might be worth while to re-evaluate effects of rectified processes in a hydrothermal system taking account of these recent results.

**Non-linear Oscillation of a Spherical Cloud of Bubbles**

Oscillation of a group of bubbles generates particular signals as well as oscillation of a single bubble. The presence of bubbles can lower the acoustic speed of the fluid by an order of magnitude [16,42], if the liquid viscosity is small enough [30,31]. Therefore, there is a sharp impedance contrast between a region of bubbly liquid and a region of pure liquid. The boundary of a bubble cloud acts like an elastic boundary that traps acoustic energy in the bub-



Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 14  
 Growth of a vapor bubble by rectified heat transfer. Bubble radius normalized by the linear resonant radius  $R_r = 2.71$  mm versus time for saturate water at 1 atm. The sound frequency is 1 kHz and the amplitude is attached to each profile (Fig. 4 in [24])



Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 15  
 Change of bubble radius (blue lines) and pressure (red lines) in a spherical cloud of bubbles after a sudden pressure rise. The initial pressure and temperature are  $3 \times 10^4$  Pa, 293 K, radii of the bubble and the cloud are  $2.5 \times 10^{-4}$  m and  $3 \times 10^{-2}$ , the void fraction is 3%, and the pressure rise is  $2.4 \times 10^4$  Pa. Values at the center (a) and at a half the radius from the center (b) are presented in dimensionless form (a unit of the dimensionless time corresponds to  $\sim$  ms). (Modified from Fig. 9 in [67])

bly region so that the bubble cloud has characteristic frequency of resonance [56,67,102]. Chouet [10] considered that it is the mechanism for the harmonic oscillations observed at volcanoes having relatively low-viscosity magma. He showed the typical values for the oscillation in the few Hz range may be generated by a columnar bubble cloud with void fraction of 1%,  $\sim 100$  m in length, and  $\sim 1$  m in

radius in a bubble-free magma with  $\sim 5$  m in radius. The radius of each bubble is assumed as  $10^{-3}$  m, which has its own resonance frequency at 8.9 kHz [10].

Omta [67] conducted numerical calculation for oscillation of a spherical cloud of bubbles with relatively large amplitude. Figure 15 shows one of his results, in which the cloud oscillation was excited by an external pressure in-

crease. Three cycles of the oscillation are presented. We can see high-frequency strong pulsation at the center of the bubble cloud (Fig. 15a). It is explained as follows [67]. The pressure perturbation is amplified and sharpened toward the center of the cloud because of the spherical geometry. Then the bubbles are excited at their resonance frequency.

The strong pulsation is observed only near the center in a spherical bubble cloud (Fig. 15). If a bubble cloud is hemispherical with its cross section attached on a solid wall, the pulsation generated at the center of the hemisphere may strongly hit the wall [86]. Generation of strong high-frequency pulses by a bubble cloud interacting with a pressure perturbation with lower frequencies is actually observed in experiments and now the phenomenon is going to be applied to medical treatment under controlled condition [33,60]. Similar mechanisms may work in a hydrothermal system, in which a low-frequency perturbation generates strong pressure impulses hitting the walls to be observed as seismic waves. In fact pressure oscillations in the bubble cloud (Fig. 15) and in the geyser (Fig. 11) have quite similar features, though their time scales are different by three orders of magnitude.

### Future Directions

We have summarized theoretical bases of the bubble dynamics, mainly based on radial motion equations of a single bubble. These theories have been established and verified by experiments for simple systems. For volcanic systems, these theories have mainly been applied to the nucleation and growth of bubbles in magma. This subject takes an important role of volcanology, though it is not included in this review paper. By comparing the theory with observation of bubbles left in natural volcanic rock [58,94] and re-producing the process by laboratory experiments [23, 52,90], researchers have determined physical parameters of the volcanic processes, which are the temperature, pressure, volatile saturation, ascent rate, and so on. The bubble dynamics theories have also been applied to explain geophysical observations, as we have reviewed several possible mechanisms of bubbles that generate pressure impulses. However, determining the effects or even existence of bubbles is more difficult in these phenomena than in the bubble growth problems, because direct observation of bubbles and re-production of the process in a laboratory are more difficult. Here we discuss how we can go forward to confirm the models and apply them to determine useful physical parameters.

It would be effective to focus on relatively simple volcanic phenomena in which the bubble dynamics theory

appears to work. Especially, for some volcanoes which erupt frequently, data taken by modern geophysical methods are being accumulated and the phenomenological cause-and-result relations between an eruption and pressure impulses before and during the eruption are well documented. For example, at Stromboli Volcano all the sequences of the repetitive small eruptions and a few proximal explosions have been taken by multi-parameter monitoring systems [84,85]. Geophysical data taken close to active craters at Sakurajima, Suwanosejima, and Semeru volcanoes have revealed common features in the pressure change before and during an explosion [32]. At Sakurajima Volcano, behaviors of a shallow gas pocket are discussed in the sequences of seismic and explosion events based on analyses of seismic data [91,92]. It might be possible and useful to have a common backbone model for these eruptions, based on which we can explain the particular detail of each case as a result of different parameters of the system.

The bubble dynamics theory which is used in the models needs to be updated, too. Responses of a bubbly fluid are sensitive to the system parameters, which determine possibilities, features, and time scales of the individual mechanisms, as are shown in the text. Although many models assume uniform system for simplicity, the natural system is considered to be non-uniform. In other words, regions having different physical parameters may coexist. Interaction of these subsystems may enhance the characteristic response, but may diminish with one another, or generate completely different effects. The inhomogeneities and their interactions may occur in various scales and manners. For example, in Sect. “**Bubbly Magma in an Elastic Rock as a Pressure Source**”, behaviors of a single uniform magma body in an elastic rock have been discussed. In the real system, the bubble size and the chemical composition are likely to be non-uniform within a small region and/or over the entire magma body. If the system is large, the hydrostatic pressure gradient is significant. Moreover, if there are multiple magma containers which have different physical parameters, each magma body will respond to the pressure perturbation differently, and pressure gradient may be generated between the two adjacent containers. Developing a model including these subscale interactions may be a subject of modern multi-scale multi-physics studies.

Laboratory experiments using analogous materials are useful in verifying and improving the models. In the procedure to construct a model system, we frequently find factors which are important in the real system but have been neglected in the idealized mathematical model. Although there may be some processes which can be realized in the

nature more easily, and there always be a scaling problem, experiments will give us more concrete idea about the mechanism of the models. Although most of the previous analogue experiments are designed to be compared with geological and petrological observations, there are some which investigate generation of pressure impulses by bubbles and are intended to explain seismic and/or acoustic observations [34,35,83]. According to the results and implications obtained by these preceding works, laboratory studies in this direction are promising.

It is also important to connect geophysical observation and geological data to understand the bubble dynamics phenomena in volcanology. Compared with other geophysical processes which occur in the earth, there is larger possibility that the source of activity appears to the surface after relatively short time. The geological samples (e.g., pyroclasts during eruptions and volcanic gasses) can inform us physical and chemical properties of the materials generating pressure impulses, and would make useful constraints on the model. On the other hand, these constraints can be tested and verified by geophysical signals of seismic, geodetic and acoustic measurements when the models are established.

By combining these theoretical, experimental, geophysical, and geological approaches, we will get better understanding on the processes which generate volcanic activities.

### Acknowledgments

The authors are grateful to Dr. B. Chouet, and Dr. H. Kawashima for useful information and advise. We also thank Dr. M. Kameda and two anonymous reviewers for their help to improve the manuscript.

### Bibliography

- Aljishi S, Tatarkevich J (1991) Why does heating water in a kettle produce sound? *Am J Phys* 59:628–632
- Barclay J, Riley DS, Sparks RSJ (1995) Analytical models for bubble growth during decompression of high viscosity magmas. *Bull Volcanol* 57:422–431
- Benoit JP, McNutt SR (1997) New constraints on source processes of volcanic tremor at Arenal volcano, Costa Rica, using broadband seismic data. *Geophys Res Lett* 24:449–452
- Blower JD, Mader HM, Wilson SDR (2001) Coupling of viscous and diffusive controls on bubble growth during explosive volcanic eruptions. *Earth Planet Sci Lett* 193:47–56
- Bowers TS (1995) Pressure-volume-temperature properties of H<sub>2</sub>O CO<sub>2</sub> fluids. In: Ahrens TJ (ed) *A Handbook of Physical Constants: Rock Physics and Phase Relations*. AGU Reference Shelf Series 3, AGU, pp 45–72
- Brodsky EE, Sturtevant B, Kanamori H (1998) Earthquakes, volcanoes, and rectified diffusion. *J Geophys Res* 103:23827–23838
- Brodsky EE, Karakostas V, Kanamori H (2000) A new observation of dynamically triggered regional seismicity: Earthquakes in Greece following the august, 1999 Izmit, Turkey earthquake. *Geophys Res Lett* 27:2741–2744
- Brodsky EE, Roeloffs E, Woodcock D, Gall I, Manga M (2003) A mechanism for sustained groundwater pressure changes induced by distant earthquakes. *J Geophys Res* 108(B8):2390. doi: 10.1029/2002JB002321
- Campos FB, Lage PLC (2000) Heat and mass transfer modeling during the formation and ascension of superheated bubbles. *Int J Heat Mass Transf* 43:2883–2894
- Chouet BA (1996) New methods and future trends in seismological volcano monitoring. In: Scarpa R, Tilling R (eds) *Monitoring and Mitigation of Volcano Hazards*. Springer, Berlin, pp 23–97
- Chouet B, Dawson P, Ohminato T, Martini M (2003) Source mechanisms of explosions at Stromboli volcano, Italy, determined from moment-tensor inversions of very-long-period data. *J Geophys Res* 108
- Chouet B, Dawson P, Arciniega-Ceballos A (2005) Source mechanism of vulcanian degassing at Popocatepetl volcano, Mexico, determined from waveform inversions of very long period signals. *J Geophys Res* 110:B07301
- Chouet B, Dawson P, Nakano M (2006) Dynamics of diffusive bubble growth and pressure recovery in a bubbly rhyolitic melt embedded in an elastic solid. *J Geophys Res* 111: B07310
- Cole RH (1948) *Underwater Explosions*. Dover, New York
- Collier L, Neuberg JW, Lensky N, Lyakhovsky V, Navon O (2006) Attenuation in gas-charged magma. *J Volcanol Geotherm Res* 153:21–36
- Commander KW, Prosperetti A (1989) Linear pressure waves in bubbly liquid – comparison between theory and experiments. *J Acoust Soc Am* 85:732–746
- Doinikov AA (2005) Equations of coupled radial and translational motions of a bubble in a weakly compressible liquid. *Phys Fluids* 17:128101
- Eller A, Flynn HG (1965) Rectified diffusion during nonlinear pulsations of cavitation bubbles. *J Acoust Soc Am* 37:493–503
- Finch RD, Neppiras EA (1973) Vapor bubble dynamics. *J Acoust Soc Am* 53:1402–1410
- Florschuetz LW, Chao BT (1965) On the mechanics of vapor bubble collapse. *Trans ASME, J Heat Transf* 87:209–220
- Fogler HS, Goddard JD (1970) Collapse of spherical cavities in viscoelastic fluids. *Phys Fluids* 13:1135–1141
- Garces MA, McNutt SR (1997) Theory of the airborne sound field generated in a resonant magma conduit. *J Volcanol Geotherm Res* 78:155–178
- Gardner JE, Hilton M, Carroll MR (1999) Experimental constraints on degassing of magma: isothermal bubble growth during continuous decompression from high pressure. *Earth Planet Sci Lett* 168:201–218
- Hao Y, Prosperetti A (1999) The dynamics of vapor bubbles in acoustic pressure fields. *Phys Fluids* 11:2008–2019
- Hao Y, Prosperetti A (2002) Rectified heat transfer into translating and pulsating vapor bubbles. *J Acoust Soc Am* 112:1787–1796
- Holloway JR, Blank JG (1994) Application of experimental results to C-O-H species in natural melts. In: Carroll MR, Holloway JR (eds) *Volatiles in Magma*. *Rev Miner*, vol 30. Mineral Soc Am, Washington, pp 187–230

27. Hsieh DY, Plesset MS (1961) Theory of rectified diffusion of mass into gas bubbles. *J Acoust Soc Am* 33:206–215
28. Ichihara M (2007) Dynamics of a spherical viscoelastic shell: Implications to a criterion for fragmentation/expansion of bubbly magma. *Earth Planet Sci Lett* 265:18–32
29. Ichihara M, Brodsky EE (2006) A limit on the effect of rectified diffusion in volcanic systems. *Geophys Res Lett* 33:L02316
30. Ichihara M, Kameda M (2004) Propagation of acoustic waves in a visco-elastic two-phase system: influences of the liquid viscosity and the internal diffusion. *J Volcanol Geotherm Res* 137:73–91
31. Ichihara M, Ohkunitani H, Ida Y, Kameda M (2004) Dynamics of bubbly oscillation and wave propagation in viscoelastic liquids. *J Volcanol Geotherm Res* 129:37–60
32. Iguchi M, Tamaguri T, Yakiwara H (2006) Source mechanisms of volcanic explosion revealed by geophysical observations at Sakurajima, Suwanosejima and Semeru volcanoes. *Eos Trans AGU* 87 (Fall Meet Suppl); Abstract V31G–03
33. Ikeda T, Yoshizawa S, Tosaki M, Allen JS, Takagi S, Ohta N, Kitamura T, Matsumoto Y (2006) Cloud cavitation control for lithotripsy using high intensity focused ultrasound. *Ultrasound Med Biol* 32:1383–1397
34. James MR, Lane SJ, Chouet B, Gilbert JS (2004) Pressure changes associated with the ascent and bursting of gas slugs in liquid-filled vertical and inclined conduits. *J Volcanol Geotherm Res* 129:61–82
35. James MR, Lane SJ, Chouet BA (2006) Gas slug ascent through changes in conduit diameter: Laboratory insights into a volcano-seismic source process in low-viscosity magmas. *J Geophys Res* 111:B05201
36. Johnson JB, Aster RC, Kyle PR (2004) Triggering of volcanic eruptions. *Geophys Res Lett* 31:L14604. doi:10.1029/2004GL020020
37. Kameda M, Matsumoto Y (1996) Shock waves in a liquid containing small gas bubbles. *Phys Fluids* 8:322–335
38. Kawashima H, Ichihara M, Kameda M (2001) Oscillation of a vapor/gas bubble with heat and mass transport. *Trans JSME, Ser B* 67:2234–2242
39. Kedar S, Sturtevant B, Kanamori H (1996) The origin of harmonic tremor at old faithful geyser. *Nature* 379:708–711
40. Kedar S, Kanamori H, Sturtevant B (1998) Bubble collapse as the source of tremor at old faithful geyser. *J Geophys Res* 103:24283–24299
41. Keller JB, Kolodner II (1956) Damping of underwater explosion bubble oscillations. *J Appl Phys* 27:1152–1161
42. Kieffer SW (1977) Sound speed in liquid-gas mixtures: water-air and water-steam. *J Geophys Res* 82:2895–2904
43. Kieffer SW (1989) Geologic nozzles. *Rev Geophysics* 27:3–38
44. Kumagai H, Chouet BA (2000) Acoustic properties of a crack containing magmatic or hydrothermal fluids. *J Geophys Res* 105:25493–25512
45. Kumagai H, Chouet BA (2001) The dependence of acoustic properties of a crack on the resonance mode and geometry. *Geophys Res Lett* 28:3325
46. Kumagai H, Chouet BA, Nakano M (2002) Temporal evolution of a hydrothermal system in Kusatsu–Shirane volcano, Japan, inferred from the complex frequencies of long-period events. *J Geophys Res* 107:2236
47. La Femina PC, Connor CB, Hill BE, Strauch W, Saballos JA (2004) Magma-tectonic interactions in Nicaragua: the 1999 seismic swarm and eruption of Cerro Negro volcano. *J Volcanol Geotherm Res* 137:187–199
48. Landau LD, Lifshitz EM (1986) *Theory of Elasticity*, 3rd edn. Butterworth, Oxford
49. Landau LD, Lifshitz EM (1987) *Fluid Mechanics*, 2nd edn. Pergamon Press, Oxford
50. Lensky NG, Lyakhovskiy V, Navon O (2001) Radial variations of melt viscosity around growing bubbles and gas overpressure in vesiculating magmas. *Earth Planet Sci Lett* 186:1–6
51. Lensky NG, Lyakhovskiy V, Navon O (2002) Expansion dynamics of volatile-supersaturated liquids and bulk viscosity of bubbly magmas. *J Fluid Mech* 460:39–56
52. Lensky NG, Navon O, Lyakhovskiy V (2004) Bubble growth during decompression of magma: experimental and theoretical investigation. *J Volcanol Geotherm Res* 129:7–22
53. Lensky NG, Niebo RW, Holloway JR, Lyakhovskiy V, Navon O (2006) Bubble nucleation as a trigger for xenolith entrapment in mantle melts. *Earth Planet Sci Lett* 245:278–288
54. Linde AT, Sacks I (1998) Triggering of volcanic eruptions. *Nature* 395:888–890
55. Linde AT, Sacks I, Johnston MJS, Hill DP, Bilham RG (1994) Increased pressure from rising bubbles as a mechanism for remotely triggered seismicity. *Nature* 371:408–410
56. Lu NQ, Prosperetti A, Yoon SW (1990) Underwater noise emissions from bubble clouds. *IEEE J Ocean Eng* 15:275–281
57. Manga M, Brodsky E (2006) Seismic triggering of eruptions in the far field: Volcanoes and geysers. *Ann Rev Earth Planet Sci* 34:263–291
58. Mangan MT, Cashman KV (1996) The structure of basaltic scoria and reticulite and inferences for vesiculation, foam formation, and fragmentation in lava fountains. *J Volcanol Geotherm Res* 73:1–18
59. Matsumoto Y, Takemura F (1994) Influence of internal phenomena on gas bubble motion (effects of thermal diffusion, phase change on the gas–liquid interface and mass diffusion between vapor and noncondensable gas in the collapsing phase). *JSME Int J, Ser B* 37:288–296
60. Matsumoto Y, Allen JS, Yoshizawa S, Ikeda T, Kaneko Y (2005) Medical ultrasound with microbubbles. *Exp Therm Fluid Sci* 29:255–265
61. Nakoryakov VE, Pokusaev BG, Shreiber IR (1993) *Wave Propagation in Gas-Liquid Media*, 2nd edn. CRC Press, Boca Raton
62. Navon O, Lyakhovskiy V (1998) Vesiculation processes in silicic magmas. In: Gilbert JS, Sparks RSJ (eds) *The Physics of Explosive Volcanic Eruption*. Geol Soc, Special Publications, 145, London, pp 27–50
63. Nigmatulin RI, Khabeev NS, Nagiev FB (1981) Dynamics, heat and mass-transfer of vapor-gas bubbles in a liquid. *Int J Heat Mass Trans* 24:1033–1044
64. Nishimura T (2004) Pressure recovery in magma due to bubble growth. *Geophys Res Lett* 31:L12613
65. Nishimura T, Ichihara M, Ueki S (2006) Investigation of the Onikobe geyser, NE Japan, by observing the ground tilt and flow parameters. *Earth Planets Space* 58:e21–e24
66. Ohl CD, Tjink A, Prosperetti A (2003) The added mass of an expanding bubble. *J Fluid Mech* 482:271–290
67. Omta R (1987) Oscillations of a cloud of bubbles of small and not so small amplitude. *J Acoust Soc Am* 82:1018–1033
68. Oura A, Yoshida S, Kudo K (1992) Rupture process of the Ito-Oki Japan earthquake of 1989 July 9 and interpretation as a trigger of volcanic-eruption. *Geophys J Int* 109:241–248

69. Plesset MS (1949) The dynamics of cavitation bubbles. *J Appl Mech* 16:277–282
70. Plesset MS, Prosperetti A (1977) Bubble dynamics and cavitation. *Ann Rev Fluid Mech* 9:145–185
71. Poritsky H (1952) The collapse or growth of a spherical bubble or cavity in a viscous fluid. *Proc First Nat Cong Appl Mech* 813–821
72. Prosperetti A (1982) A generalization of the rayleigh-plesset equation of bubble dynamics. *Phys Fluids* 25:409–410
73. Prosperetti A (1984) Acoustic cavitation series: part three, bubble phenomena in sound fields: part two. *Ultrasonics* 22:115–124
74. Prosperetti A (1984) Acoustic cavitation series: part two, bubble phenomena in sound fields: part one. *Ultrasonics* 22:69–78
75. Prosperetti A (1991) The thermal behavior of oscillating gas bubbles. *J Fluid Mech* 222:587–616
76. Prosperetti A (2004) Bubbles. *Phys Fluids* 16:1852–1865
77. Prosperetti A, Hao Y (2002) Vapor bubbles in flow and acoustic fields. *Ann NY Acad Sci* 974:328–347
78. Prosperetti A, Lezzi A (1986) Bubble dynamics in a compressible liquid, 1. 1st-order theory. *J Fluid Mech* 168:457–478
79. Prousevitch AA, Sahagian DL, Anderson AT (1993) Dynamics of diffusive bubble growth in magmas: Isothermal case. *J Geophys Res* 98:22283–22307
80. Rayleigh L (1917) On the pressure developed in a liquid during the collapse of a spherical cavity. *Philos Mag* 34:94–98
81. Ripepe M, Gordeev E (1999) Gas bubble dynamics model for shallow volcanic tremor at Stromboli. *J Geophys Res* 104:10639–10654
82. Ripepe M, Poggi P, Braun T, Gordeev E (1996) Infrasonic waves and volcanic tremor at Stromboli. *Geophys Res Lett* 23:181–184
83. Ripepe M, Ciliberto S, Della Schiava M (2001) Time constraints for modeling source dynamics of volcanic explosions at Stromboli. *J Geophys Res* 106:8713–8727
84. Ripepe M, Marchetti E, Poggi P, Harris A, Fiaschi AJL, Olivieri G (2004) Seismic, acoustic and thermal network monitors the 2003 eruption of Stromboli volcano. *EOS, Trans AGU* 85:329
85. Ripepe M, Marchetti E, Olivieri G, DelleDonne D, Genco R, Laccana G (2007) Monitoring the 2007 Stromboli effusive eruption by an integrated geophysical network. *The 21st Century COE Earth Sci Int Symp Abstr Z* 327
86. Shimada M, Matsumoto Y, Kobayashi T (2006) Influence of the nuclei size distribution on the collapsing behavior of the cloud cavitation. *JSME Int J Ser B* 155:307–322
87. Shimomura Y, Nishimura T, Sato H (2006) Bubble growth processes in magma surrounded by an elastic medium. *J Volcanol Geotherm Res* 155:307–322
88. Sparks RSJ (1978) Dynamics of bubble formation and growth in magmas – review and analysis. *J Volcanol Geotherm Res* 3:1–37
89. Sturtevant B, Kanamori H, Brodsky EE (1996) Seismic triggering by rectified diffusion in geothermal systems. *J Geophys Res* 101:25269–25282
90. Suzuki Y, Gardner JE, Larsen JF (2007) Experimental constraints on syneruptive magma ascent related to the phreatomagmatic phase of the 2000AD eruption of Usu volcano, Japan. *Bull Volcanol* 69:423–444
91. Tameguri T, Iguchi M, Ishihara K (2002) Mechanism of explosive eruptions from moment tensor analyses of explosion earthquakes at Sakurajima volcano. *Bull Volcanol Soc Japan* 47:197–215
92. Tameguri T, Maryanto S, Iguchi M (2007) Source mechanisms of harmonic tremors at Sakurajima volcano. *Bull Volcanol Soc Japan* 52:273–279
93. Toramaru A (1995) Numerical study of nucleation and growth of bubbles in viscous magmas. *J Geophys Res* 100:1913–1931
94. Toramaru A (2006) BND (bubble number density) decompression rate meter for explosive volcanic eruptions. *J Volcanol Geotherm Res* 154:303–316
95. Vergnolle S, Brandeis G (1994) Origin of the sound generated by Strombolian explosions. *Geophys Res Lett* 21:1959–1962
96. Vergnolle S, Brandeis G (1996) Strombolian explosions. 1. A large bubble breaking at the surface of a lava column as a source of sound. *J Geophys Res* 101:20433–20447
97. Voight B, Linde AT, Sacks IS, Mattioli GS, Sparks RSJ, Elsworth D, Hidayat D, Malin PE, Shalev E, Widiwijayanti C, Young SR, Bass V, Clarke A, Dunkley P, Johnston W, McWhorter N, Neuberger J, Williams P (2006) Unprecedented pressure increase in deep magma reservoir triggered by lava-dome collapse. *Geophys Res Lett* 33:L03,312
98. Wang T (1974) Rectified heat transfer. *J Acoust Soc Am* 56:1131–1143
99. Webb SL (1997) Silicate melts: Relaxation, rheology, and the glass transition. *Rev Geophys* 35:191–218
100. Yamada K, Emori H, Nakasawa N (2006) Bubble expansion rates in viscous compressible liquid. *Earth Planets Space* 58:865–872
101. Yang B, Prosperetti A, Takagi S (2003) The transient rise of a bubble subject to shape or volume changes. *Phys Fluids* 15:2640–2648
102. Yoon SW, Crum LA, Prosperetti A, Lu NQ (1991) An investigation of the collective oscillations of a bubble cloud. *J Acoust Soc Am* 89:700–706
103. Zuber N (1961) The dynamics of vapor bubbles in nonuniform temperature fields. *Int J Heat Mass Transf* 2:83–98

## Regional Climate Models: Linking Global Climate Change to Local Impacts

DANIELA JACOB  
Max-Planck-Institute for Meteorology,  
Hamburg, Germany

### Article Outline

Glossary  
Definition of the Subject  
Introduction  
Basic Features and Model Characteristics  
Validation  
IPCC-Scenarios  
Regional Climate Change  
Regional Extremes  
Future Directions  
Acknowledgments  
Bibliography

### Glossary

**Climate models** They are mathematical representations of the Earth system, in which physical and biogeochemical processes are described numerically. Climate models can be of a global scale or focus on a sub-region (regional climate model).

**Downscaling** Dynamical and statistical techniques to interpret global climatic changes in specific regions.

**IPCC emission scenario** Description of possible developments of the socio-economic system expressed in terms of emissions into the atmosphere.

**Projection** Simulation of possible climatic changes in the future, dependent on emission scenarios, land-use changes and natural variability in the climate system.

**Validation** Comparison of observed data against model result for quality assessment of the model.

### Definition of the Subject

A variety of observations demonstrates that during the last decades the climate has changed. As reported by the *Intergovernmental Panel on Climate Change* (IPCC, 2001, 2007), a mean increase of temperature by 0.09 K per decade was observed globally from 1951 to 1989. Up to now, 2007, this trend has continued. Europe experienced an extraordinary heat wave in summer 2003, with daily mean temperatures being about 10° warmer locally than the long term mean. The increase of temperature varies depending on the region and season.

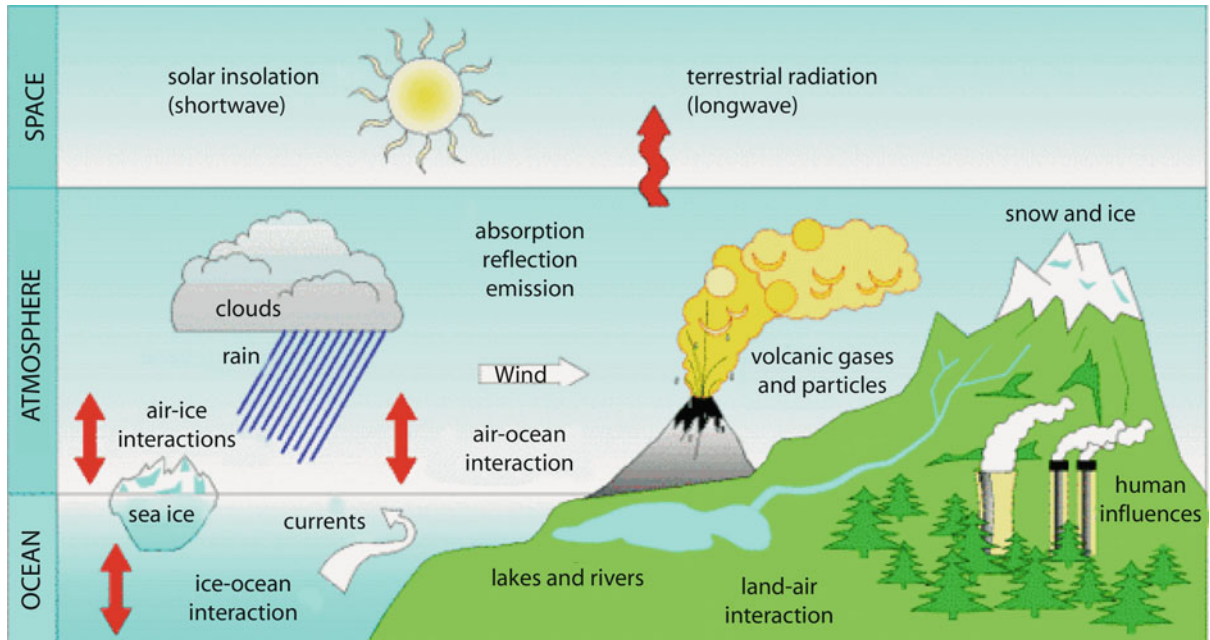
The temperature change seems to be accompanied by changes in several meteorological and hydrological quantities, like number and duration of heat waves, frost periods, storminess or monthly mean precipitation. In Germany, for example, winter precipitation has increased in parts by more than 30% within the last four decades. In addition, very intense precipitation was observed in summer 2002 in parts of the Elbe drainage basin, which faced a severe flooding.

It can be expected that extreme weather situations will occur more often in a warming world. Therefore, a growing demand from decision-makers and the general public for detailed information on possible future climate development is evident, worldwide. The quantification of risks associated with changing climates is a prerequisite for the formulation and implementation of realistic adaptation and mitigation strategies [1].

Global climate models (GCM) have been developed to study the Earth's climate system in the past and future. Unfortunately, even today, global climate models provide information only at a relatively coarse spatial scale, which is often not suitable for regional climate change assessments. To fill this gap, two different principles to transfer the information from a global model to the region of interest have been developed accordingly: statistical downscaling and dynamical downscaling. Statistical downscaling techniques connect the climate change signal provided by the GCM with observations from measurement stations in the region to achieve higher resolved climate change signals.

Dynamical downscaling uses high resolution three-dimensional regional climate models (RCM), which are nested into GCMs. RCMs are similar to numerical weather forecasting models, which are taken into account non-linear processes in the climate system. The results of both downscaling methods depend on both the quality of the global and regional models. In the following, the focus will be on dynamical downscaling, in order to be able to also detect more easily new extremes, which have not been observed so far, and to take into account possible feedback mechanisms, which might appear under climate change conditions, and which influence the extent of regional climatic changes. Regional feedback mechanisms are, for example, snow-albedo/temperature feedbacks or soil moisture-temperature feedbacks. If snow melts the reflectivity of the surface changes from bright (white snow) to dark (vegetation or soil). This enhances the absorption of incoming radiation and leads to warming of the surface, which in turn accelerates the snow melt nearby. Evaporation from soils and vegetation increases with temperature and decreases soil moisture. Drier soils evaporate less,





Regional Climate Models: Linking Global Climate Change to Local Impacts, Figure 1  
The physical climate system

so that cooling due to evaporation is decreasing, which in turn increases temperatures regionally.

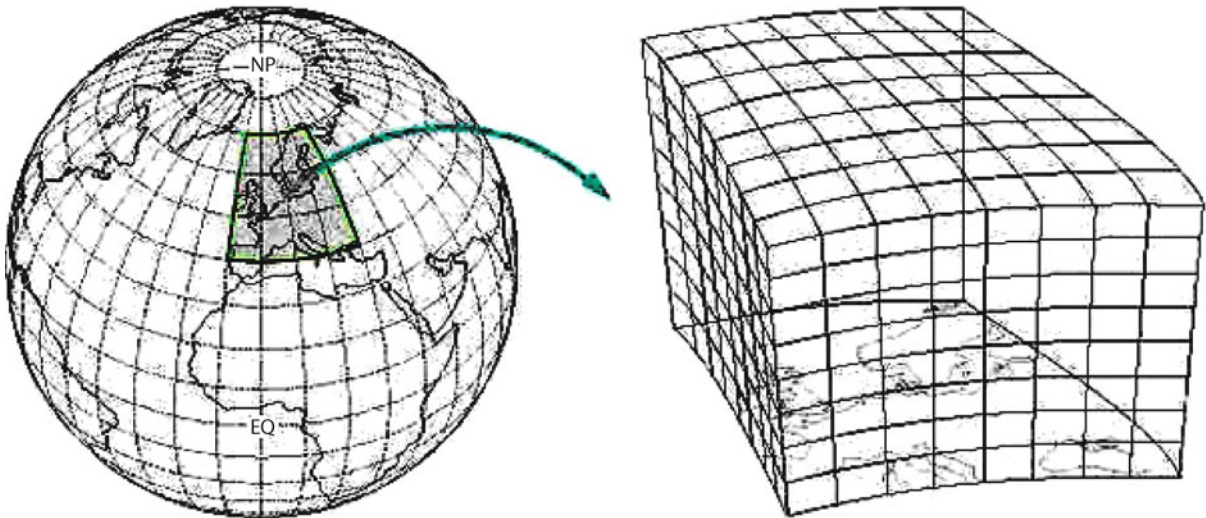
### Introduction

The climate on Earth varies from region to region, and is characterized by sequences of weather events. These events are determined by the atmospheric flow, established through a wide range of interacting scales. The interacting scales cover large-scale features of the order of thousands to hundreds of km, mostly determined by the distribution of the continents and oceans, solar radiation and the composition of the atmosphere, regional features of the order of a few hundred km to a few km, forced through complex topography and vegetation distribution, and small scale features, like convection. The description of the Earth's climate needs to consider all scales, respectively. Therefore global climate models (GCMs) have been developed. They are mathematical representations of the Earth system (Fig. 1), in which physical and biogeochemical processes are described numerically to simulate the climate system as realistically as possible. Today GCMs develop into Earth system models (ESM), which are not only coupled atmosphere-ocean general circulation models (AOGCM), but also take into account some biogeochemical feedbacks, like the carbon cycle or dynamical

vegetation. They are the most advanced numerical tools for climate modeling and describe changes due to large scale forcing.

Increasing greenhouse gas (GHG) concentration, changing aerosol composition and load as well as land surface changes are influencing the climate of the Earth, globally and regionally. Therefore the demands for fine scale regional climate information were growing and in response to this the development of regional climate models started about 20 years ago. It is obvious that the simulations of regional climate changes requires the simulations of processes from global to local scales, so that very high resolution AOGCMs with grid sizes of about 10 km could be seen as the best solution. However, until today the horizontal resolution is still relatively coarse (100 to 250 km grid size) due to limitations in computer resources. Hence AOGCMs cannot provide regional details.

To overcome the deficiency two different approaches are used: statistical downscaling and dynamical downscaling. Both translate the information from the global model to the region of interest. Statistical downscaling techniques connect the climate change signal provided by the GCM with observations from measurement stations in the region to achieve higher resolved climate change signals. Dynamical downscaling uses high resolution three-dimensional regional climate models (RCM), which are



Regional Climate Models: Linking Global Climate Change to Local Impacts, Figure 2  
Nesting technique

nested into GCMs (Fig. 2). RCMs are similar to numerical weather prediction models, in which non-linear processes in the climate system are taken into account. The results of both methods depend on both the quality of the global and regional models. In the following, the focus will be on dynamical downscaling, to be able to detect more easily new extremes, which have not been observed so far, and to take into account possible feedback mechanisms, which might appear under climate change conditions, like the snow-albedo/temperature feedback or the soil moisture-temperature feedback.

Regional climate models are limited area models. They only cover the area of interest, which can be resolved to a much higher degree than GCMs. Therefore RCMs describe the effects of regional and small scale processes within the simulation domain and are connected to the global flow using the nesting technique. This technique was developed to get higher resolution climate information on regional scales, and it is very similar to the nesting procedure in NWP. For initialization and at the lateral boundaries the GCM, in which the RCM is nested, provides information about the state of the atmosphere and the surface conditions. Usually atmospheric fields like wind, pressure, temperature and humidity are provided as well as sea surface temperatures [14,17]. Soil temperature and soil moisture are initialized once, but calculated within the RCMs during the simulation.

The development of regional climate models started in the USA. Filippo Giorgi at NCAR was the first one running the MM4 model in a so-called climate mode [10],

which means simulations longer than a few days, as it was common for numerical weather prediction (NWP). For many years RCMs were applied to simulations covering one month. The extension of NWP models to month-long simulations required changes in the formulation of physical processes which were taken into account within the model. In NWP models processes acting on time scales longer than weeks are not important and so not included.

The development of longer-term climate simulations happened very fast and simultaneously in several modeling centers of the world. In the early 1990s, the first multi-year simulations were carried out by Giorgi et al. [11,12], whereas Jones et al. [21,22] and McGregor et al. [27] succeeded in ten-year simulations. Nowadays regional climate simulations stretch from several decades, first achieved by Machenhauer et al. [25], up to more than a century in transient climate change mode [20]. Currently RCMs are widely used for regional climate studies for almost all regions of the world, with horizontal grid spacing ranging from more than 100 km to 10 km. A more detailed overview can be found in Giorgi [9] and for example in [7] focusing on regional climate modeling in the Arctic.

The basic features of RCMs will be explained in Sect. “**Basic Features and Model Characteristics**” and examples of applications will be presented in Sect. “**Validation**”, climate scenarios in Sect. “**IPCC-Scenarios**”, and examples of applications in Sects. “**Regional Climate Change**” and “**Regional Extremes**”. A discussion of future perspectives and concluding remarks follow in Sect. “**Future Directions**”.

## Basic Features and Model Characteristics

Until today, most RCMs are three-dimensional hydrostatic circulation models, solving the discretized primitive equations of the atmospheric motion. Summaries of model characteristics can be found in many publications, e. g. Jacob et al. [18,19]. As an example for the development and characteristics of many RCMs, the standard set-up of REMO, the regional climate model developed and used at the Max-Planck-Institute for Meteorology is described in more detail below.

The development of REMO started in 1994 utilizing the existing NWP model (EM) of the German Weather Service DWD [26]. Additionally, the physical parametrization package of the general circulation model ECHAM4 [33] has been implemented. During the last decade it could be shown in several applications that the combination of the EM dynamical core plus the ECHAM4 physical parametrization scheme is able to realistically reproduce regional climatic features and therefore became the standard setup.

The atmospheric prognostic variables of REMO are the horizontal wind components, surface pressure, temperature and specific humidity, as well as cloud liquid water. The temporal integration is accomplished by a leap-frog scheme with semi-implicit correction and time filtering after Asselin [2]. REMO is a grid box model, with grid box centers defined on a rotated latitude–longitude coordinate system. For horizontal discretization the model uses a spherical Arakawa-C grid in which all variables except the wind components are defined in the center of the respective grid box. In the vertical, a hybrid vertical coordinate system is applied [35]. Details about the physical parameterizations can be found in Jacob 2001 [16], but will not be explained here in more detail, since they vary slightly from RCM to RCM (see for example [18,19]).

The resolution of the horizontal grids in RCMs varies from about 100 km to 10 km and has increased constantly. For many years  $1/2^\circ$  grid size could be seen as a standard horizontal resolution, which was used in many experiments, even for model inter-comparison studies (e. g. [4,18,25,32]).

REMO uses horizontal grids with  $1/12^\circ$ ,  $1/6^\circ$  or  $1/2^\circ$ , corresponding to horizontal resolutions of about 10 km, 18 km and 55 km. In the vertical 20 to 40 levels are applied.

Applying the nesting technique for regional climate models requires large scale atmospheric flow fields to *drive* the RCMs at their lateral boundaries. These fields can be derived from different sources depending on the application. Regional climate simulations require climate change information from AOGCMs, whereas the simulations of

the last decades are driven by global analyzes of observations. The analyzes consist of observations, which have been interpolated in space and time using global models. They can be interpreted as the best available representation of the observed atmospheric flow conditions; however, systematic biases cannot be excluded due to the utilization of numerical models for interpolation. In regional climate modeling the use of driving data from analyzes or re-analyzes products are referred to as simulations with perfect boundary conditions (PBC). These experiments have the clear advantage to be directly comparable to observations for the actual time periods and they build the basis for model validation experiments.

In all cases, the relaxation scheme according to Davies [5] is applied in REMO, meaning that the prognostic variables are adjusted towards the large-scale forcing in a lateral sponge zone of 8 grid boxes. Within this zone the influence of the lateral boundary conditions decreases exponentially towards the inner model domain.

At the lower boundary, RCMs are determined through the interaction with the land surface and, over sea, by the sea surface temperature (SST) and sea ice distribution. The SST can either be interpolated from the large-scale forcing or from observational datasets, or it can be calculated online by a regional ocean model coupled to the RCM, e. g. [23]. The same is true for the sea ice extent, which can as a further option also be diagnosed from the SST. The land surface with its ongoing changes plays a major role in the climate system. Therefore, in all RCMs the exchange between surface and atmosphere is realized by the implementation of a land surface scheme [30]. Generally, one surface grid box can either be covered by water, sea ice or land or can include fractions of land and water areas, all characterized by their own roughness length and albedo. The land fraction of the surface can be covered by bare soil or by vegetation of different type. Depending on the complexity of the land surface scheme the exchange between the atmosphere and the underlying surface is realized through turbulent surface fluxes and the surface radiation flux, which are calculated separately for each fraction and weighted averages of the fluxes are used within the lowest atmospheric model level. Physical properties of the soil and vegetation control the exchange of heat, moisture and momentum over land. In REMO these properties include for example the surface roughness length, the soil field capacity, the water holding capacity of the vegetation, the background albedo, the fractional vegetation cover and the leaf area index (LAI). Some of these parameters strongly depend on the physiological state of the vegetation and are variable between the growing and the dormancy season [31].

There are two options to use the nesting technique. Within the one-way mode a GCM drives a RCM at the lateral boundaries, but no information is given back to the GCM. This method is the standard one used until today in regional climate modeling. It is relatively easy to implement and allows the use of RCMs without running a GCM. The RCM adds information on scales smaller than the driving GCM (e. g. topographical forcing), but is strongly dependent on the superimposed large scale flow. Hence RCMs cannot correct large scale flows originating from GCMs, which might have large errors. However, Giorgi et al. [13], showed that some modulation of the large scale flow is possible within the RCM simulation, stimulated by regional scale forcing.

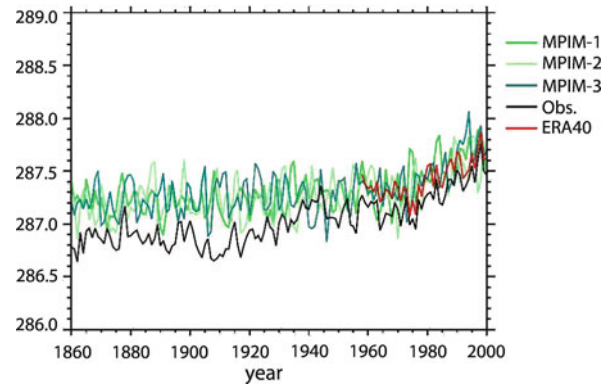
In the two-way mode, both models GCM and RCM run simultaneously and the RCM feeds back information to the GCM every GCM time step. This method has recently been established for regional climate modeling studies [24]; it has the clear advantage that the atmospheric flow generated within the RCM domain can modulate the large scale flow in areas with strong energetic input from the surface to the atmosphere (e. g. the maritime continent).

If RCM experiments are carried out with high horizontal resolution, like 10 or 20 km, it can be required to use the so-called double nesting technique to avoid mismatch in scales along the lateral boundaries due to the coarse resolution of the driving GCMs. Double nesting means that first a RCM simulation will be carried out with a relatively coarse resolution to generate lateral boundary conditions for further nesting. For REMO, sometimes a sequence of nests is calculated [20].

Finally there are two independent options to run a regional climate model with PBC: the forecast mode and the climate mode. In climate mode the RCM simulation is initialized once from analyzes, and then it is continuously calculated forward in time, driven by regularly up-dated lateral boundaries. In forecast mode, a sequence of short runs (e. g. 30 hours), each initialized every 30 hours from analyzes, is carried out. The forecast mode has the advantage to force the RCM flow to be very close to the observed one, but it has the disadvantage to suppress mesoscale flow features. These mesoscale processes can be excited within the RCM domain by land–sea contrasts or topography and are too small to be taken into account in the GCM.

## Validation

The quality of the RCM simulations depends strongly on the performance of the driving model due to the one-way nesting procedure. Therefore, it is extremely important



Regional Climate Models: Linking Global Climate Change to Local Impacts, Figure 3

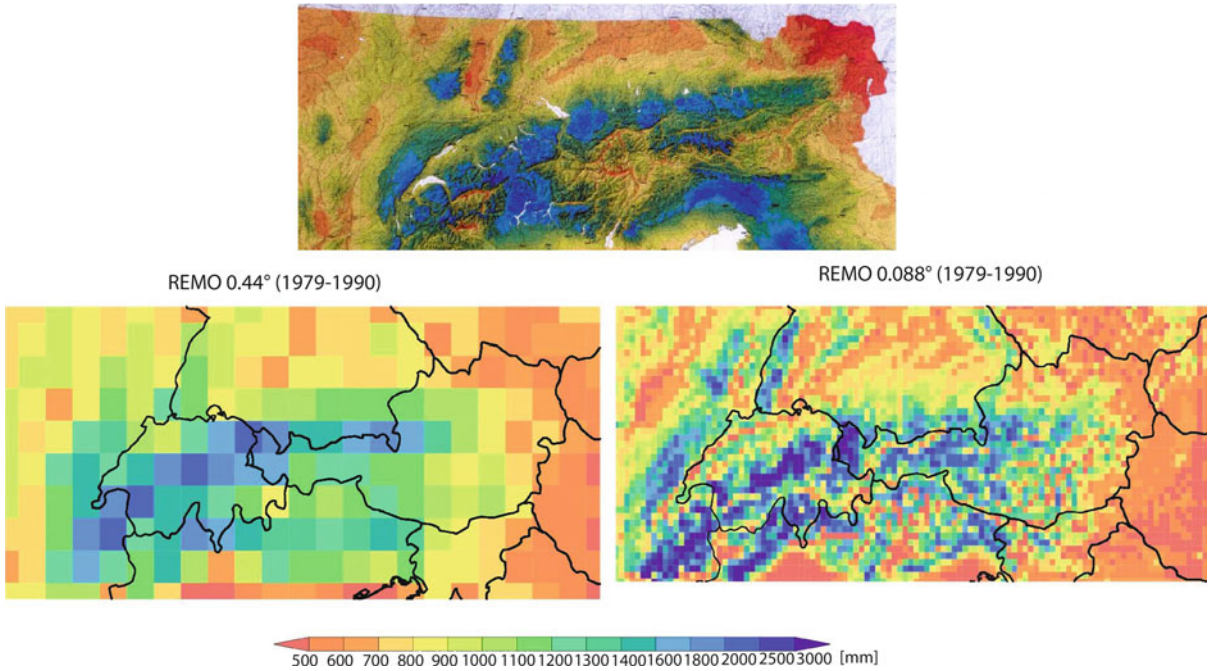
**Time series of the global mean near surface air temperature (K): observed (black), from reanalyzes (red), and from the MPI global climate model simulations (green)**

to validate the driving large scale fields before applying RCMs. The model quality, however, can only be judged in comparison with independent observations. Therefore, time periods of the past are simulated and the model results are compared against measurements before the models are used for climate change studies. These comparisons are also part of model development and testing.

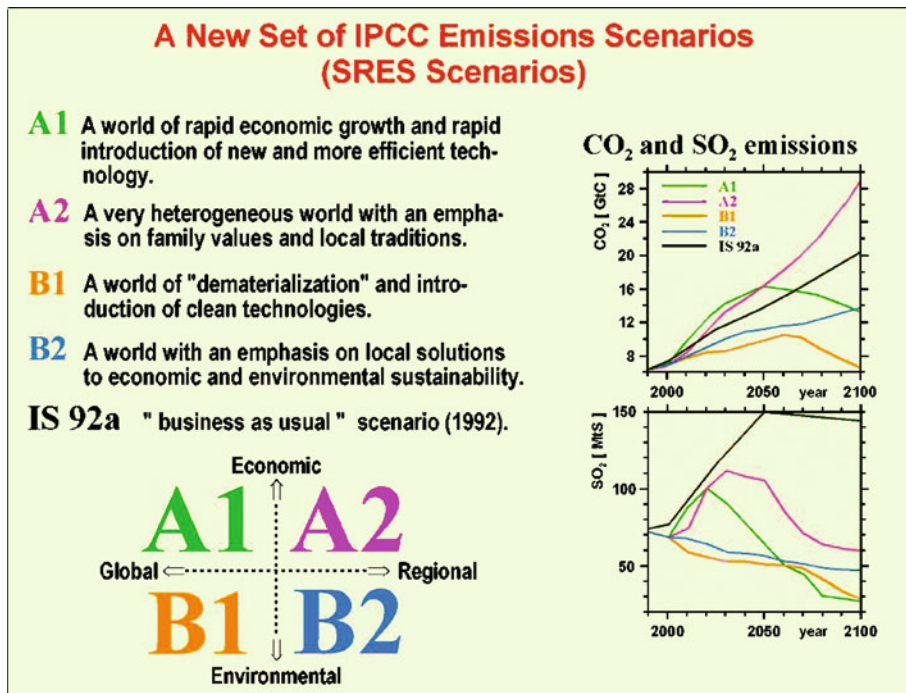
As an example, Fig. 3 shows time series of observed and simulated global mean near surface temperatures for the period 1860 to 2000. The simulated results from the global coupled climate model ECHAM5/MPI-OM (Max-Planck-Institute for Meteorology) are in good agreement with ERA40 data, but about  $0.5^\circ$  warmer than the reconstructed observations. The observed increase during the last decades is clearly visible.

As for GCMs, the model quality of RCMs needs to be analyzed. Therefore RCMs are nested into re-analyzed data, which can be seen as close to reality as possible (see above). The results of the RCM simulations of the last decades are compared against independent observations, means as well as extremes are considered. As an example, simulated precipitation climatologies calculated with REMO with two different horizontal grid sizes are compared against observations [8].

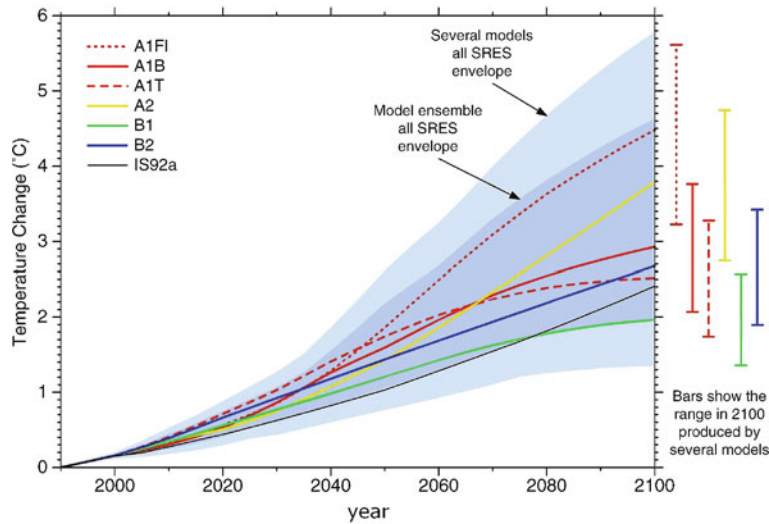
The total amount of precipitation and the horizontal pattern are much better resolved using the very high horizontal resolution of about 10 km (Fig. 4). Regional maxima, like the one in the Black Forest, and minima, like in the central valleys of the Alps, are detectable. However, the resolution is still too coarse for climate change studies in individual alpine valleys. The 50 km grid is much finer than standard GCM grids (about 150 to 250 km), but is still



Regional Climate Models: Linking Global Climate Change to Local Impacts, Figure 4  
 Annual total precipitation (mm), observed (1971–1990, upper panel) and simulated with about 50 km grid size (left) and 10 km grid size (right)



Regional Climate Models: Linking Global Climate Change to Local Impacts, Figure 5  
 SRES Scenarios, which shows the four major storylines together with the associated developments of CO<sub>2</sub> and SO<sub>2</sub> emissions from 2000 until 2100



Regional Climate Models: Linking Global Climate Change to Local Impacts, Figure 6  
Changes in global mean near surface air temperature as calculated by several GCMs under seven emissions scenarios until 2100

insufficient for studying regional details if the regions are too small.

### IPCC-Scenarios

The investigation of possible future climate changes requires information about possible changes in the *drivers* of climate change. So-called *drivers* are for example, amount and distribution of aerosols and green house gases (GHG) in the atmosphere, which depend directly on natural and man-made emissions. The IPCC emissions scenarios (Fig. 5) follow so-called story lines, describing possible developments of the socioeconomic system [29].

The emissions are directly used within GCMs and RCMs and they initiate changes in global and regional climates through numerous non-linear feedback mechanisms. As an example, Fig. 6 shows possible developments of global mean near surface temperatures calculated by several models for different scenarios.

The global mean changes in near surface temperature until 2050 is about 1.5°C, whereas until the end of the century a wide spread appears from 1.5°C to 5.5°C.

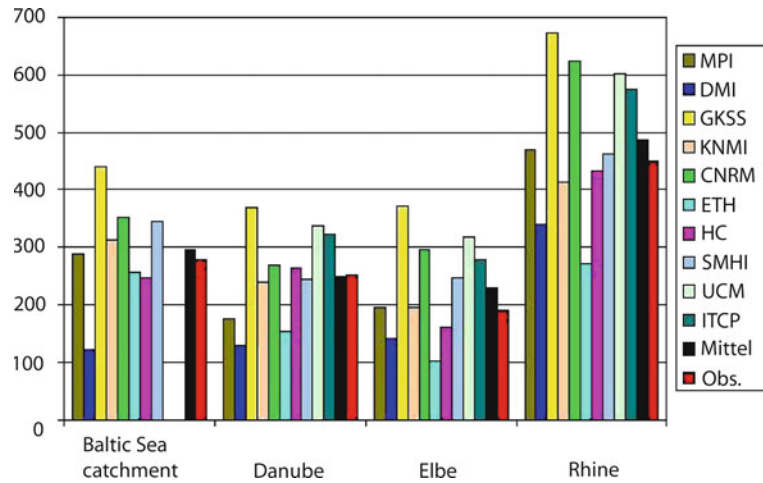
### Regional Climate Change

In order to achieve information about the probability, e. g. for the intensification of the hydrological cycle over Europe, several models from different European climate research institutes are used, as it was done in the EU project PRUDENCE [4].

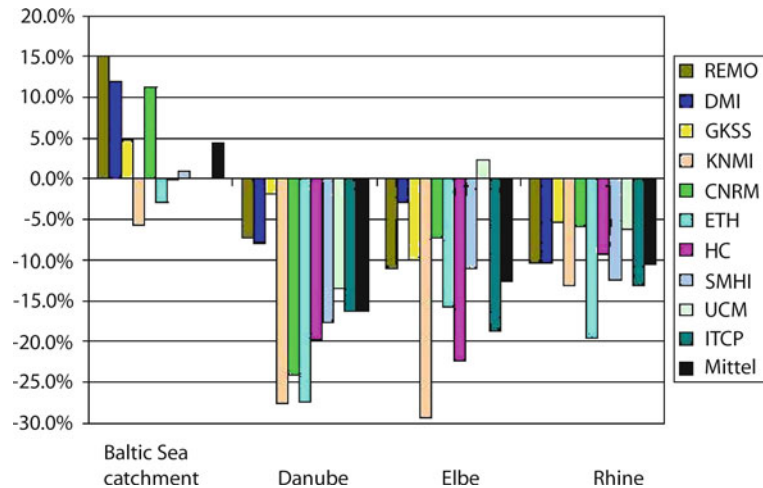
Following the climate change scenario A2 projecting a strong future increase of greenhouse gases until the year 2100 (IPCC, 2001) and a subsequent global mean temperature increase of about 3.5°, numerous simulations were conducted within PRUDENCE [19]. An analysis of their results for different river catchments [15] shows significant differences between the projected changes over northern and central Europe for the time period 2070–2100 compared to the current climate (1961–1990, Fig. 7).

For the Baltic Sea catchment, a precipitation increase of about +10% for the annual mean is projected, with the largest increase of up to +40% in winter, while a slight reduction of precipitation is calculated for the late summer. Evapotranspiration will increase during the entire year with a maximum increase in winter. These rises in precipitation and evapotranspiration would lead to an increase of river discharge into the Baltic Sea of more than 20% in winter and early spring. Here, the seasonal distribution of discharge is largely influenced by the onset of spring snowmelt.

For the catchments of Rhine, Elbe and Danube, a different change in the water balance components is projected. While the annual mean precipitation will remain almost unchanged, it will increase in late winter (January–March) and decrease significantly in summer. The evapotranspiration will rise during the entire year, except for the summer, with a maximum increase in winter. These changes lead to a large reduction of 10 to 20% in the annual mean discharge (Fig. 8). Especially for the Danube,



Regional Climate Models: Linking Global Climate Change to Local Impacts, Figure 7  
 Simulated and observed river run-off (precipitation P – evaporation E) for 1961 to 1990 [15] in the Baltic Sea, Danube, Elbe and Rhine catchments



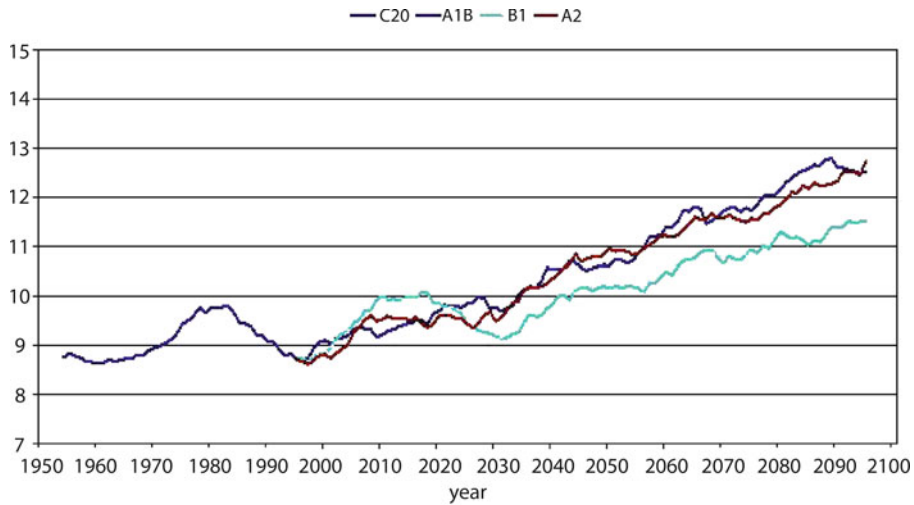
Regional Climate Models: Linking Global Climate Change to Local Impacts, Figure 8  
 Simulated and observed change in river run-off (precipitation P – evaporation E) for the period 2071 to 2100 compared to 1961 to 1990 [15]. Baltic Sea, Danube, Elbe and Rhine catchments

the projected summer drying has a strong impact on the discharge that is reduced up to 20% throughout the year except for the late winter (February/March) when the increased winter precipitation causes a discharge increase of about 10%. These projected changes in the mean discharge will have significant impacts on water availability and usability in the affected regions.

Under climate change conditions not only the absolute amounts of precipitation may change but also the precipitation intensities, i. e. the amount of precipitation within a certain time period. The simulation of precipitation intensities or extreme precipitation events requires however a considerably higher resolution than the A2 results pre-

sented above so that for example the influence of the topography of the Alps on the formation of precipitation over the Rhine catchment could be adequately calculated. High resolution RCM results show that the global warming until 2050 will lead to an increase of high precipitation events over the Alpine part of the Rhine catchment, especially in summer. This climate change signal becomes clearly visible in the Pre-Alps, but a similar trend is also seen in the high resolution simulations over large parts of Europe.

A major breakthrough was possible with the regional climate change simulations on 10 km grid scale. Within a co-operation with the national environmental



Regional Climate Models: Linking Global Climate Change to Local Impacts, Figure 9  
Changes in annual mean near surface temperature (°C) from 1950 to 2100 for three different IPCC SRES scenarios

agency (UBA), REMO was used for a control simulation from 1950 to 2000 and three transient runs for the IPCC SRES scenarios A2, A1B and B1. The simulation domain covers Germany, Austria and Switzerland [20]. As an example the most important results for Germany at the end of this century are summarized as follows:

The simulated annual mean near surface temperature is increasing up to 3.5°C depending on the emission scenario (Fig. 9). The regional pattern of temperature changes shows that the south and southeast warm more than all other areas in the simulation domain. The warming is associated with a decrease of precipitation amount in wide areas of Germany during summer and an increase of precipitation in south and southwest regions during the winter (Fig. 10). The winter precipitation is mostly rain and less precipitation falls as snow.

### Regional Extremes

The calculated rapid and strong changes of climate parameters can have severe impacts on humans and the environment. As an example, REMO results for the Rhine basin are presented for a B2 scenario until 2050. Between 1960 and 2050 the near surface temperature will rise by about 3°C and the number of summer days and hot days will increase (Fig. 11). In addition, the number of periods with summer days, this is the period of consecutive days with a daily maximum temperature above 25°C (not shown), will be higher in the future decades. Winter temperature also increases, leading to a decrease in frost and ice days.

The investigation of probability distribution functions for temperature and precipitation using the 10 km horizontal resolutions simulations for Germany shows possible monthly mean temperature of more than 30°C for July and >10°C for January appearing in the Rhine valley under the assumption of A1B scenario until the end of this century. In addition, possible increases in monthly mean precipitations are projected for A1B until 2100 in the area of Leipzig (Elbe drainage basin) for January as well as for July (Bülow, Ph D thesis, in preparation).

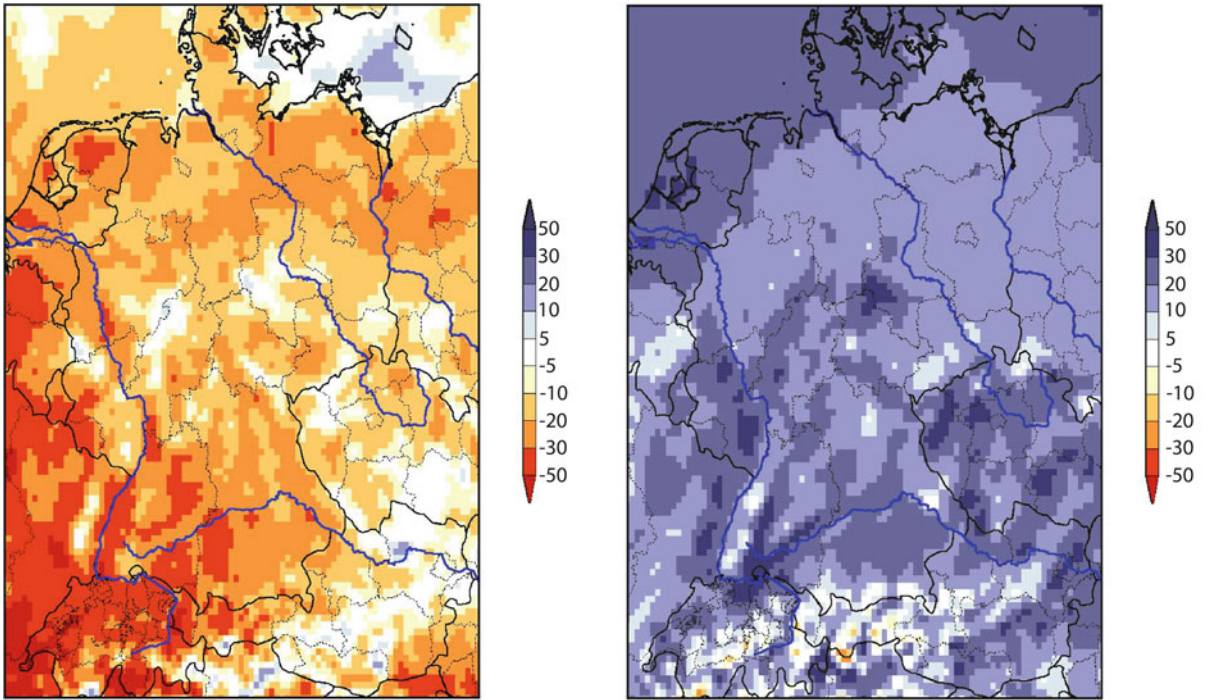
The projected changes in extremes are coherent with findings of Schär et al. [34] who studied the occurrence of summer heat waves in Europe today and in the future. He showed that the summer 2003 heat wave was extraordinary but can appear much more often in the future.

Another robust finding is the increase of heavy precipitation events in summer, which goes together with a decrease in monthly mean summer precipitation in central Europe [3]. Such short-term, strong convective summer precipitation events have the potential of causing damages, e. g. for agriculture, but also in cities, when sewage systems might be flooded.

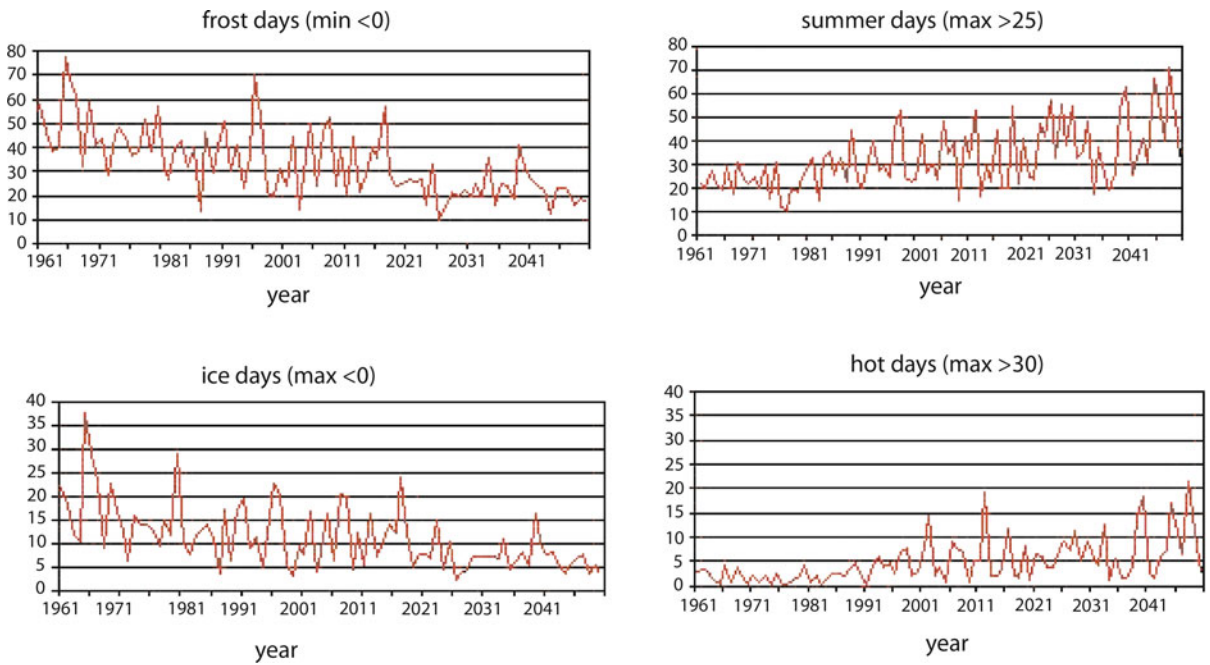
### Future Directions

Regional climate modeling has made major progress during the last decades and it could be shown that the added value lies mainly in the provision of details in space and time, which is demanded for impact assessments as well as by the public. RCMs are ready for operational use as powerful tools to simulate regional climatic features and their





Regional Climate Models: Linking Global Climate Change to Local Impacts, Figure 10  
 Climate change signals for summer (*left*) and winter (*right*) precipitation (%) in scenario A1B for 2071 to 2100 compared to 1961 to 1990



Regional Climate Models: Linking Global Climate Change to Local Impacts, Figure 11  
 REMO B2 scenario for the Rhine catchment: frost days (*upper left*), ice days (*lower left*), summer days (*upper right*) and hot days (*lower right*)

changes in all regions of the world, for time periods from today until the end of this century.

Now, further research on regional climate changes should focus on the reliability in projected regional and local climate change pattern. This can be done by ensemble calculations of GCMs-RCMs, which are also needed for the establishment of probability distribution functions to analyze extreme situations. Here special focus will be on the use of RCMs for regional climate prediction for the next 10 to 30 years, for which it is unclear if weather and climate extremes may change, where it might happen and if regions will face extreme situations in the future, in which this never happened before. The dissemination of regional climate “change information” can efficiently be done with the establishment of climate information systems in individual regions of the World.

The use of RCMs as intelligent interpolators of observed data in data sparse regions need to be proven, but RCMs have a great potential for this. In addition, the added value of RCMs compared to GCMs needs to be proven by applying the two-way nesting technique. In mountainous regions like the Alps an impact on the large scale flow can be expected.

The extension of regional climate models (RCM) to regional system models (RSM) is a major challenge for the next years. The coupling to regional ocean models, land use and hydrological models has recently started. Along with this the carbon and nitrogen cycles will be implemented on a regional scale. This allows a much better simulation of additional regional feedbacks as mentioned in Sect. “**Definition of the Subject**”, which might have the potential to modulate the regional climate signal projected by GCMs for the region of interest.

Finally, it must be stated that along with improvements of RCMs, the development of GCMs has to be continued. Their performance in individual regions can be quite poor and an improvement is urgently needed. Here RCMs can deliver important information about regional climatic details and help advancing global climate change simulations.

## Acknowledgments

I like to thank the REMO-grooup at the Max Planck Institute for Meteorology for their enthusiastic support.

## Bibliography

- Arnell NM (1996) Global warming, river flows and water resources. Wiley, Chichester
- Asselin R (1972) Frequency filter for time integrations. *Mon Weather Rev* 100:487–490
- Christensen JH, Christensen OB (2004) Climate modelling: severe summer flooding in Europe. *Nature* 421:805–806
- Christensen JH, Christensen OB (2007) A summary of the PRUDENCE model projections of changes in European climate by the end of this century. *Clim Chan* 81(1):7–30
- Davies HC (1976) A later boundary formulation for multi-level prediction models. *Quart J R Meteorol Soc* 102:405–418
- Dethloff K, Abegg C, Rinke A, Hebestad I, Romanov V (2001) Sensitivity of Arctic climate simulations to different boundary layer parameterizations in a regional climate model. *Tellus* 53(A):1–26
- Dethloff K, Rinke A, Lynch A, Dorn W, Saha S, Handorf D (2008) Arctic climate change – The ACSYS decade and beyond. Chapter 8: Arctic regional climate models (in press)
- Frei C, Christensen JH, Deque M, Jacob D, Jones RG, Vidale PL (2003) Daily precipitation statistics in regional climate models: Evaluation and intercomparison for the European Alps. *J Geophys Res* 108(D3):4124. doi: [10.1029/2002JD002287](https://doi.org/10.1029/2002JD002287)
- Giorgi F (2006) Regional climate modeling: Status and perspectives. *J Phys IV France* 139(2006):101–118. doi: [10.1051/jp4:2006139008](https://doi.org/10.1051/jp4:2006139008)
- Giorgi F, Bates GT (1989) The climatological skill of a regional model over complex terrain. *Mon Weather Rev* 117:2325–2347
- Giorgi F, Bates GT, Niemann SJ (1993) The multi-year surface climatology of a regional atmospheric model over the western United States. *J Clim* 6:75–95
- Giorgi F, Brodeur CS, Bates GT (1994) Regional climate change scenarios over the United States produced with a nested regional climate model. *J Clim* 7:375–399
- Giorgi F, Mearns LO, Shields C, McDaniel L (1998) Regional nested model simulations of present day and 2XCO2 climate over the Central Plains of the US. *Clim Chan* 40:457–493
- Giorgi F, Mearns LO (1999) Introduction to special section: Regional climate modelling revisited. *J Geophys Res* 104:6335–6352
- Hagemann S, Jacob D (2007) Gradient in the climate change signal of European discharge predicted by a multi-model ensemble. *Clim Chan* 81(1):309–327
- Jacob D (2001) A note to the simulation of the annual and inter-annual variability of the water budget over the Baltic Sea drainage basin. *Meteorol Atmos Phys* 77:61–73
- Jacob D, Podzun R (1997) Sensitivity studies with the regional climate model REMO. *Meteorol Atmos Phys* 63:119–129
- Jacob D, Van den Hurk BJJM, Andrae U, Elgered G, Fortelius C, Graham LP, Jackson SD, Karstens U, Köpken C, Lindau R, Podzun R, Rockel B, Rubel F, Sass BH, Smith RNB, Yang X (2001) A comprehensive model inter-comparison study investigating the water budget during the BALTEX-PIDCAP period. *Meteorol Atmos Phys* 77:19–43
- Jacob D, Bärring L, Christensen OB, Christensen JH, Hagemann S, Hirschi M, Kjellström E, Lenderink G, Rockel B, Schär C, Seneviratne SI, Somot S, van Ulden A, van den Hurk B (2007) An inter-comparison of regional climate models for Europe: Design of the experiments and model performance. *Clim Chan* 81(1):31–52
- Jacob D, Göttel H, Kotlarski S, Lorenz P, Sieck K (2008) Klimaauswirkungen und Anpassung in Deutschland – Phase 1: Erstellung regionaler Klimaszenarien für Deutschland. Abschlussbericht zum UFOPLAN-Vorhaben 204 41 138, Berichtszeitraum: 1. Oktober 2004 bis 30. September 2007. Max-Planck-Institut für Meteorologie (MPI-M), Hamburg

21. Jones RG, Murphy JM, Noguer M (1995) Simulations of climate change over Europe using a nested regional climate model. I: Assessment of control climate, including sensitivity to location of lateral boundaries. *Quart J R Meteorol Soc* 121:1413–1449
22. Jones RG, Murphy JM, Noguer M, Keen AB (1997) Simulation of climate change over Europe using a nested regional climate model. II: Comparison of driving and regional model responses to a doubling of carbon dioxide. *Quart J R Meteorol Soc* 123:265–292
23. Lehmann A, Lorenz P, Jacob D (2004) Modelling the exceptional Baltic Sea flow events in 2002–2003. *Geophys Res Lett* 31:L21308. doi: [0.1029/2004GL020830](https://doi.org/10.1029/2004GL020830)
24. Lorenz P, Jacob D (2005) Influence of regional scale information on the global circulation: A two-way nesting climate simulation. *Geophys Res Lett* 32:L18706. doi: [0.1029/2005GL023351](https://doi.org/10.1029/2005GL023351)
25. Machenhauer B, Windelband M, Botzet M, Christensen JH, Déqué M, Jones RG, Ruti PM, Visconti G (1998) Validation and analysis of regional present-day climate and climate change simulations over Europe. MPI Report No. 275. MPI, Hamburg
26. Majewski D (1991) The Europa-Modell of the Deutscher Wetterdienst. In: ECMWF seminar on numerical methods in atmospheric models, vol 2. ECMWF, Reading
27. McGregor JL, Katzfey JJ, Nguyen KC (1995) Seasonally varying nested climate simulations over the Australian region. 3rd Int Conf Model Glob Clim Chan Var. Hamburg, Germany, 4–8 Sept 1995
28. McGregor JL, Katzfey JJ, Nguyen KC (1999) Recent regional climate modelling experiments at CISRO. In: Ritchie H (ed) Research activities in atmospheric and oceanic modelling. CAS/JSC Working Group on Numerical Experimentation Report 28. WMO/TD – no. 942. WMO, Geneva, pp 7.37–7.38
29. Nakicenovic N, Alcamo J, Davis G, de Vries B, Fenhann J, Gaffin S, Gregory K, Grübler A, Jung TY, Kram T, La Rovere EL, Michaelis L, Mori S, Morita T, Pepper W, Pitcher H, Price L, Raihi K, Roehrl A, Rogner HH, Sankovski A, Schlesinger M, Shukla P, Smith S, Swart R, van Rooijen S, Victor N, Dadi Z (2000) IPCC special report on emissions scenarios. Cambridge University Press, Cambridge
30. Pitman A (2003) Review: The evolution of, and revolution in, land surface schemes designed for climate models. *Int J Climatol* 23:479–510
31. Rechid D, Jacob D (2006) Influence of seasonally varying vegetation on the simulated climate in Europe. *Meteorol Z* 15:99–116
32. Rinke A, Marbaix P, Dethloff K (2004) Internal variability in Arctic regional climate simulations: Case study for the SHEBA year. *Clim Res* 27:197–209
33. Roeckner E, Arpe K, Bengtsson L, Christoph M, Claussen M, Dümenil L, Esch M, Giorgetta M, Schlese U, Schulzweida U (1996) The atmospheric general simulation model ECHMA-4: Model description and simulation of present-day climate. Report 218. Max Planck Institute for Meteorology, Hamburg
34. Schär C, Vidale PL, Lüthi D, Frei C, Häberli C, Liniger MA, Appenzeller C (2004) The role of increasing temperature variability in European summer heatwaves. *Nature* 427:332–336
35. Simmons AJ, Burridge DM (1981) An energy and angular-momentum conserving vertical finite-difference scheme and hybrid vertical coordinate. *Mon Weather Rev* 109:758–766

## Seismic Wave Propagation in Media with Complex Geometries, Simulation of

HEINER IGEL<sup>1</sup>, MARTIN KÄSER<sup>1</sup>, MARCO STUPAZZINI<sup>2</sup>

<sup>1</sup> Department of Earth and Environmental Sciences, Ludwig-Maximilians-University, Munich, Germany

<sup>2</sup> Department of Structural Engineering, Politecnico di Milano, Milano, Italy

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[The Evolution of Numerical Methods and Grids](#)

[3D Wave Propagation on Hexahedral Grids:](#)

[Soil-Structure Interactions](#)

[3D Wave Propagation on Tetrahedral Grids:](#)

[Application to Volcanology](#)

[Local Time Stepping:  \$\Delta t\$ -Adaptation](#)

[Discussion and Future Directions](#)

[Acknowledgments](#)

[Bibliography](#)

### Glossary

**Numerical methods** Processes in nature are often described by partial differential equations. Finding solutions to those equations is at the heart of many studies aiming at the explanation of observed data. Simulations of realistic physical processes requires generally the use of numerical methods – a special branch of applied mathematics – that approximate the partial differential equations and allows solving them on computers. Examples are the finite-difference, finite-element, or finite-volume methods.

**Spectral elements** The spectral element method is an extension of the finite element method that makes use of specific basis functions describing the solutions inside each element. These basis functions (e. g., Chebyshev or Legendre polynomials) allow the interpolation of functions exactly at certain collocation points. This is often termed spectral accuracy.

**Discontinuous Galerkin method** The discontinuous Galerkin method is a flavor of the finite-element method that allows discontinuous behavior of the spatial or temporal fields at the element boundaries. The discontinuities – that might be small in the case of continuous physical fields such as seismic waves – then define so-called Riemann problems that can be handled using

the concepts from finite-volume techniques. Therefore, the approximate solution is updated via numerical fluxes across the element boundaries.

**Parallel algorithms** All modern supercomputers make use of parallel architectures. This means that a large number of processors are performing (different) tasks on different data at the same time. Numerical algorithms need to be adapted to these hardware architectures by using specific programming paradigms (e. g., the message passing interface MPI). The computational efficiency of such algorithms strongly depends on the specific parallel nature of the problem to be solved, and the requirement for inter-processor communication.

**Grid generation** Most numerical methods are based on the calculation of the solutions at a large set of points (grids) that are either static or depend on time (adaptive grids). These grids often need to be adapted to the specific geometrical properties of the objects to be modeled (volcano, reservoir, globe). Grids may be designed to follow domain boundaries and internal surfaces. Before specific numerical solvers are employed the grid points are usually connected to form triangles or rectangles in 2D or hexahedra or tetrahedra in 3D.

### Definition of the Subject

Seismology is the science that aims at understanding the Earth's interior and its seismic sources from measurements of vibrations of the solid Earth. The resulting images of the physical properties of internal structures and the spatio-temporal behavior of earthquake rupture processes are prerequisites to understanding the dynamic evolution of our planet and the physics of earthquakes. One of the key ingredients to obtain these images is the calculation of synthetic (or theoretical) seismograms for given earthquake sources and internal structures. These synthetic seismograms can then be compared quantitatively with observations and acceptable models be searched for using the theory of inverse problems. The methodologies to calculate synthetic seismograms have evolved dramatically over the past decades in parallel with the evolution of computational resources and the ever increasing volumes of permanent seismic observations in global and regional seismic networks, volcano monitoring networks, and experimental campaigns. Today it is a tremendous challenge to extract an optimal amount of information from seismograms. The imaging process is still primarily carried out using ray theory or extensions thereof not fully taking into account the complex scattering processes that are occurring in nature.

To model seismic observations in their full complexity we need to be able to simulate wave propagation through 3D structures with constitutive relations that account for anisotropic elasticity, attenuation, porous media as well as complex internal interfaces such as layer boundaries or fault systems. This implies that numerical methods have to be employed that solve the underlying partial differential equations on computational grids. The high-frequency oscillatory nature of seismic wave fields makes this an expensive endeavor as far as computational resources are concerned. As seismic waves are propagating hundreds of wavelengths through scattering media, the required accuracy of the numerical approximations has to be of the highest possible order. Despite the fact that the physics of wave propagation is well understood, only recently computational algorithms are becoming available that allow us to accurately simulate wave propagation on many scales such as reservoirs, volcanoes, sedimentary basins, continents, and whole planets.

In addition to the imaging problem for subsurface structure and earthquake sources, the possibilities for 3D wave simulations have opened a new route to forecasting strong ground motions following large earthquakes in seismically active regions. In the absence of any hope to deterministically predict earthquakes, the calculation of earthquake scenarios in regions with sufficiently well known crustal structures and fault locations will play an important role in mitigating damage particularly due to potentially amplifying local velocity structures. However, to be able to employ the advanced 3D simulation technology in an efficient way, and to make use of the fast advance of supercomputing infrastructure, a paradigm shift in the concept of wave simulation software is necessary: The Earth science community has to build soft infrastructures that enable massive use of those simulation tools on the available high-performance computing infrastructure.

In this paper we want to present the state of the art of computational wave propagation and point to necessary developments in the coming years, particularly in connection with finding efficient ways to generate computational grids for models with complex topography, faults, and the combined simulation of soil and structures.

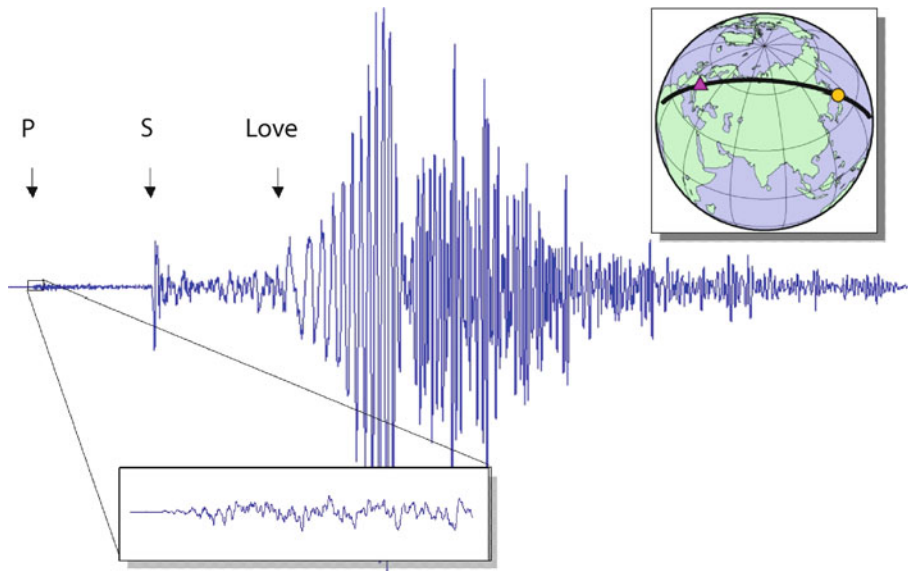
## Introduction

We first illustrate the evolution of methodologies to calculate and model aspects of seismic observations for the case of global wave propagation. Seismology can look back at almost 50 years of systematic observations of earthquake induced teleseismic ground motions with the standardized

global seismic and regional networks. The digital revolution in the past decades has altered the recording culture such that now seismometers are recording ground motions permanently rather than in trigger-mode, observations are becoming available in near-real time, and – because of the required sampling rates – the daily amount of observations automatically sent to the data centers is gigantic. If we take a qualitative look at a seismic observation (Fig. 1) we can illustrate what it takes to model either part or the whole information contained in such physical measurements.

In Fig. 1 a seismogram observed using a broadband seismometer (station WET in Germany) is shown. Globally observed seismograms following large earthquakes contain frequencies up to 1 Hz (P-wave motions) down to periods of around one hour (eigenmodes of the Earth) in which case modeling is carried out in the frequency domain. Seismograms of the kind shown in Fig. 1 contain many types of information. For large earthquakes the first part of the seismogram (inlet) contains valuable information on the spatio-temporal evolution of the earthquake rupture on a finite-size fault. A model of the fault slip history is a prerequisite to model the complete wave form of seismograms as the whole seismogram is affected by it unless severe low-pass filtering is applied. Information on the global seismic velocity structure is contained in the arrival times of numerous body-wave phases (here only P- and S-wave arrivals are indicated) and in the dispersive behavior of the surface waves (here the onset of the low-frequency Love waves is indicated). Further information is contained in the characteristics of the coda to body wave phases indicative of scattering in various parts of the Earth (see [62] for an account of modern observational seismology).

Adding a temporal and spatial scale to the above qualitative discussion reveals some important insight what it takes to simulate wave propagation on a planetary scale using grid-based numerical methods. Given the maximum frequency of around 1 Hz (P-waves) and 0.2 Hz (S-waves) the minimum wavelength in the Earth is expected to be  $O(\text{km})$ , requiring  $O(100 \text{ m})$  type grid spacing at least in the crustal part of the Earth leading to  $O(10^{12})$  necessary grid points (or volume elements) for accurate numerical simulations. This would lead to memory requirements  $O(100 \text{ TByte})$  that are today possible on some of the world's largest supercomputers. The message here is that despite the rapid evolution of computational power, the complete modeling of teleseismic observations using approaches such as spectral elements (e. g., [63,64]) requiring tremendous numbers of calculations to constrain structure and sources will remain a grand challenge for some



Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 1

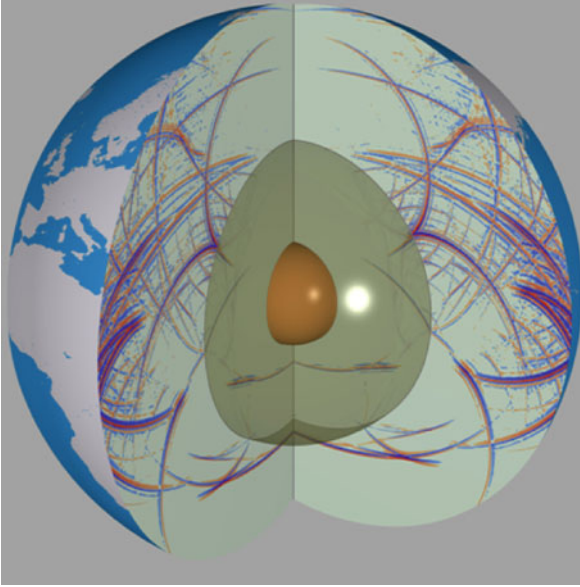
Transverse velocity seismogram of the M8.3 Tokachi-Oki earthquake near Hokkaido observed at station WET in Germany with a broadband seismometer. The total seismogram length is one hour. Arrival times of body wave phases (P, S) and the onset of transversely polarized surface (Love) waves are indicated

time to come. However, in many cases it is not necessary or not even desirable to simulate or model the whole seismogram, i. e. the complete observed frequency band. If we lower the cutoff frequency to 0.1 Hz (period 10 s), the required memory drops down to  $O(100 \text{ GByte})$ . Such calculations can be done today on PC-clusters that can be inexpensively assembled and run on an institutional level (e. g., [8]). In addition, it means that the massive use of such forward simulations for imaging purposes and phenomenological investigations of wavefield effects is around the corner. This does not only apply to wave propagation or imaging on a planetary scale but in the same way to problems in volcanology, regional seismology, and exploration geophysics.

An illustration of global wave simulations using the finite difference method (e. g., [14,54,55,58,109,110,114]) is shown in Fig. 2 (more details on the methodologies are given in Sect. “The Evolution of Numerical Methods and Grids”). The snapshot of the radial component of motion at a time when the direct P-wave has almost crossed the Earth reveals the tremendous complexity the wave field exhibits even in the case of a spherically symmetric Earth model (PREM, [37]). The wavefield with a dominant period of ca. 15 seconds also highlights the short wavelengths that need to be propagated over very large distances. This is the special requirement for computational wave propagation that is quite different in other fields of compu-

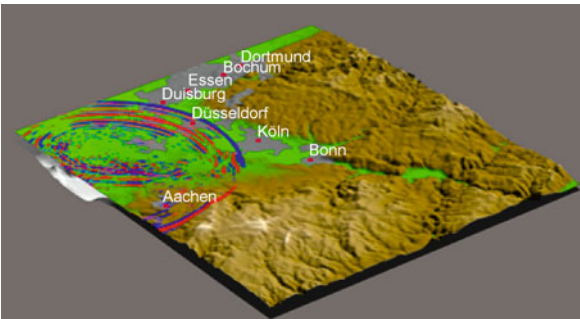
tational Earth Sciences. While the theory of linear elastic wave propagation is well understood and most numerical methods have been applied to it in various forms, the accuracy requirements are so high that – particularly when models with complex geometrical features need to be modeled – there are still open questions as to what works best. One of the main goals of this paper is to highlight the need to focus on the grid generation process for various types of computational grid cells (e. g., rectangular, triangular in 2D, and hexahedral and tetrahedral in 3D) and the interface to appropriate highly accurate solvers for wave propagation problems.

As mentioned above computational modeling of strong ground motions following large earthquakes (see Fig. 3 for an illustration) is expected to play an increasingly important role in producing realistic estimates of shaking hazard. There are several problems that are currently unsolved: (1) to achieve frequencies that are interesting for earthquake engineers in connection with structural damage the near surface velocity structure needs to be known and frequencies beyond 5 Hz need to be calculated. In most cases this structure is not well known (on top of the uncertainties of the lower basin structures) and the required frequencies demand extremely large computational models. (2) In addition to structural uncertainties, there are strong dependencies on the particular earthquake rupture process that influence the observed ground



Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 2

Snapshot of wave propagation inside the Earth approx. 25 minutes after an earthquake occurs at the top part of the model. The radial component of motion is shown (blue and red denote positive and negative velocity, resp.). The simulation was carried out using an axi-symmetric approximation to the wave equation [55,58] and high-order finite-differences. Motion is allowed in the radial and horizontal directions. This corresponds to the P-SV case in 2D cartesian calculations. Therefore the wavefield contains both P- and S-waves and phase conversions



Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 3

Snapshot (horizontal component) for a simulation of the M5.9 Roermond earthquake in the Cologne Basin in 1992 [38]. The 3D sedimentary basin (maximum depth 2 km) leads to strong amplification and prolongation of the shaking duration that correlates well with basin depth. Systematic calculations may help mitigating earthquake induced damage

motions. This suggests that many 3D calculations should be carried out for any characteristic earthquake of interest, to account for such variations (e. g., Wang et al. 2006). (3)

The large velocity variations (e. g., 300 m/s up to 8 km/s) require locally varying grid densities which is difficult to achieve with some of the classical numerical methods in use (e. g. finite differences). Some of the potential routes are developed below.

In summary, computational simulation of 3D wave propagation will be more and more a central tool for seismology with application in imaging problems, earthquake rupture problems, questions of shaking hazard, volcano seismology and planetary seismology. In the following we briefly review the history of the application of numerical methods to wave propagation problems and the evolution of computational grids. The increasing complexity of models in terms of geometrical features and range of physical properties imposes the use of novel methodologies that go far beyond the initial approximations based on finite differences.

### The Evolution of Numerical Methods and Grids

In this section we give a brief history of the application of numerical methods to the problem of seismic wave propagation. Such a review can not be complete, certainly gives a limited perspective, and only some key references are given. One of the points we would like to highlight is the evolution of the computational grids that are being employed for wave propagation problems and the consequences on the numerical methods of choice now and in the future.

Why do we need numerical approximations to elastic wave propagation problems at all? It is remarkable what we learned about the Earth without them! In the first decades in seismology, modeling of seismic observations was restricted to the calculation of ray-theoretical travel times in spherically symmetric Earth models (e. g., [13,16]). With the advent of computing machines these approaches could be extended to 2D and 3D media leading to ray-theoretical tomography and the images of the Earth's interior that we know today (e. g., [115]). The analytical solution of wave propagation in spherical coordinates naturally leads to spherical harmonics and the possible quasi-analytical solution of wave propagation problems in spherically symmetric media using normal modes. As this methodology leads to complete waveforms the term "waveform inversion" was coined for fitting the waveforms of surface waves by correcting the phase differences for surface waves at particular frequencies (e. g., [118]). This allowed the recovery of seismic velocity models particularly of crust and upper mantle (surface wave tomography). A similar approach in Cartesian layered geometry led to complete solutions of the wave

equation in cylindrical coordinates through the summation of Bessel functions, the reflectivity method [46]. This method was later extended to spherical media through the Earth-flattening transformation [85]. Recently, ray-theory was extended allowing the incorporation of finite-frequency effects (e. g., [84]). The impact on the imaging process is still being debated.

Most of these methods are still today extremely valuable in providing first estimates of 2D or 3D effects and are important for the use in standard seismic processing due to their computational efficiency. Nevertheless, with the tremendous improvements of the quality of seismic observations we strive today to extract much more information on Earth's structure and sources from recorded waveforms. As waveforms are in most places strongly affected by 3D structural variations the application of numerical methods that solve "directly" the partial differential equations descriptive of wave propagation becomes mandatory. This necessity was recognized early on and the developments of numerical wave propagation began in the sixties of the 20th century.

### Numerical Methods Applied to Wave Propagation Problems

The finite-difference technique was the first numerical method to be intensively applied to the wave propagation problem (e. g., [1,6,61,77,82,83,88,89,116,117]). The partial differentials in the wave equation are replaced by finite differences leading to an extrapolation scheme in time that can either be implicit or explicit. The analysis of such simple numerical schemes led to concepts that are central to basically all numerical solutions of wave propagation problems. First, the discretization in space and time introduces a scale into the problem with the consequence that the numerical scheme becomes dispersive. This numerical dispersion – for the originally non-dispersive problem of purely elastic wave propagation – has the consequence that for long propagation distances wave pulses are no longer stable but disperse. The consequence is, that in any simulation one has to ascertain that enough grid points per wavelength are employed so that numerical dispersion is reduced sufficiently. Finding numerical schemes that minimize these effects has been at the heart of any new methodology ever since. Second, the so-called CFL criterion [24] that follows from the same theoretical analysis of the numerical scheme basically relates a "grid velocity" – the ratio between the space and time increments  $dx$  and  $dt$ , respectively – to the largest physical velocity  $c$  in the model. In order to have a stable calculation, this ratio has to be smaller than a constant  $\varepsilon$  that depends on the specific scheme and

the space dimension

$$c \frac{dt}{dx} \leq \varepsilon. \quad (1)$$

This simple relationship has important consequences: When the grid spacing  $dx$  must be small, because of model areas with low seismic velocities, then the time step  $dt$  has to be made smaller accordingly leading to an overall increase in the number of time steps and thus overall computational requirements. In addition, the early implementations were based on regular rectangular grids, implying that large parts of the model were carrying out unnecessary calculations. As shown below local time-stepping and local accuracy are important ingredients in efficient modern algorithms.

The fairly inaccurate low order spatial finite-difference schemes were later extended to high-order operators [26,48,49,50,51,56,76,103]. Nevertheless, the required number of grid points per wavelength was still large, particularly for long propagation distances. This has led to the introduction of pseudo-spectral schemes, "pseudo" because only the calculations of the derivatives were done in the spectral domain, but the wave equation was still solved in the space-time domain with a time-extrapolation scheme based on finite differences [10,45,47,67]). The advantage of the calculation of derivatives in the spectral domain is at hand: The Fourier theorem tells us that by multiplying the spectrum with  $ik$ ,  $i$  being the imaginary unit and  $k$  the wavenumber, we obtain an *exact* derivative (exact to numerical precision) on a regular set of grid points. This sounds attractive. However, there are always two sides to the coin. The calculation requires FFTs to be carried out extensively and the original "local" scheme becomes a "global" scheme. This implies that the derivative at a particular point in the computational grid becomes dependent on any other point in the grid. This turns out to be computationally inefficient, in particular on parallel hardware. In addition, the Fourier approximations imply periodicity which makes the implementation of boundary conditions (like the free surface, or absorbing boundary conditions) difficult.

By replacing the basis functions (Fourier series) in the classical pseudo-spectral method with Chebyshev polynomials that are defined in a limited domain  $(-1,1)$  the problem with the implementation of boundary problems found an elegant solution (e. g., [66,107,108]). However, through the irregular spacing of the Chebyshev collocation points (grid densification at the domain boundaries, see section below) new problems arose with the consequence that this approach was not much further pursued except in combination with a multi-domain approach in which the field



variables exchange their values at the domain boundaries (e. g., [108]).

So far, the numerical solutions described are all based on the *strong* form of the wave equation. The finite-element method is another main scheme that found immediate applications to wave propagation problems (e. g., [79]). Finite element schemes are based on solving the *weak* form of the wave equation. This implies that the space- and time-dependent fields are replaced by weighted sums of basis (also called trial) functions defined inside elements. The main advantage of finite element schemes is that elements can have arbitrary shape (e. g., triangular, trapezoidal, hexahedral, tetrahedral, etc.). Depending on the polynomial order chosen inside the elements the spatial accuracy can be as desired. The time-extrapolation schemes are usually based on standard finite differences. There are several reasons why finite-element schemes were less widely used in the field of wave propagation. First, in the process a large system matrix needs to be assembled and must be inverted. Matrix inversion in principle requires global communication and is therefore not optimal on parallel hardware. Second, in comparison with the finite-element method, finite-difference schemes are more easily coded and implemented due to their algorithmic simplicity.

A tremendous step forward was the introduction of basis functions inside the elements that have spectral accuracy, e. g., Chebyshev or Legendre polynomials [11,15,39,40,65,86,90,98]. The so-called spectral element scheme became particularly attractive with the discovery that – by using Legendre polynomials – the matrices that required inversion became diagonal [65]. This implies that the scheme does no longer need global communication, it is a local scheme in which extrapolation to the next time step can be naturally parallelized. With the extension of this scheme to spherical grids using the cubed-sphere discretization [63,64] this scheme is today the method of choice on many scales unless highly complex models need to be initiated.

Most numerical schemes for wave propagation problems were based on regular, regular stretched, or hexahedral grids. The numerical solution to unstructured grids had much less attention, despite the fact that highly complex models with large structural heterogeneities seem to be more readily described with unstructured point clouds. Attempts were made to apply finite volume schemes to this problem [31], and other concepts (like natural neighbor coordinates [7] to find numerical operators that are applicable on unstructured grids [72,73,78]). These approaches were unfortunately not accurate enough to be relevant for 3D problems. Recently, a new flavor of nu-

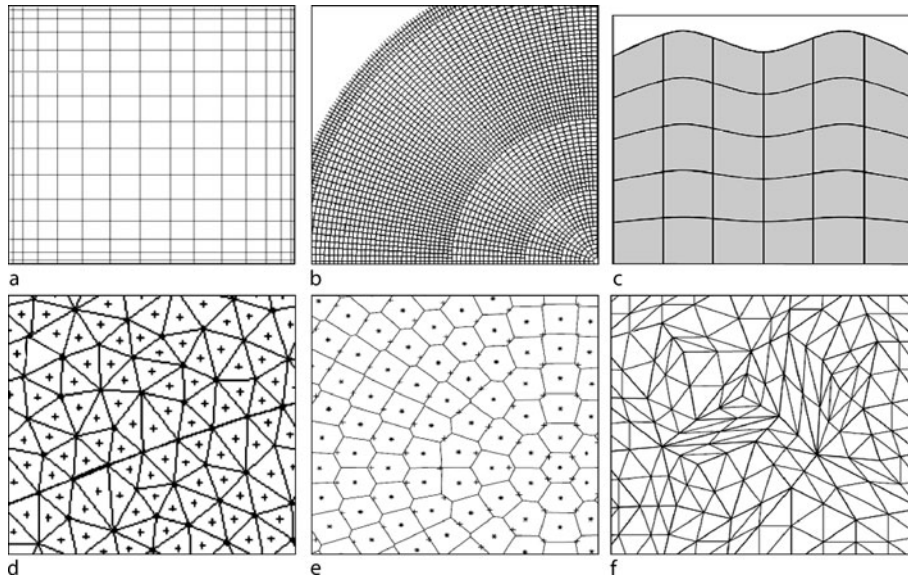
merical method found application to wave propagation on triangular or tetrahedral grids. This combination of a discontinuous Galerkin method with ideas from finite volume schemes [33,70] allows for the first time arbitrary accuracy in space and time on unstructured grids. While the numerical solution on tetrahedral grids remains computationally slower, there is a tremendous advantage in generating computational grids for complex Earth models. Details on this novel scheme are given below.

Before presenting two schemes (spectral elements and the discontinuous Galerkin method) and some applications in more detail we want to review the evolution of grids used in wave propagation problems.

### Grids for Wave Propagation Problems

The history of grid types used for problems in computational wave propagation is tightly linked to the evolution of numerical algorithms and available computational resources. The latter in the sense that – as motivated in the introduction – even today realistic simulations of wave propagation are still computationally expensive. This implies that it is not sufficient to apply stable and simple numerical schemes and just use enough grid points per wavelength and/or extremely fine grids for geometrically complex models. Optimal mathematical algorithms that minimize the computational effort are still sought for as the recent developments show that are outlined in the following sections.

In Fig. 4 a number of different computational grids in two space dimensions is illustrated. The simple-most equally-spaced regular finite-difference grid is only of practical use in situations without strong material discontinuities. With the introduction of the pseudospectral method based on Chebyshev polynomials grids as shown in Fig. 4a grids appeared that are denser near the domain boundaries and coarse in the interior. While this enabled a much more efficient implementation of boundary conditions the ratio between the size of the largest to the smallest cell depends on the overall number of grid points per dimension and can be very large. This leads to very small time steps, that can in some way be compensated by grid stretching [9] but overall the problem remains. An elegant way of allowing grids to be of more practical shape is by stretching the grids using analytical functions (Fig. 4c, this basically corresponds to a coordinate transformation, e. g., [50,107]). By doing this either smooth surface topography or smoothly varying internal interfaces can be followed by the grid allowing a more efficient simulation of geometrical features compared to a blocky representation on standard finite difference grids.



Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 4

Examples of 2D grids used for wave propagation simulations. **a** Chebyshev grid with grid densification near the domain boundaries. **b** Multidomain finite-difference grid in regular spherical coordinates. **c** Stretched regular finite-difference grid that allows following smoothly varying interface or surface boundaries. **d** Triangular staggered grid following an interface that allows finite-difference type operators. **e** Unstructured grid with associated Voronoi cells for calculations using the finite-volume method. **f** Triangular cells for finite-element type calculations. See text for details and references

The problem of global wave propagation using spherical coordinates (here in the two-dimensional, axi-symmetric approximation) nicely illustrates the necessity to have spatially varying grid density (e. g., [42,43,53,59,89,109]). The grid shown in Fig. 4b demonstrates that in spherical coordinates a regular discretization leads to grid distances that get smaller and smaller towards the center of the Earth. This is in contrast to what is required to efficiently model the Earth's velocity structure: Velocities are small near the surface (requiring high grid density) and increase towards the center of the Earth (requiring low grid density). One way of adjusting is by re-gridding the mesh every now and then, in this case doubling the grid spacing appropriately. This is possible, yet it requires interpolation at the domain boundaries that slightly degrades the accuracy of the scheme.

The problems with grid density and complex surfaces cry for the use of so-called unstructured grids. Let us define an unstructured grid as an initial set of points (a point cloud), each point characterized by its spatial coordinates. We wish to solve our partial differential equations on this point set. It is clear that – with appropriate grid generation software – it is fairly easy to generate such grids that obey exactly any given geometrical constraints be it in connection with surfaces or velocity models (i. e., varying grid density). It is important to note that such point clouds can-

not be represented by 2D or 3D matrices as is the case for regular or regular stretched grid types. This has important consequences for the parallelization of numerical schemes. The first step after defining a point set is to use concepts from computational geometry to handle the previously unconnected points. This is done through the idea of Voronoi cells, that uniquely define triangles and their neighbors (Delauney triangulation). In Fig. 4d an example is shown for a triangular grid that follows an internal interface [72]. For finite-difference type operators on triangular grids a grid-staggering makes sense. Therefore, velocities would be defined in the center of triangles and stresses at the triangle vertices. Voronoi cells (Fig. 4e) can be used as volumetric elements for finite volume schemes [31,73]. For finite-element schemes triangular elements (Fig. 4f, e. g., [70]) with appropriate triangular shape functions are quite standard but have not found wide applications in seismology.

If the grid spacing of a regular finite-difference grid scheme in 3D would have to be halved this would result in an overall increase of computation time by a factor of 8 (a factor 2 per space dimension and another factor 2 because of the necessary halving of the time step). This simply means that the accuracy of a specific numerical scheme and the saving in memory or computation time is much more relevant in three dimensions. The evolu-

tion of grids in three dimensions is illustrated with examples in Fig. 5. A geometrical feature that needs to be modeled correctly particularly in volcanic environments is the free surface. With standard regular-spaced finite-difference schemes only a stair-step representation of the surface is possible (Fig. 5a, e. g., [87,92]). While the specific numerical implementation is stable and converges to the correct solution a tremendous number of grid points is necessary to achieve high accuracy.

Chebyshev grids and regular grids were applied to the problem of wave propagation in spherical sections (Fig. 5b, e. g., [52,57]). The advantage of solving the problem in spherical coordinates is the natural orthogonal coordinate system that facilitates the implementation of boundary conditions. However, due to the nature of spherical coordinates the physical domain should be close to the equator and geographical models have to be rotated accordingly. A highly successful concept for wave propagation in spherical media was possible through the adoption of the cubed-sphere approach in combination with spectral-elements (Fig. 5c, [63,64]). The cubed-sphere discretization is based on hexahedral grids. Towards the center of the Earth the grid spacing is altered to keep the number of elements per wavelength approximately constant.

Computational grids for wave propagation based on tetrahedra (Fig. 5d,e) are only recently being used for seismic wave propagation in combination with appropriate numerical algorithms such as finite volumes [34] or discontinuous Galerkin (e. g., [70]). The main advantage is that the grid generation process is greatly facilitated when using tetrahedra compared to hexahedra. Generating point clouds that follow internal velocity structures and connecting them to tetrahedra are straight forward and efficient mathematical computations. However, as described in more detail below, tetrahedral grids require more involved computations and are thus less efficient than hexahedral grids. Complex hexahedral grids – even for combined modeling of structure and soil (Fig. 5f) are possible but – at least at present – require a large amount of manual interaction during the grid generation process. It is likely that the combination of both grid types (tetrahedral in complex regions, hexahedral in less complex regions) will play an important role in future developments.

In the following we would like to present two of the most competitive schemes presently under development, (1) the spectral element method and (2) the discontinuous Galerkin approach combined with finite-volume flux schemes. The aim is to particularly illustrate the role of the grid generation process and the pros and cons of the specific methodologies.

### 3D Wave Propagation on Hexahedral Grids: Soil-Structure Interactions

We briefly present the spectral element method (SEM) based on Lagrange polynomials, focusing only on its main features and on its implementation for the solution of the elasto-dynamic equations. The SEM can be regarded as a generalization of the finite element method (FEM) based on the use of high order piecewise polynomial functions. The crucial aspect of the method is the capability of providing an arbitrary increase in spatial accuracy simply enhancing the algebraic degree of these functions (the spectral degree SD). On practical ground, this operation is completely transparent to the users, who limit themselves to choosing the spectral degree at runtime, leaving to the computational code the task of building up suitable quadrature points for integration and new degrees of freedom. Obviously, the increasing spectral degree implies raising the required computational effort.

On the other hand, one can also play on the grid refinement to improve the accuracy of the numerical solution, thus following the standard finite element approach. Spectral elements are therefore a so-called “ $h-p$ ” method, where “ $h$ ” refers to the grid size and “ $p$ ” to the degree of polynomials. Referring to Faccioli et al. [40], Komatitsch and Vilotte [65], Chaljub et al. [15] for further details, we briefly remind in the sequel the key features of the spectral element method adopted. We start from the wave equation for the displacement  $\mathbf{u}$ :

$$\rho \frac{\partial \mathbf{u}^2}{\partial t^2} = \text{div } \sigma_{ij}(\mathbf{u}) + f, \quad i, j = 1 \dots d (d = 2, 3) \quad (2)$$

where  $t$  is the time,  $\rho = \rho(x)$  the material density,  $f = f(x, t)$  a known body force distribution and  $\sigma_{ij}$  the stress tensor. Introducing Hooke’s law:

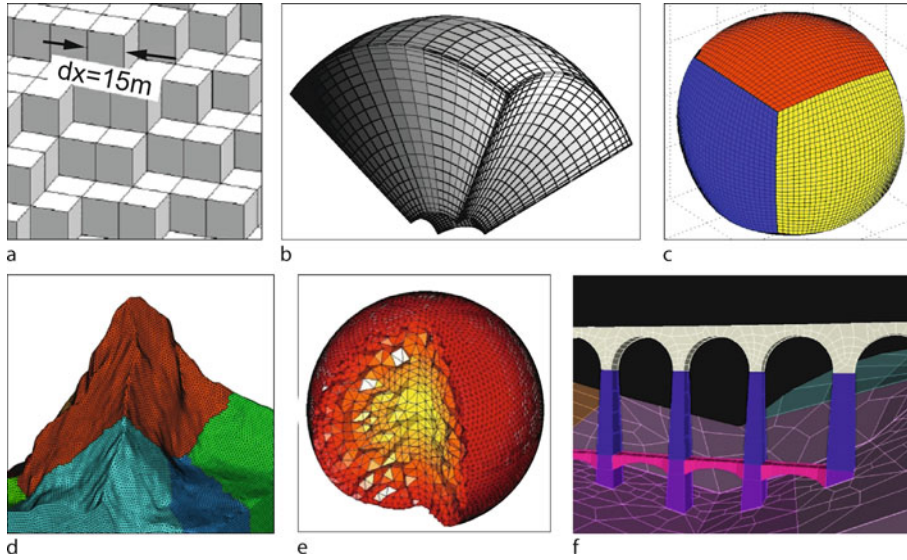
$$\sigma_{ij}(\mathbf{u}) = \lambda \text{div } \mathbf{u} \delta_{ij} + 2\mu \varepsilon_{ij}(\mathbf{u}), \quad (3)$$

where

$$\varepsilon_{ij}(\mathbf{u}) = \frac{1}{2} \left( \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) \quad (4)$$

is the strain tensor,  $\lambda$  and  $\mu$  are the Lamé coefficients, and  $\delta_{ij}$  is the Kronecker symbol, i. e.  $\delta_{ij} = 1$  if  $i = j$  and  $\delta_{ij} = 0$ , otherwise.

As in the FEM approach, the dynamic equilibrium problem for the medium can be stated in the weak, or variational form, through the principle of virtual work [121] and through a suitable discretization procedure that depends on the numerical approach adopted, can be written as an ordinary differential equations system with respect



Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 5

Examples of 3D grids. **a** Stair-step representation of a complex free surface with finite-difference cells. **b** Chebyshev grid in spherical coordinates for a spherical section. **c** Cubed sphere grid used for spectral-element and multi-domain Chebyshev calculations. **d** Tetrahedral grid of the Matterhorn. **e** Tetrahedral grid of the Earth's interior with the grid density tied to the velocity model. **f** Hexahedral grid of bridge structure and subsurface structure for spectral-element calculations. See text for details and references

to time:

$$[M] \ddot{\mathbf{U}}(t) + [K] \mathbf{U}(t) = \mathbf{F}(t) + \mathbf{T}(t) \quad (5)$$

with matrices  $[M]$  and  $[K]$ , respectively the mass and stiffness matrices, and vectors  $\mathbf{F}$  and  $\mathbf{T}$  representing the contributions of external forces and traction conditions, respectively. In our SE approach,  $\mathbf{U}$  denotes the displacement vector at the Gauss–Lobatto–Legendre (GLL) nodes, that correspond to the zeroes of the first derivatives of Legendre polynomial of degree  $N$ . The advancement of the numerical solution in time is provided by the explicit 2nd order leap-frog scheme. This scheme is conditionally stable and must satisfy the well known and already mentioned Courant–Friedrichs–Levy (CFL) condition. The key features of the SE discretization are described in the following.

Like in the FEM standard technique, the computational domain may be split into quadrilaterals in 2D or hexahedra in 3D, both the local distribution of grid points within the single element and the global mesh of all the grid points in the domain must be assigned. Many of these grid points are shared amongst several spectral elements. Each spectral element is obtained by a mapping of a master element through a suitable transformation and all computations are performed on the master element. Research is in progress regarding the introduction of triangular spec-

tral elements [80]. The nodes within the element where displacements and spatial derivatives are computed, on which volume integrals are evaluated, are not necessarily equally spaced (similar to the Chebyshev approach in pseudospectral methods mentioned above). The interpolation of the solution within the element is done by Lagrange polynomials of suitable degree. The integration in space is done through Gauss–Lobatto–Legendre quadrature formula.

Thanks to this numerical strategy, the exponential accuracy of the method is ensured and the computational effort minimized, since the mass matrix results to be diagonal. The spectral element (SE) approach developed by Faccioli et al. [40] has been recently implemented in the computational code GeoELSE (GeoElasticity by Spectral Elements) [93,102,120] for 2D/3D wave propagation analysis. The most recent version of the code includes: (i) the capability of dealing with fully unstructured computational domains, (ii) the parallel architecture, and (iii) visco-plastic constitutive behavior [30]. The mesh can be created through an external software (e.g., CUBIT [25]) and the mesh partitioning is handled by METIS [81].

### Hexahedral Grids

As already mentioned in the SEM here presented the computational domain is decomposed into a family of non

overlapping quadrilaterals in 2D or hexahedra in 3D. The grid discretization should be suitable to accurately propagate up to certain frequencies. Obviously, owing to the strong difference of the mechanical properties between soft-soil and stiff-soil (or building construction material) and to the different geometrical details as well, the grid refinement needed in the various parts of the model varies substantially. Therefore, a highly unstructured mesh is needed to minimize the number of elements. While 3D unstructured tetrahedral meshes can be achieved quite easily with commercial or non commercial software, the creation of a 3D non structured hexahedra mesh is still recognized as a challenging problem. In the following paragraph we provide state of the art results concerning the mesh creation.

### Grid Generation

Hexahedral grids have more severe restrictions in meshing efficiently. This is basically related to the intrinsic difficulty that arises from the mapping of the computational domain with this particular element. As a consequence automatic procedures have difficulty capturing specific boundaries, create poor quality elements, the assigned size is difficult to be preserved and the generation process is usually much slower compared to the tetrahedral mesh generation algorithms. On the other hand the advantages of hexahedral meshes are usually related to the lower computational cost of the wave propagation solutions with respect to the one based on triangular meshes or hexahedral structured grids (like in the finite difference method).

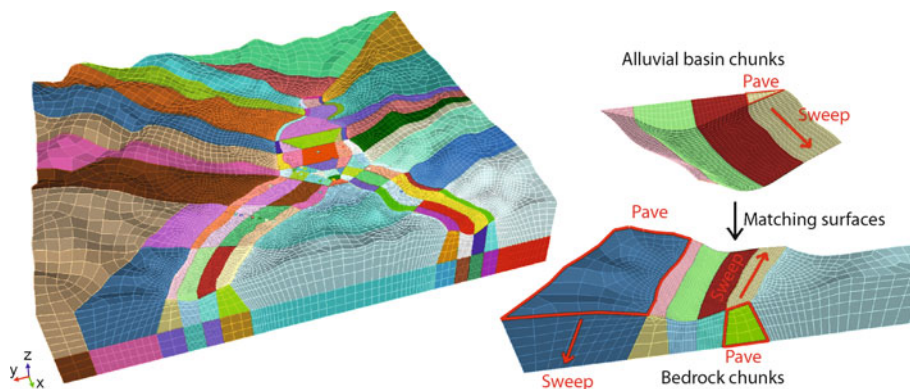
Nevertheless certain problems can be addressed reasonably well with specific solutions. A quite typical case

in earthquake seismology is the study of the alluvial basin response under seismic excitation. In handling this problem, a first strategy is to try to “honor” the interface between the sediment (soft soil) and the bedrock (stiff soil). The two materials are divided by a physical interface and the jump in the mechanical properties is strictly preserved. The major drawback of this approach is that usually it requires strong skills from the user to build-up the mesh and a significant amount of working time (Fig. 6). Given that the “honoring approach” is not always feasible in a reasonable time (or with a reasonable effort) a second strategy is worth to be mentioned: The so called “not honoring” procedure. In this second case the mesh is refined in proximity of the area where the soft deposit are localized but the elements do not respect the interface. On a practical ground the mechanical properties are assigned node by node and the sharp jump is smoothed through the Lagrange interpolation polynomial and substituted with smeared interfaces (Fig. 7). At the present time it is still strongly under debate if it is worth to honor or not the physical interfaces.

Finally, we highlight the fact that meshing software (e.g., CUBIT [25]) is available that seems to be extremely promising and potentially very powerful for the creation of geophysical and seismic engineering unstructured hexahedral meshes. Further very interesting mesh generation procedures based on hexahedral are under investigation [99].

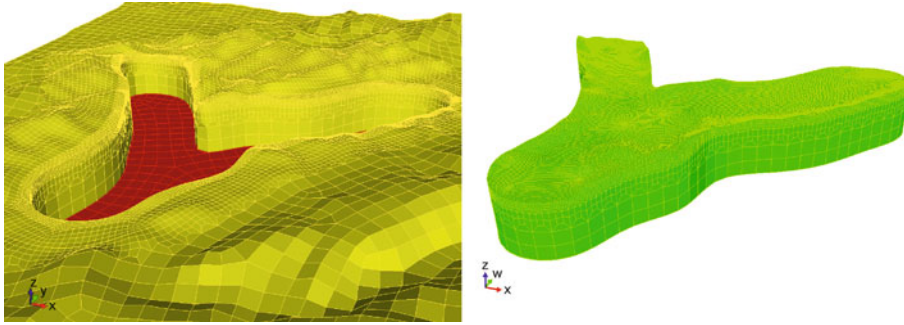
### Scale Problem with Structure and Soil

In engineering practice one of the most common approaches to design buildings under seismic load is the imposition of an acceleration time history to the structure,



Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 6

3D numerical model used for the simulations of ESG06 “Grenoble Benchmark”. “Honoring” technique: The computational domain is subdivided into small chunks and each one is meshed starting from the alluvial basin down to the bedrock. For simplicity only the spectral elements are shown without GLL nodes



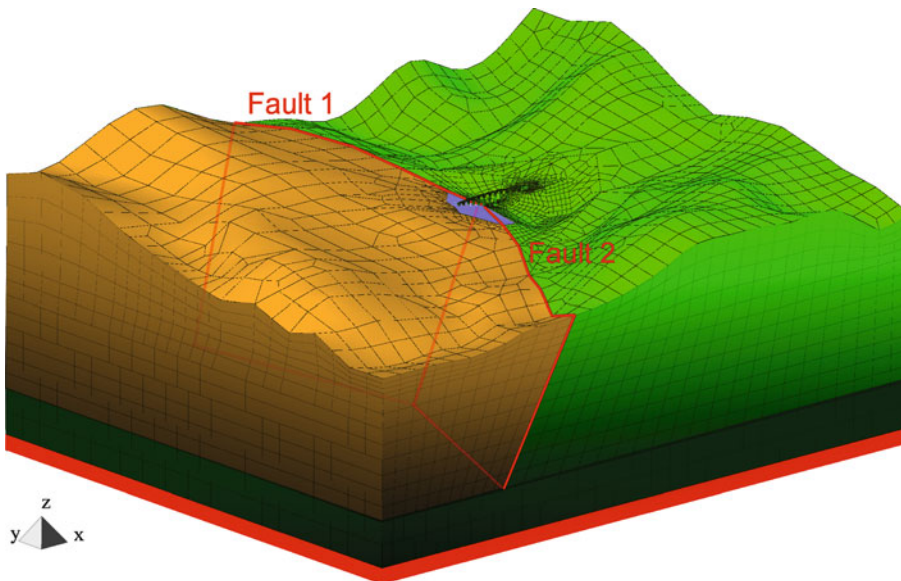
Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 7

3D numerical model used for the simulations of ESG06 “Grenoble Benchmark”. “Not Honoring” technique: The computational domain is meshed with a coarse mesh and then refined twice approximately in the area where the alluvial basin is located

basically acting like an external load. An excellent example of this technique can be found in recent publications (e. g., [68,69]) and in the study of the so-called “urban-seismology”, recently presented by Fernandez-Ares et al., [44]. In this case the goal is to understand how the presence of an entire city can modify the incident wave-field. Due to the size of the simulation and the number of buildings, the latter are modeled as single degrees of freedom oscillators. The interaction between soil and structure is preserved but the buildings are simplified. For important structure (e. g.: Historical buildings, world heritage buildings, hospitals,

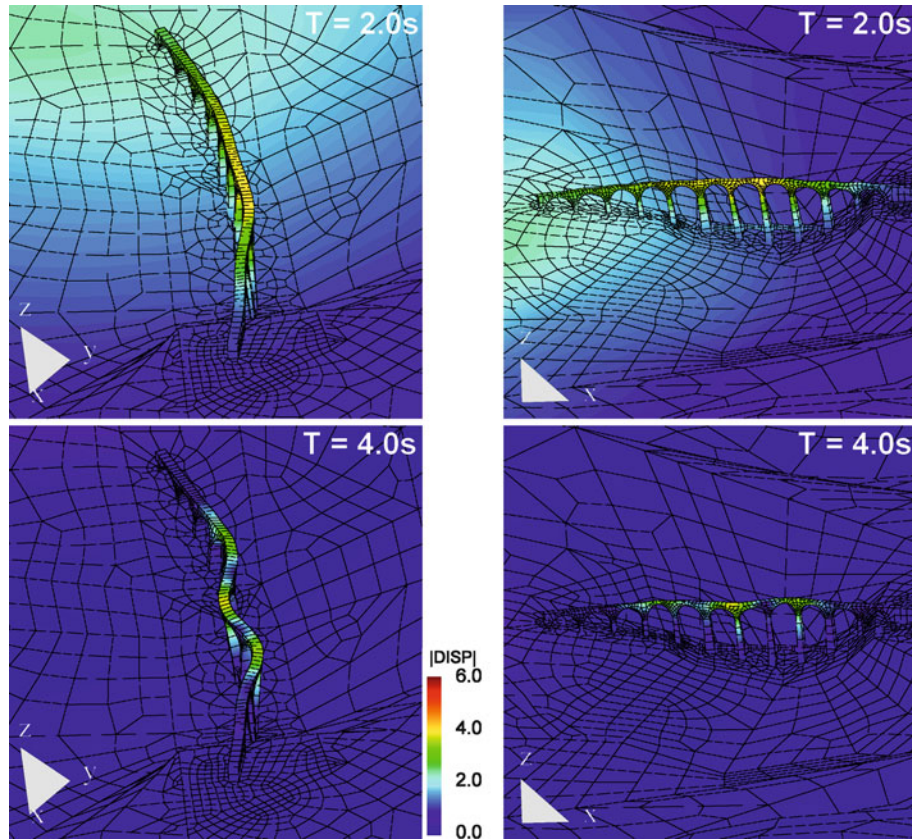
schools, theaters, railway and highways) it is worth to provide an ad-hoc analysis capable to take into account the full complexity of the phenomena.

Here we present an example of a fully coupled modeling (Fig. 8): A railway bridge and its geotechnical-topographical surroundings. The Acquasanta bridge on the Genoa-Ovada railway, North Italy, is located in the Genoa district and represents a typical structure the ancestor of which can be traced back to the Roman “Pont du Gard”. This structural type did not change significantly along the centuries, thanks to the excellent design achieved no less



Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 8

3D model of Acquasanta bridge and the surrounding geological configuration. The investigated area is 2 km in length, 1.75 km in width and 0.86 km in depth. The model was designed to propagate waves up to 5 Hz with a  $SD = 3$  (Order 4) and has 38,569 hexahedral elements and 1,075,276 grid points. The contact between calcareous schists (brown color) and serpentine rocks (green color) is modeled with two sub-vertical faults (red-line). Cyan color represents the alluvial and weathered deposits



Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 9  
 Snapshots of the modulus of the displacement vector and the magnified deformed shape of the bridge (in mm)

than 1900 years ago. The Acquasanta bridge structure is remarkable both for the site features and the local geological and geomorphological conditions. The foundations of several of the piers rest on weak rock; moreover, some instability problems have been detected in the past on the valley slope towards Ovada.

Several simulations have been performed with GeoELSE, in order to evaluate the influence of seismic site effects on the dynamic response of the Acquasanta bridge. A fully coupled 3D soil-structure model was designed: The grid is characterized by a “subvertical fault” between calcareous schists and serpentine rocks. This is in accordance with available data, even if further investigations in future should identify more in detail the tectonic structure of the area. The geometry of weathered materials overlaying the calcareous schists on the Ovada side has been assumed according to available information. The dimension of hexahedral elements ranges some tens of centimeters to about 1000 m. With such a model, the problem can be handled in its 3D complexity and we can exam-

ine the following aspects that are usually analyzed under restrictive and simplified assumptions: (i) soil-structure interaction, (ii) topographic amplification, (iii) soft soil amplification (caused by the superficial alluvium deposit shown in cyan), (iv) subvertical fault (red line) between the schists, on the Ovada side, and serpentine rock, on the Genoa side. For excitation a shear plane wave ( $x$ -direction) was used (Ricker wavelet,  $f_{max} = 3$  Hz,  $t_0 = 1.0$  s, and amplitude = 1 mm) propagating vertically from the bottom (red elements in Fig. 8).

In Fig. 9 we present some snapshots of the modulus of the displacement vector and the magnified deformed shape of the bridge. It is worth to note that at  $T = 2$  s the motion of the bridge is almost in-plane (direction  $x$ ), while at  $T = 4$  s is clearly evident how the coupling between the in-plane and out-plane ( $y$ -direction) motion starts to be important.

The study of the soil-structure interaction problem could be easily enhanced (i) improving the input excitation of the model here presented and (ii) taking into ac-

count complex constitutive behavior both from the soil and the structure side. The former is already available in GeoELSE thanks to the recent implementation [41,93] of the domain reduction method (DRM), a methodology that divides the original problem into two simpler ones [4,119], to overcome the problem of multiple physical scales that is created by a seismic source located some kilometers away from the structure with typical element size of the order of meters and located over a relatively small area (less than 1 km<sup>2</sup>) on soft deposit. The latter still needs to be improved because of the lack of a complete tool capable to handle in 3D non linear soil behavior, non-linear structural behavior and the presence of the water, that play a crucial role in the failure of buildings. Partial response to this problem can be found in the recent work of Bonilla et al. [5] and in the visco-plastic rheology recently introduced in GeoELSE [30].

### 3D Wave Propagation on Tetrahedral Grids: Application to Volcanology

As indicated above, the simulation of a complete, highly accurate wave field in realistic media with complex geometry is still a great challenge. Therefore, in the last years a new, highly flexible and powerful simulation method has been developed that combines the Discontinuous Galerkin (DG) Method with a time integration method using Arbitrary high order DERivatives (ADER) of the approximation polynomials. The unique property of this numerical scheme is, that it achieves arbitrarily high approximation order for the solution of the governing seismic wave equation in space and time on structured and unstructured meshes in two and three space dimensions.

Originally, this new ADER-DG approach [32,35] was introduced for general linear hyperbolic equation systems with constant coefficients or for linear systems with variable coefficients in conservative form. Then, the extension to non-conservative systems with variable coefficients and source terms and its particular application to the simulation of seismic waves on unstructured triangular meshes in two space dimensions was presented [70]. And finally, the further extension of this approach to three-dimensional tetrahedral meshes has been achieved [33]. Furthermore, the accurate treatment of viscoelastic attenuation, anisotropy and poroelasticity has been included to handle more complex rheologies [28,29,71]. The governing system of the three-dimensional seismic wave equations is hereby formulated in velocity-stress and leads to the hyperbolic system of the form

$$\frac{\partial \mathbf{Q}_p}{\partial t} + A_{pq} \frac{\partial \mathbf{Q}_q}{\partial \xi} + B_{pq} \frac{\partial \mathbf{Q}_q}{\partial \eta} + C_{pq} \frac{\partial \mathbf{Q}_q}{\partial \zeta} = S_p, \quad (6)$$

where the vector  $\mathbf{Q}$  of unknowns contains the six stress and the three velocity components and  $S$  is the source term. The Jacobian matrices  $A$ ,  $B$  and  $C$  include the material values as explained in detail in [33,70].

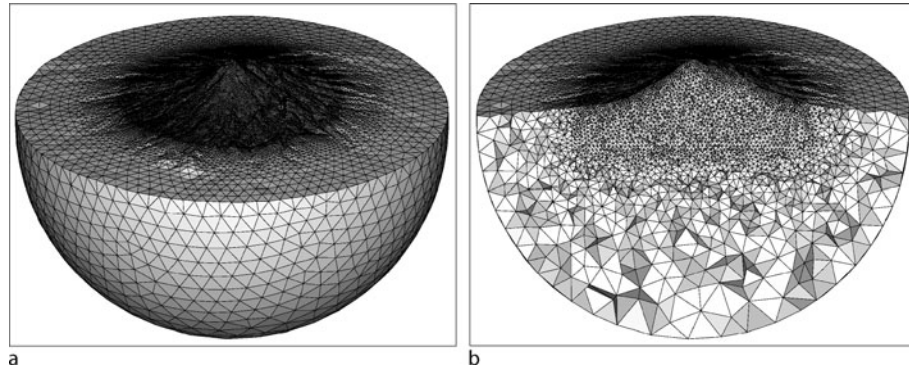
### The ADER-DG Method: Basic Concepts

The ADER-DG method is based on the combination of the ADER time integration approach [113], originally developed in the finite volume (FV) framework [96,97,111] and the Discontinuous Galerkin finite element method [18,19,20,21,22,23,91]. As described in detail in [33] in the ADER-DG approach the solution is approximated inside each tetrahedron by a linear combination of space-dependent polynomial basis functions and time-dependent degrees of freedom as expressed through

$$(\mathbf{Q}_h)_p(\xi, \eta, \zeta, t) = \hat{\mathbf{Q}}_{pl}(t) \Phi_l(\xi, \eta, \zeta), \quad (7)$$

where the basis functions  $\Phi_l$  form an orthogonal basis and are defined on the canonical reference tetrahedron. The unknown solution inside each element is then approximated by a polynomial, whose coefficients – the degrees of freedom  $\hat{\mathbf{Q}}_{pl}$  – are advanced in time. Hereby, the solution can be discontinuous across the element interfaces, which allows the incorporation of the well-established ideas of numerical flux functions from the finite volume framework [75,112]. To define a suitable flux over the element surfaces, so-called Generalized Riemann Problems (GRP) are solved at the element interfaces. The GRP solution provides simultaneously a numerical flux function as well as a time-integration method. The main idea is a Taylor expansion in time in which all time derivatives are replaced by space derivatives using the so-called Cauchy–Kovalevski procedure which makes recursive use of the governing differential Eq. (6). The numerical solution of Eq. (6) can thus be advanced by one time step without intermediate stages as typical e. g. for classical Runge–Kutta time stepping schemes. Due to the ADER time integration technique the same approximation order in space and time is achieved automatically. Furthermore, the projection of the elements in physical space onto a canonical reference element allows for an efficient implementation, as many computations of three-dimensional integrals can be carried out analytically beforehand. Based on a numerical convergence analysis this new scheme provides arbitrary high order accuracy on unstructured meshes. Moreover, due to the choice of the basis functions in Eq. (7) for the piecewise polynomial approximation [23], the ADER-DG method shows even spectral convergence.





Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 10

Tetrahedral mesh for the model of the volcano Merapi. The zone of interest, such as the free surface topography and the volcano's interior are discretized by a fine mesh, whereas the spatial mesh is gradually coarsened towards the model boundaries

### Grid Generation: Unstructured Triangulations and Tetrahedralization

Both tetrahedral and hexahedral elements are effectively used to discretize three-dimensional computational domains and model wave propagation with finite element type methods. Tetrahedra can be the right choice because of the robustness when meshing any general shape. Hexahedra can be the element of choice due to their ability to provide more efficiency and accuracy in the computational process. Furthermore, techniques for automatic mesh generation, gradual mesh refinement and coarsening are generally much more robust for tetrahedral meshes in comparison to hexahedral meshes. Straightforward tetrahedral refinement schemes, based on longest-edge division, as well as the extension to adaptive refinement or coarsening procedures of a refined mesh exist [3,12]. In addition, parallel strategies for refinement and coarsening of tetrahedral meshes have been developed [27].

Less attention has been given to the modification of hexahedral meshes. Methods using iterative octrees have been proposed [74,95], but these methods often result in nonconformal elements that cannot be accommodated by some solvers. Lately also conformal refinement and coarsening strategies for hexahedral meshes have been proposed [2]. Other techniques insert non-hexahedral elements that result in hybrid meshes that need special solvers that can handle different mesh topologies. Commonly, the geometrical problems in geosciences arise through rough surface topography, as shown for the Merapi volcano in Fig. 10, and internal material boundaries of complex shape that lead to wedges and overturned or discontinuous surfaces due to folding and faulting. However, once the geometry of the problem is defined by the help of modern computer aided design (CAD) software,

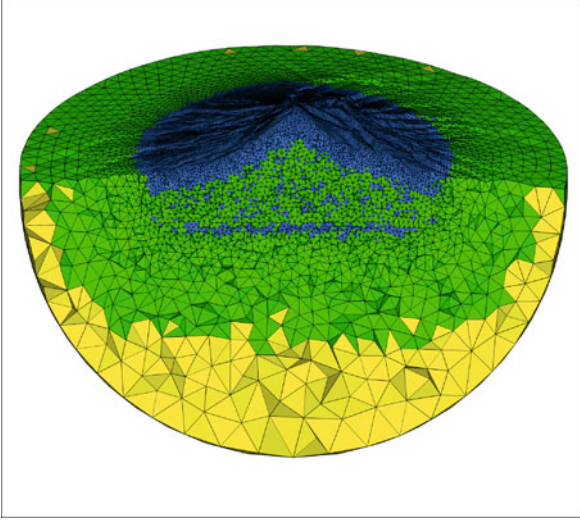
the meshing process using tetrahedral elements is automatic and stable. After the mesh generation process, the mesh vertices, the connectivity matrix and particular information about boundary surfaces are typically imported to a solver.

The computational possibilities and algorithmic flexibility of a particular solver using the ADER-DG approach for tetrahedral meshes are presented in the following.

### Local Accuracy: $p$ -Adaptation

In many large scale applications the computational domain is much larger than the particular zone of interest. Often such an enlarged domain is chosen to avoid effects from the boundaries that can pollute the seismic wave field with possible, spurious reflections. Therefore, a greater number of elements has to be used to discretize the domain describing the entire model. However, in most cases the high order accuracy is only required in a restricted area of the computational domain and it is desirable to choose the accuracy that locally varies in space. This means, that it must be possible to vary the degree  $p$  of the approximation polynomials locally from one element to the other [36]. As the ADER-DG method uses a hierarchical order of the basis functions to construct the approximation polynomials, the corresponding polynomial coefficients, i. e. the degrees of freedom, for a lower order polynomial are always a subset of those of a higher-order one. Therefore, the computation of fluxes between elements of different approximation orders can be carried out by using only the necessary part of the flux matrices.

Furthermore, the direct coupling of the time and space accuracy via the ADER approach automatically leads to a local adaptation also in time accuracy, which often is referred to as  $p_t$ -adaptivity. In general, the distribution of the



Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 11

The local degree  $p$  of the approximation polynomial depends on the insphere radius of each tetrahedral element and is given in color code. Close to the surface topography an approximation polynomial of degree  $p = 5$  (blue) is used, whereas in depth the degree is reduced to  $p = 4$  (green) and  $p = 3$  (yellow)

degree  $p$  might be connected to the mesh size  $h$ , i. e. the radius of the inscribed sphere of a tetrahedral element. In particular, the local degree  $p$  can be coupled to the mesh size  $h$  via the relations

$$p = p_{\min} + (p_{\max} - p_{\min}) \left( \frac{h - h_{\min}}{h_{\max} - h_{\min}} \right)^r, \quad (8)$$

$$p = p_{\max} - (p_{\max} - p_{\min}) \left( \frac{h - h_{\min}}{h_{\max} - h_{\min}} \right)^r, \quad (9)$$

where the choice of the power  $r$  determines the shape of the  $p$ -distribution. Note, that depending on the choice of the first term and the sign the degree  $p$  can increase as in Eq. (8) or decrease as in Eq. (9) with increasing  $h$ , starting from a minimum degree  $p_{\min}$  up to a maximum degree  $p_{\max}$ . This provides additional flexibility for the distribution of  $p$  inside the computational domain. An example of a  $p$ -distribution for the volcano Merapi is given in Fig. 11.

Here the idea is to resolve the slowly propagating surface waves with high accuracy, whereas the waves propagating towards the absorbing model boundaries pass through a zone of low spatial resolution. This approach leads to numerical damping due to an amplitude decay that reduces possible boundary reflections. Furthermore, the computational cost is reduced significantly due to the strongly reduced number of total degrees of freedom in the model.

### Local Time Stepping: $\Delta t$ -Adaptation

Geometrically complex computational domains or spatial resolution requirements often lead to meshes with small or even degenerate elements. Therefore, the time step for explicit numerical schemes is restricted by the ratio of the size  $h$  of the smallest element and the corresponding maximum wave speed in this element. For global time stepping schemes all elements are updated with this extremely restrictive time step length leading to a large amount of iterations. With the ADER-DG approach, time accurate local time stepping can be used, such that each element is updated by its own, optimal time step [36]. Local time-stepping was used in combination with the finite-difference method [42,106].

An element can be updated to the next time level if its actual time level and its local time step  $\Delta t$  fulfill the following condition with respect to all neighboring tetrahedra  $n$ :

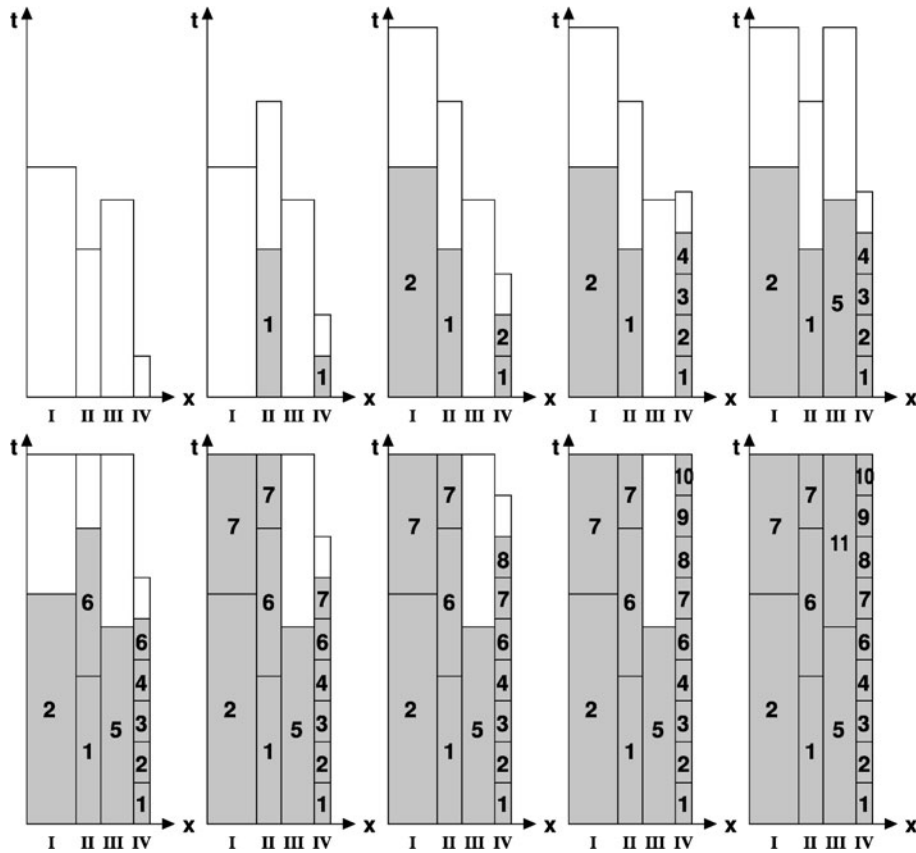
$$t + \Delta t \leq \min(t_n + \Delta t_n). \quad (10)$$

Figure 12 is visualizing the evolution of four elements (I, II, III and IV) in time using the suggested local time stepping scheme. A loop cycles over all elements and checks for each element, if condition (10) is fulfilled. At the initial state all elements are at the same time level, however, element II and IV fulfill condition (10) and therefore can be updated. In the next cycle, these elements are already advanced in time (grey shaded) in cycle 1. Now elements I and IV fulfill condition (10) and can be updated to their next local time level in cycle 2. This procedure continues and it is obvious, that the small element IV has to be updated more frequently than the others. A synchronization to a common global time level is only necessary, when data output at a particular time level is required as shown in Fig. 12.

Information exchange between elements across interfaces appears when numerical fluxes are calculated. These fluxes depend on the length of the local time interval over which a flux is integrated and the corresponding element is evolved in time. Therefore, when the update criterion (10) is fulfilled for an element, the flux between the element itself and its neighbor  $n$  has to be computed over the local time interval:

$$\tau_n = [\max(t, t_n), \min(t + \Delta t, t_n + \Delta t_n)]. \quad (11)$$

As example, the element III fulfills the update criterion (10) in cycle 5 (see Fig. 12). Therefore, when computing the fluxes only the remaining part of the flux given by the intervals in Eq. (11) has to be calculated. The other flux contribution was already computed by the neighbors II and IV during their previous local updates. These flux



Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 12

Visualization of the local time stepping scheme. The actual local time level  $t$  is at the top of the *gray shaded* area with numbers indicating the cycle, in which the update was done. *Dotted lines* indicate the local time step length  $\Delta t$  with which an element is updated

contributions have been accumulated and were stored into a memory variable and therefore just have to be added.

Note that e. g. element IV reaches the output time after 10 cycles and 9 local updates, which for a global time stepping scheme would require  $9 \times 4 = 36$  updates for the all four elements. With the proposed local time stepping scheme only 16 updates are necessary to reach the same output time with all elements as indicated by the final number of grey shaded space time elements in Fig. 12.

Comparing these numbers leads to a speedup factor of 2.25. For strongly heterogeneous models and local time step lengths this factor can become even more pronounced. However, due to the asynchronous update of elements that might be spatially very close to each other the mesh partitioning for parallel computations becomes an important and difficult issue. Achieving a satisfying load balancing is a non-trivial task and still poses some unresolved problems as explained in the following.

### Mesh Partitioning and Load Balancing

For large scale applications it is essential to design a parallel code that can be run on massively parallel supercomputing facilities. Therefore, the load balancing is an important issue to use the available computational resources efficiently. For global time stepping schemes without  $p$ -adaptation standard mesh partitioning as done e. g. by METIS [60] is sufficient to get satisfying load balancing. The unstructured tetrahedral mesh is partitioned into subdomains that contain an equal or at least very similar number of elements as shown in Fig. 13. Therefore, each processor has to carry out a similar amount of calculations. However, if  $p$ -adaptation is applied, the partitioning is more sophisticated as one subdomain might have many elements of high order polynomials whereas another might have the same number of elements but with lower order polynomials. Therefore, the parallel efficiency is re-

stricted by the processor with the highest work load. However, this problem can usually be solved by weighted partitioning algorithms, e. g. METIS.

In the case of local time stepping, mesh partitioning is becoming a much more difficult task. One solution is to divide the computational domain into a number of zones, that usually contain a geometrical body or a geological zone that typically is meshed individually with a particular mesh spacing  $h$  and contains a dominant polynomial order. Then each of these zones is partitioned separately into subdomains of approximately equal numbers of elements. Then each processor receives a subdomain of each zone, which requires a similar amount of computational work as shown in Fig. 13. In particular, the equal distribution of tetrahedra with different sizes is essential in combination with the local time stepping technique. Only if each processor receives subdomains with a similar amount of small and large elements, the work load is balanced. The large elements have to be updated less frequently than the smaller elements and therefore are computationally cheaper. Note, that the separately partitioned and afterwards merged zones lead to non-connected subdomains for each processor (see Fig. 13). This increases the number of element surfaces between subdomains of different processors and therefore increases the communication required. However, communication is typically low as the degrees of freedom have to be exchanged only once per time step and only for tetrahedra that have an interface at the boundary between subdomains. Therefore, the improvements due to the new load balancing approach are dominant and outweigh the increase in communication.

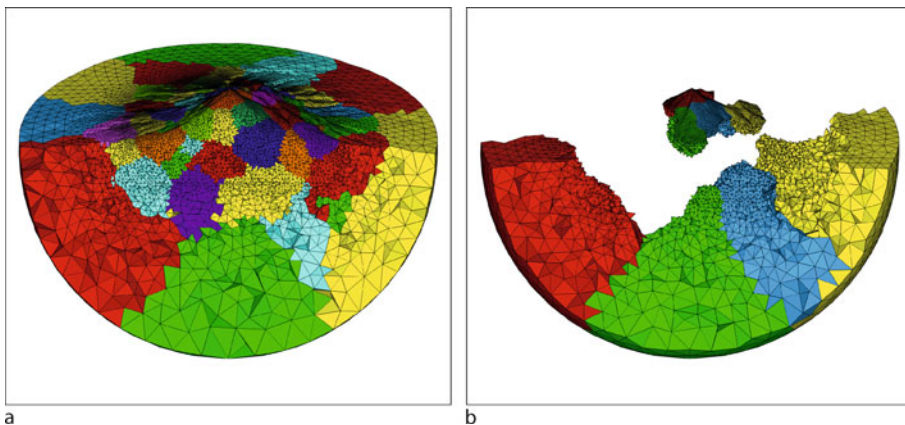
However, care has to be taken as the distribution of the polynomial degree  $p$  or the seismic velocity structure

might influence the efficiency of this grouped partitioning technique. A profound and thorough mesh partitioning method is still a pending task as the combination of local time stepping and  $p$ -adaptivity requires a new weighting strategy of the computational cost for each tetrahedral element considering also the asynchronous element update. The automatic partitioning of unstructured meshes with such heterogeneous properties together with the constraint of keeping the subdomains as compact as possible to avoid further increase of communication is still subject to future work.

In Fig. 13 an example of a grouped partition of the tetrahedral mesh is shown for 4 processors. Two non-connected subdomains indicated by the same color are assigned to each processor including small – and therefore computationally expensive – tetrahedra that are updated frequently due to their small time step, and much larger elements that typically are cheap due to their large time step. This way, the work load often is balanced sufficiently well over the different processors.

#### Relevance of High Performance Computing: Application to Merapi Volcano

In recent years the development of the ADER-DG algorithm including the high order numerical approximation in space and time, the mesh generation, mesh adaptation, parameterization, and data visualization created the basis of an efficient and highly accurate seismic simulation tool. Realistic large scale applications and their specific requirements will further guide these developments. On the other hand, the study and incorporation of geophysical processes that govern seismic wave propagation insures,



Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 13

Standard partitioning of the computational domain (left) and an example of 4 subdomains grouped together for more efficient local time stepping

that the simulation technology matches the needs and addresses latest challenges in modern computational seismology. Hereby, the accurate modeling of different source mechanisms as well as the correct treatment of realistic material properties like anelasticity, viscoplasticity, porosity and highly heterogeneous, scattering media will play an important role.

However, only the combination of this state-of-the-art simulation technology with the most powerful supercomputing facilities actually available can provide excellent conditions to achieve scientific progress for realistic, large scale applications. This combination of modern technologies will substantially contribute to resolve current problems, not only in numerical seismology, but will also influence other disciplines. The phenomenon of acoustic, elastic or seismic wave propagation is encountered in many different fields. Beginning with the classical geophysical sciences seismology, oceanography, and volcanology such waves also appear in environmental geophysics, atmospheric physics, fluid dynamics, exploration geophysics, aerospace engineering or even medicine.

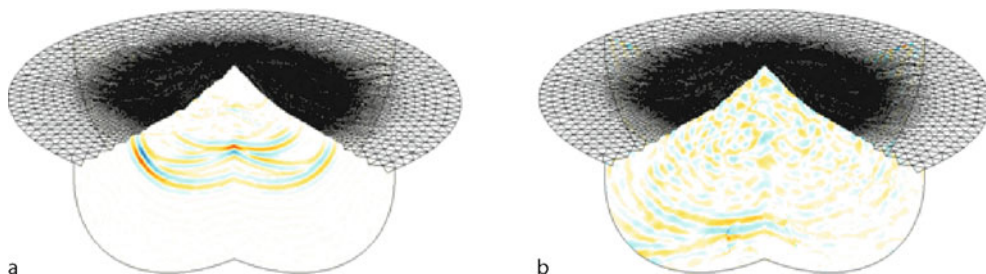
With the rapid development of modern computer technology and the development of new highly accurate simulation algorithms computer modeling just started to herald a new era in many applied sciences. The 3D wave propagation simulations in realistic media require a substantial amount of computation time even on large parallel computers. Extremely powerful national supercomputers already allow us to run simulations with unrivaled accuracy and resolution. However, using the extremely high accuracy and flexibility of new simulation methods on such massively parallel machines the professional support of experts in supercomputing is absolutely essential. Only professional porting, specific CPU-time and storage optimizations of current software with respect to continuously changing compilers, operating systems, hardware ar-

chitectures or simply personnel, will ensure the lifetime of new simulation technologies accompanied by ongoing improvements and further developments. Additionally, the expertise and support in the visualization of scientific results using technologies of Virtual Reality for full 3D models not only enhances the value of simulation results but will support data interpretation and awake great interest in the new technology within a wide research community.

As an example, volcano monitoring plays an increasingly important role in hazard estimation in many densely populated areas in the world. Highly accurate computer modeling today is a key issue to understand the processes and driving forces that can lead to dome building, eruptions or pyroclastic flows. However, data of seismic observations at volcanoes are often very difficult to interpret. Inverting for the source mechanism, i. e. seismic moment tensor inversion, or just locating an exact source position is often impossible due to the strongly scattered wave field caused by an extremely heterogeneous material distribution inside the volcano. Furthermore, the rough topography alone can affect the wave field by its strongly scattering properties as shown in Fig. 14.

Therefore, it is fundamental to understand the effects of topography and scattering media and their influence on the seismic wave field. A systematic study of a large number of scenarios computed by highly accurate simulation methods to provide reliable synthetic data sets is necessary to test the capabilities of currently used inversion tools. Slight changes in parameters like the source position, the source mechanism or the elastic and geometric properties of the medium can then reveal the limits of such tools and provide more precise bounds of their applicability in volcano seismology.

Finally, the implementation of the ADER-DG method is still much more expensive than other state-of-the-art implementations of existing methods. However, a fair



Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 14

Snapshots of the seismic wave field after an explosive event close to the summit of Merapi volcano. The free surface topography introduces strong scattering of the waves making it extremely difficult to invert for the seismic source mechanism or the exact source location

comparison between accuracy and computational cost is still a pending task. The main reason for the CPU-time difference is the much larger number of tetrahedral elements than hexahedra that have to be used to cover the same volume. Furthermore, due to the choice of the basis functions, the flux computations are expensive, as the matrix-matrix multiplications involved are not sparse.

However, the ADER-DG method is currently implemented on hexahedral meshes to make fair comparisons possible. Preliminary tests show, that the change of mesh topology from tetrahedra to hexahedra significantly reduces the computational cost. However, final results are subject to future investigations.

### Discussion and Future Directions

As indicated in the introduction and highlighted in the previous sections, computational tools for wave propagation problems are getting increasingly sophisticated to meet the needs of current scientific problems. We are far away from simple finite-difference time schemes that are solving problems on regular grids on serial computers in which case the particular programming approach did not affect dramatically the overall performance. Today, competitive algorithms are results of years of partly highly professional coding. Implementations on high-performance computing hardware requires in-depth knowledge of parallel algorithms, profiling, and many technical aspects of modern computations. To make complex scientific software available to other researchers requires implementation and testing on many different (parallel) platforms. This may involve parallelization using different programming paradigms (e. g., the combination of OpenMP and MPI on nodes of shared memory machines), and interoperability on heterogeneous computational GRIDs.

This has dramatic consequences particularly for young researchers in the Earth Sciences who want to use advanced computational tools to model observations. While in the early days a finite-difference type algorithm could be understood, coded, implemented and tested in a few weeks, this is no longer possible. In addition, standard curricula do not offer training in computational methods allowing them to efficiently write and test codes. This suggests that at least for some, well-defined computational problems verified and professionally engineered scientific software solutions should be provided to the community and also professionally extended and maintained in close collaboration with scientists. In seismology we are in a quite fortunate situation. In contrast to many other fields of physical sciences, our constitutive relations (e. g., stress-strain) are fairly well understood, and – as indicated

in this paper – numerical solutions for 3D problems and their implementation on parallel hardware are well advanced. Another argument for stable tested “community”-codes for wave propagation is the fact that advancement in many scientific problems (e. g., imaging the Earth’s interior, quantifying earthquake-induced shaking hazard) relies on zillions of forward modeling runs with only slight variations of the internal velocity models.

As far as technical developments are concerned, the efficient initialization of complex 3D models on computational grids is still a great challenge. Realistic models may be composed of complex topography, families of overlapping fault surfaces, discontinuous interfaces, and varying rheologies (e. g., elastic, anisotropic, viscoelastic, viscoplastic, porous). This may require the combination of tetrahedral and hexahedral grid in models with strongly varying degree of complexity. Ideally, standards for Earth models (and synthetic data) formats should be established by the communities that allow easy exchange and multiple use of models with different simulation tools (e. g., wave propagation, deformation, earthquake rupture). In addition, the rapid developments towards PetaFlop computing opens new questions about the scalability and efficient parallelization of current and future algorithms.

As the forward problem of wave propagation is at the core of the seismic imaging problem for both source and Earth’s structure, in the near future we will see the incorporation of 3D simulation technology into the imaging process. Provided that the background seismic velocity models are fairly well known (e. g., reservoirs, global Earth, sedimentary basins), adjoint methods provide a powerful analytical tool to (1) relate model deficiencies to misfit in observations and (2) quantify the sensitivities to specific aspects of the observations (e. g., [100,104,105]). As the core of the adjoint calculations is the seismic forward problem, the challenge is the actual application to real data and the appropriate parametrizations of the model and the data that optimize the data fitting process.

In summary, while we look back at (and forward to) exciting developments in computational seismology, a paradigm shift in the conception of one of the central tools of seismology – the calculation of 3D synthetic seismograms – is necessary. To extract a maximum amount of information from our high-quality observations scientists should have access to high-quality simulation tools. It is time to accept that “*software is infrastructure*” and provide the means to professionally develop and maintain community codes and model libraries at least for basic Earth science problems and specific focus regions. Developments are one the way along those lines in the SPICE project (Seismic Wave Propagation and Imaging in Complex Me-

dia, a European Network [101]), the Southern California Earthquake Center (SCEC [94]) and the CIG Project (Computational infrastructure in geodynamics [17]).

### Acknowledgments

We would like to acknowledge partial support towards this research from: The European Human Resources and Mobility Program (SPICE-Network), the German Research Foundation (Emmy Noether-Programme), the Bavarian Government (KOHNIWIHR, graduate college THESIS, BaCaTec), and MunichRe. We would also like to thank J. Tromp supporting MS's visit to CalTech. We also thank two anonymous reviewers for constructive comments on the manuscript.

### Bibliography

#### Primary Literature

- Alterman Z, Karal FC (1968) Propagation of elastic waves in layered media by finite-difference methods. *Bull Seism Soc Am* 58:367–398
- Benzley SE, Harris NJ, Scott M, Borden M, Owen SJ (2005) Conformal refinement and coarsening of unstructured hexahedral meshes. *J Comput Inf Sci Eng* 5:330–337
- Bey J (1995) Tetrahedral grid refinement. *Computing* 55:355–378
- Bielak J, Loukakis K, Hisada Y, Yoshimura C (2003) Domain reduction method for three-dimensional earthquake modeling in localized regions, Part I: Theory. *Bull Seism Soc Am* 93:817–824
- Bonilla LF, Archuleta RJ, Lavallée D (2005) Hysteretic and dilatant behavior of cohesionless soils and their effects on non-linear site response: Field data observations and modelling. *Bull Seism Soc Am* 95(6):2373–2395
- Boore D (1972) Finite-difference methods for seismic wave propagation in heterogeneous materials. In: Bolt BA (ed) *Methods in Computational Physics*, vol 11. Academic Press, New York
- Braun J, Sambridge MS (1995) A numerical method for solving partial differential equations on highly irregular evolving grids. *Nature* 376:655–660
- Bunge HP, Tromp J (2003) Supercomputing moves to universities and makes possible new ways to organize computational research. *EOS* 84(4):30, 33
- Carcione JM, Wang J-P (1993) A Chebyshev collocation method for the elastodynamic equation in generalised coordinates. *Comp Fluid Dyn* 2:269–290
- Carcione JM, Kosloff D, Kosloff R (1988) Viscoacoustic wave propagation simulation in the earth. *Geophysics* 53:769–777
- Carcione JM, Kosloff D, Behle A, Seriani G (1992) A spectral scheme for wave propagation simulation in 3-D elastic-anisotropic media. *Geophysics* 57:1593–1607
- Carey G (1997) *Computational grids: Generation, adaptation, and solution strategies*. Taylor Francis, New York
- Cerveny V (2001) *Seismic ray theory*. Cambridge University Press, Cambridge
- Chaljub E, Tarantola A (1997) Sensitivity of SS precursors to topography on the upper-mantle 660-km discontinuity. *Geophys Res Lett* 24(21):2613–2616
- Chaljub E, Komatitsch D, Vilotte JP, Capdeville Y, Valette B, Festa G (2007) Spectral element analysis in seismology. In: Wu R-S, Maupin V (eds) *Advances in wave propagation in heterogeneous media*. *Advances in Geophysics*, vol 48. Elsevier, London, pp 365–419
- Chapman CH (2004) *Fundamentals of seismic wave propagation*. Cambridge University Press, Cambridge
- CIG [www.geodynamics.org](http://www.geodynamics.org). Accessed 1 Jul 2008
- Cockburn B, Shu CW (1989) TVB Runge–Kutta local projection discontinuous Galerkin finite element method for conservation laws II: General framework. *Math Comp* 52:411–435
- Cockburn B, Shu CW (1991) The Runge–Kutta local projection P1-Discontinuous Galerkin finite element method for scalar conservation laws. *Math Model Numer Anal* 25:337–361
- Cockburn B, Shu CW (1998) The Runge–Kutta discontinuous Galerkin method for conservation laws V: Multidimensional systems. *J Comput Phys* 141:199–224
- Cockburn B, Lin SY, Shu CW (1989) TVB Runge–Kutta local projection discontinuous Galerkin finite element method for conservation laws III: One dimensional systems. *J Comput Phys* 84:90–113
- Cockburn B, Hou S, Shu CW (1990) The Runge–Kutta local projection discontinuous Galerkin finite element method for conservation laws IV: The multidimensional case. *Math Comp* 54:545–581
- Cockburn B, Karniadakis GE, Shu CW (2000) *Discontinuous Galerkin methods, theory, computation and applications*. LNCSE, vol 11. Springer, New York
- Courant R, Friedrichs KO, Lewy H (1928) Über die partiellen Differenzialgleichungen der mathematischen Physik. *Mathematische Annalen* 100:32–74
- CUBIT [cubit.sandia.gov](http://cubit.sandia.gov). Accessed 1 Jul 2008
- Dablain MA (1986) The application of high-order differencing to the scalar wave equation. *Geophysics* 51:54–66
- De Cougny HL, Shephard MS (1999) Parallel refinement and coarsening of tetrahedral meshes. *Int J Numer Methods Eng* 46:1101–1125
- de la Puente J, Dumbser M, Käser M, Igel H (2007) Discontinuous Galerkin methods for wave propagation in poroelastic media. to appear in *Geophysics*
- de la Puente J, Käser M, Dumbser M, Igel H (2007) An arbitrary high order discontinuous Galerkin method for elastic waves on unstructured meshes IV: Anisotropy. *Geophys J Int* 169(3):1210–1228
- di Prisco C, Stupazzini M, Zambelli C (2007) Non-linear SEM numerical analyses of dry dense sand specimens under rapid and dynamic loading. *Int J Numer Anal Methods Geomech* 31(6):757–788
- Dormy E, Tarantola A (1995) Numerical simulation of elastic wave propagation using a finite volume method. *J Geophys Res* 100(B2):2123–2134
- Dumbser M (2005) Arbitrary high order schemes for the solution of hyperbolic conservation laws in complex domains. Shaker, Aachen
- Dumbser M, Käser M (2006) An arbitrary high order discontinuous galerkin method for elastic waves on unstructured meshes II: The three-dimensional isotropic case. *Geophys J Int* 167:319–336

34. Dumbser M, Käser M (2007) Arbitrary high order non-oscillatory finite volume schemes on unstructured meshes for linear hyperbolic systems. *J Comput Phys* 221:693–723. doi:10.1016/j.jcp.2006.06.043
35. Dumbser M, Munz CD (2005) Arbitrary high order discontinuous Galerkin schemes. In: Cordier S, Goudon T, Gutnic M, Sonnendruker E (eds) *Numerical methods for hyperbolic and kinetic problems*. IRMA series in mathematics and theoretical physics. EMS Publishing, Zurich, pp 295–333
36. Dumbser M, Käser M, Toro EF (2007) An arbitrary high-order discontinuous Galerkin method for elastic waves on unstructured meshes – V. Local time stepping and p-adaptivity. *Geophys J Int* 171:695–717
37. Dziewonski AM, Anderson DL (1981) Preliminary reference earth model. *Phys Earth Planet Inter* 25:297–356
38. Ewald M, Igel H, Hinzen K-G, Scherbaum F (2006) Basin-related effects on ground motion for earthquake scenarios in the lower rhine embayment. *Geophys J Int* 166:197–212
39. Faccioli E, Maggio F, Quarteroni A, Tagliani A (1996) Spectral-domain decomposition methods for the solution of acoustic and elastic wave equation. *Geophysics* 61:1160–1174
40. Faccioli E, Maggio F, Paolucci R, Quarteroni A (1997) 2D and 3D elastic wave propagation by a pseudo-spectral domain decomposition method. *J Seismol* 1:237–251
41. Faccioli E, Vanini M, Paolucci R, Stupazzini M (2005) Comment on “Domain reduction method for three-dimensional earthquake modeling in localized regions, part I: Theory.” by Bielak J, Loukakis K, Hisada Y, Yoshimura C, and “Part II: Verification and Applications.” by Yoshimura C, Bielak J, Hisada Y, Fernández A. *Bull Seism Soc Am* 95:763–769
42. Falk J, Tessmer E, Gajewski D (1996) Efficient finite-difference modelling of seismic waves using locally adjustable time steps. *Geophys Prosp* 46:603–616
43. Falk J, Tessmer E, Gajewski D (1996) Tube wave modelling by the finite differences method with varying grid spacing. *Pure Appl Geoph* 148:77–93
44. Fernandez A, Bielak J, Prentice C (2006) Urban seismology; City effects on earthquake ground motion and effects of spatial distribution of ground motion on structural response paper presented at 2006 annual meeting. *Seism Res Lett* 77(2):305
45. Fornberg B (1996) *A practical guide to pseudospectral methods*. Cambridge University Press, Cambridge
46. Fuchs K, Müller G (1971) Computation of synthetic seismograms with the reflectivity method and comparison with observations. *Geophys J Royal Astronom Soc* 23(4):417–33
47. Furumura T, Takenaka H (1996) 2.5-D modeling of elastic waves using the pseudospectral method. *Geophys J Int* 124:820–832
48. Geller RJ, Takeuchi N (1998) Optimally accurate second-order time-domain finite difference scheme for the elastic equation of motion: One-dimensional case. *Geophys J Int* 135:48–62
49. Graves RW (1993) Modeling three-dimensional site response effects in the Marina district basin, San Francisco, California. *Bull Seism Soc Am* 83:1042–1063
50. Hestholm SO, Ruud BO (1998) 3-D finite-difference elastic wave modeling including surface topography. *Geophysics* 63:613–622
51. Holberg O (1987) Computational aspects of the choice of operator and sampling interval for numerical differentiation in large-scale simulation of wave phenomena. *Geophys Prospect* 35:629–655
52. Igel H (1999) Wave propagation through 3-D spherical sections using the Chebyshev spectral method. *Geop J Int* 136:559–567
53. Igel H, Gudmundsson O (1997) Frequency-dependent effects on travel times and waveforms of long-period S and SS waves. *Phys Earth Planet Inter* 104:229–246
54. Igel H, Weber M (1995) SH-wave propagation in the whole mantle using high-order finite differences. *Geophys Res Lett* 22(6):731–734
55. Igel H, Weber M (1996) P-SV wave propagation in the Earth’s mantle using finite-differences: Application to heterogeneous lowermost mantle structure. *Geophys Res Lett* 23:415–418
56. Igel H, Mora P, Rioulet B (1995) Anisotropic wave propagation through finite-difference grids. *Geophysics* 60:1203–1216
57. Igel H, Nissen-Meyer T, Jahnke G (2001) Wave propagation in 3-D spherical sections: Effects of subduction zones. *Phys Earth Planet Inter* 132:219–234
58. Jahnke G, Igel H, Cochard A, Thorne M (2007) Parallel implementation of axisymmetric SH wave propagation in spherical geometry. *Geophys J Int* (in print)
59. Jastram C, Tessmer E (1994) Elastic modelling on a grid with vertically varying spacing. *Geophys Prosp* 42:357–370
60. Karypis G, Kumar V (1998) Multilevel k-way Partitioning Scheme for Irregular Graphs. *J Parallel Distrib Comput* 48(1):96–129
61. Kelly KR, Ward RW, Treitel S, Alford RM (1976) Synthetic seismograms: A finite-difference approach. *Geophysics* 41:2–27
62. Kennett BLN (2002) *The seismic wavefield, vol I + II*. Cambridge University Press, Cambridge
63. Komatitsch D, Tromp J (2002) Spectral-element simulations of global seismic wave propagation, part I: Validation. *Geophys J Int* 149:390–412
64. Komatitsch D, Tromp J (2002) Spectral-element simulations of global seismic wave propagation, part II: 3-D models, oceans, rotation, and gravity. *Geophys J Int* 150:303–318
65. Komatitsch D, Vilotte JP (1998) The spectral-element method: An efficient tool to simulate the seismic response of 2D and 3D geological structures. *Bull Seism Soc Am* 88:368–392
66. Komatitsch D, Couteil F, Mora P (1996) Tensorial formulation of the wave equation for modelling curved interfaces. *Geophys J Int* 127(1):156–168
67. Kosloff D, Baysal E (1982) Forward modeling by a fourier method. *Geophysics* 47(10):1402–1412
68. Krishnan S, Ji C, Komatitsch D, Tromp J (2006) Case studies of damage to tall steel moment-frame buildings in Southern California during large San Andreas earthquakes. *Bull Seismol Soc Am* 96(4A):1523–1537
69. Krishnan S, Ji C, Komatitsch D, Tromp J (2006) Performance of two 18-story steel moment-frame buildings in Southern California during two large simulated San Andreas earthquakes. *Earthq Spectra* 22(4):1035–106
70. Käser M, Dumbser M (2006) An arbitrary high order discontinuous Galerkin method for elastic waves on unstructured meshes I: The two-dimensional isotropic case with external source terms. *Geophys J Int* 166:855–877
71. Käser M, Dumbser M, de la Puente J, Igel H (2007) An arbitrary high order discontinuous Galerkin method for elastic waves



- on unstructured meshes III: Viscoelastic attenuation. *Geophys J Int* 168(1):224–242
72. Käser M, Igel H (2001) Numerical simulation of 2D wave propagation on unstructured grids using explicit differential operators. *Geophys Prospect* 49(5):607–619
  73. Käser M, Igel H, Sambridge M, Braun J (2001) A comparative study of explicit differential operators on arbitrary grids. *J Comput Acoust* 9(3):1111–1125
  74. Kwak D-Y, Im Y-T (2002) Remeshing for metal forming simulations – part II: Three dimensional hexahedral mesh generation. *Int J Numer Methods Eng* 53:2501–2528
  75. LeVeque RL (2002) *Finite volume methods for hyperbolic problems*. Cambridge University Press, Cambridge
  76. Levander AR (1988) Fourth-order finite-difference P-SV seismograms. *Geophysics* 53:1425–1436
  77. Madariaga R (1976) Dynamics of an expanding circular fault. *Bull Seismol Soc Am* 66(3):639–66
  78. Magnier S-A, Mora P, Tarantola A (1994) Finite differences on minimal grids. *Geophysics* 59:1435–1443
  79. Marfurt KJ (1984) Accuracy of finite-difference and finite-element modeling of the scalar and elastic wave equations. *Geophysics* 49:533–549
  80. Mercierat ED, Vilotte JP, Sanchez-Sesma FJ (2006) Triangular spectral element simulation of two-dimensional elastic wave propagation using unstructured triangular grids. *Geophys J Int* 166(2):679–698
  81. METIS [glaros.dtc.umn.edu/gkhome/views/metis](http://glaros.dtc.umn.edu/gkhome/views/metis). Accessed 1 Jul 2008
  82. Moczo P (1989) Finite-difference techniques for SH-waves in 2-D media using irregular grids – Application to the seismic response problem. *Geophys J Int* 99:321–329
  83. Moczo P, Kristek J, Halada L (2000) 3D 4th-order staggered grid finite-difference schemes: Stability and grid dispersion. *Bull Seism Soc Am* 90:587–603
  84. Montelli R, Nolet G, Dahlen FA, Masters G, Engdahl ER, Hung S (2004) Finite-frequency tomography reveals a variety of plumes in the mantle. *Science* 303(5656):338–343
  85. Müller G (1977) Earth-flattening approximation for body waves derived from geometric ray theory – improvements, corrections and range of applicability. *J Geophys* 42:429–436
  86. Nissen-Meyer T, Fournier A, Dahlen FA (2007) A 2-D spectral-element method for computing spherical-earth seismograms – I. Moment-tensor source. *Geophys J Int* 168:1067–1092
  87. Ohminato T, Chouet BA (1997) A free-surface boundary condition for including 3D topography in the finite-difference method. *Bull Seism Soc Am* 87:494–515
  88. Opršal I, Zahradník (1999) Elastic finite-difference method for irregular grids. *Geophysics* 64:240–250
  89. Pitarka A (1999) 3D elastic finite-difference modeling of seismic motion using staggered grids with nonuniform spacing. *Bull Seism Soc Am* 89:54–68
  90. Priolo E, Carcione JM, Seriani G (1996) Numerical simulation of interface waves by high-order spectral modeling techniques. *J Acoust Soc Am* 95:681–693
  91. Reed WH, Hill TR (1973) *Triangular mesh methods for the neutron transport equation*. Technical Report, LA-UR-73-479, Los Alamos Scientific Laboratory
  92. Ripperger J, Igel H, Wassermann J (2004) Seismic wave simulation in the presence of real volcano topography. *J Volcanol Geotherm Res* 128:31–44
  93. Scandella L (2007) Numerical evaluation of transient ground strains for the seismic response analyses of underground structures. Ph D Thesis, Milan University of Technology, Milan
  94. SCEC [www.scec.org](http://www.scec.org). Accessed 1 Jul 2008
  95. Schneiders R (2000) Octree-Based Hexahedral Mesh Generation. *Int J Comput Geom Appl* 10(4):383–398
  96. Schwartzkopff T, Munz CD, Toro EF (2002) ADER: A high-order approach for linear hyperbolic systems in 2D. *J Sci Comput* 17:231–240
  97. Schwartzkopff T, Dumbser M, Munz CD (2004) Fast high order ADER schemes for linear hyperbolic equations. *J Comput Phys* 197:532–539
  98. Seriani G, Priolo E, Carcione JM, Padovani E (1992) High-order spectral element method for elastic wave modeling: 62nd Ann. Internat. Mtg., Soc. Expl. Geophys., Expanded Abstracts, 1285–1288
  99. Shepherd JF (2007) Topologic and geometric constraint-based hexahedral mesh generation. Ph.D. Thesis on Computer Science, School of Computing The University of Utah, Salt Lake City
  100. Sieminski A, Liu Q, Trampert J, Tromp J (2007) Finite-frequency sensitivity of surface waves to anisotropy based upon adjoint methods. *Geophys J Int* 168:1153–1174
  101. SPICE [www.spice-rtn.org](http://www.spice-rtn.org). Accessed 1 Jul 2008
  102. Stupazzini M (2004) A spectral element approach for 3D dynamic soil-structure interaction problems. Ph D Thesis, Milan University of Technology, Milan
  103. Takeuchi N, Geller RJ (2000) Optimally accurate second order time-domain finite difference scheme for computing synthetic seismograms in 2-D and 3-D media. *Phys Earth Planet Int* 119:99–131
  104. Tape C, Liu Q, Tromp J (2007) Finite-frequency tomography using adjoint methods: Methodology and examples using membrane surface waves. *Geophys J Int* 168:1105–1129
  105. Tarantola A (1986) A strategy for nonlinear elastic inversion of seismic reflection data. *Geophysics* 51(10):1893–1903
  106. Tessmer E (2000) Seismic finite-difference modeling with spatially varying time steps. *Geophysics* 65:1290–1293
  107. Tessmer K, Kosloff D (1996) 3-D elastic modeling with surface topography by a Chebyshev spectral method. *Geophysics* 59:464–473
  108. Tessmer E, Kessler D, Kosloff K, Behle A (1996) Multi-domain Chebyshev–Fourier method for the solution of the equations of motion of dynamic elasticity. *J Comput Phys* 100:355–363
  109. Thomas C, Igel H, Weber M, Scherbaum F (2000) Acoustic simulation of P-wave propagation in a heterogeneous spherical earth: Numerical method and application to precursor energy to PKP. *Geophys J Int* 141:307–320
  110. Thorne M, Lay T, Garnero E, Jahnke G, Igel H (2007) 3-D seismic imaging of the D'' region beneath the Cocos Plate. *Geophys J Int* 170:635–648
  111. Titarev VA, Toro EF (2002) ADER: Arbitrary high order Godunov approach. *J Sci Comput* 17:609–618
  112. Toro EF (1999) *Riemann solvers and numerical methods for fluid dynamics*. Springer, Berlin
  113. Toro EF, Millington AC, Nejad LA (2001) Towards very high order Godunov schemes, in *Godunov methods; Theory and applications*. Kluwer/Plenum, Oxford, pp 907–940
  114. Toyokuni G, Takenaka H, Wang Y, Kennett BLN (2005) Quasi-spherical approach for seismic wave modeling in a 2-D slice

- of a global earth model with lateral heterogeneity. *Geophys Res Lett* 32:L09305
115. Van der Hilst RD (2004) Changing views on Earth's deep mantle. *Science* 306:817–818
116. Virieux J (1984) SH-wave propagation in heterogeneous media: Velocity-stress finite-difference method. *Geophysics* 49:1933–1957
117. Virieux J (1986) P-SV wave propagation in heterogeneous media: Velocity-stress finite-difference method. *Geophysics* 51:889–901
118. Woodhouse JH, Dziewonski AM (1984) Mapping the upper mantle: Three dimensional modelling of earth structure by inversion of seismic waveforms. *J Geophys Res* 89:5953–5986
119. Yoshimura C, Bielak J, Hisada Y, Fernández A (2003) Domain reduction method for three-dimensional earthquake modeling in localized regions, part II: Verification and applications. *Bull Seism Soc Am* 93:825–841
120. Zambelli C (2006) Experimental and theoretical analysis of the mechanical behaviour of cohesionless soils under cyclic-dynamic loading. Ph D Thesis, Milan University of Technology, Milan
121. Zienkiewicz O, Taylor RL (1989) *The finite element method*, vol 1. McGraw-Hill, London

### Books and Reviews

- Carcione JM, Herman GC, ten Kroode APE (2002) Seismic modelling. *Geophysics* 67:1304–1325
- Mozco P, Kristek J, Halada L (2004) *The finite-difference method for seismologists: An introduction*. Comenius University, Bratislava. Available in pdf format at <ftp://ftp.nuquake.eu/pub/Papers>
- Mozco P, Kristek J, Galis M, Pazak P, Balazovjeh M (2007) The finite difference and finite-element modelling of seismic wave propagation and earthquake motion. *Acta Physica Slovaca*, 57(2)177–406
- Wu RS, Maupin V (eds) (2006) *Advances in wave propagation in heterogeneous earth*. In: Dmowska R (ed) *Advances in geophysics*, vol 48. Academic/Elsevier, London

## Seismic Waves in Heterogeneous Earth, Scattering of

HARUO SATO

Department of Geophysics, Graduate School of Science,  
Tohoku University, Sendai-shi, Miyagi-ken, Japan

### Article Outline

Glossary

Definition of the Subject

Introduction

Radiative Transfer Theory for a Scattering Medium

Wave Envelopes in Random Media  
and Statistical Characterization

Envelope Broadening of a High-Frequency Seismogram

Spatial Variation of Scattering Characteristics

Temporal Change in the Earth Medium Structure

Future Directions

Acknowledgments

Bibliography

### Glossary

**Attenuation factor  $Q^{-1}$**  A measure of attenuation characteristics of a medium caused by intrinsic absorption and scattering loss. The former means the transfer of vibration energy into heat and the latter means the transfer of vibration energy from the direct wave to coda waves caused by scattering due to medium heterogeneity.

**Coda waves** Wave trains that follow the arrival of the direct S-wave phase are called S-coda waves or simply coda waves. Coda waves are interpreted as a superposition of S waves scattered by distributed heterogeneities. Wave trains between direct P and S wave arrivals are called P-coda waves.

**Coda attenuation factor  $Q_C^{-1}$**  This parameter characterizes the amplitude decay of S coda of a local earthquake with the lapse time increasing based on the S-to-S single scattering. The coda duration shortens for a larger coda attenuation factor.

**Envelope broadening** The source duration time of a microearthquake is short; however, the apparent duration time of the S-wave seismogram increases with the travel distance increasing because of diffraction and scattering by medium heterogeneities. This phenomenon is called envelope broadening.

**Radiative transfer theory** A phenomenological theory that describes scattering process of wave energy in a scattering medium on the basis of causality, geomet-

rical spreading and the energy conservation. It neglects the interference of waves but focuses on the intensity only. This theory admits various types of scattering patterns. It is often applied to model the energy propagation of high-frequency seismic-waves in heterogeneous Earth media.

**Random media** A mathematical model for media whose parameters are described by random functions of space coordinates. The stochastic properties of the ensemble of random media are characterized by their autocorrelation function or the power spectral density function.

**Scattering coefficient  $g$**  A measure of the scattering power in a unit solid angle at a certain direction by a unit volume of heterogeneous media for the incidence of unit energy flux density. The average of  $g$  over the solid angle gives the total scattering coefficient  $g_0$ , of which the reciprocal gives the mean free path. This quantity characterizes the coda excitation and the scattering loss in the heterogeneous media.

### Definition of the Subject

The structure of the solid Earth was extensively studied by using seismic waves such as travel time analysis based on Snell's law, dispersion analysis of surface waves, and spectral analysis of free oscillation, where the notion of a horizontally stratified structure or a spherical shell structure prevailed among the geophysical community. This means the acceptance of the dominance of gravity in geodynamic process. Velocity tomography revealed that the solid Earth structure is three-dimensionally inhomogeneous with various ranges of scales; however, the resolution of velocity tomography is much coarser than the wavelength of seismic waves. In 1970s, the existence of distributed inhomogeneities having the order of the wavelength of seismic waves was recognized from the observation of coda waves of local earthquakes, which are long-lasting wave trains following the direct S-wave arrival in high-frequency seismograms. Here, we use "high-frequency" for frequency higher than about 1 Hz. The long duration time of coda waves can be interpreted as a direct evidence of wide-angle scattering caused by distributed small-scale heterogeneities since the source duration time is generally very short. S-wave seismograms of microearthquakes show broadened envelopes with travel distance increasing. This envelope broadening phenomenon is also an evidence of scattering around the forward direction due to random velocity inhomogeneities.

Since then, focusing on the frequency dependence of seismogram envelopes, geophysicists have extensively

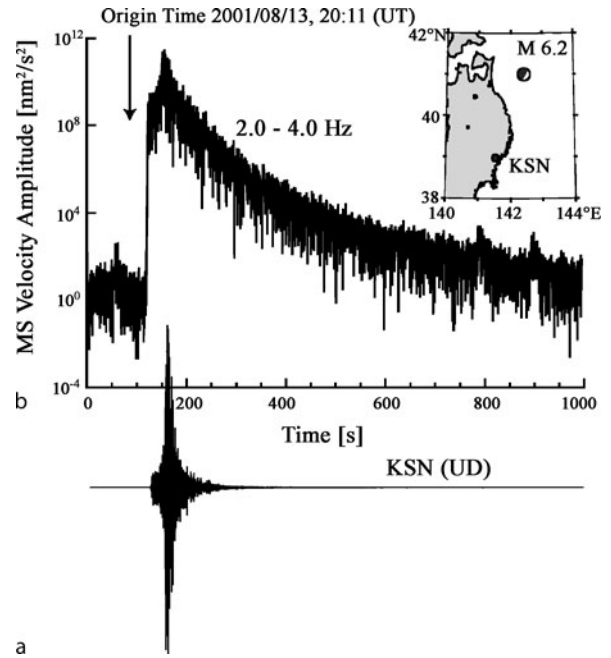
studied the scattering process of high-frequency seismic waves in relation to the spectral structure of velocity inhomogeneities, where the statistical characterization of the medium inhomogeneity is inevitable. The radiative transfer theory and the stochastic Markov approximation have been developed as mathematical tools for the analyzes of seismogram envelopes. The strength of scattering and/or the spectral structure of random inhomogeneity have been measured in various regions of the solid Earth. The scattering approach is found to be also useful for detecting temporal changes in the crustal medium associated with earthquake occurrences. Thus, scattering of high-frequency seismic waves in the heterogeneous Earth medium is important for understanding the physical structure and the geodynamic process reflecting the evolution of the solid Earth.

## Introduction

### Coda Waves

The high-frequency seismogram of a local earthquake has a long tail after the direct S-coda arrival. The tail portion of seismogram is called “S-coda waves” or simply “coda”. As an example, Fig. 1a and b show the raw seismogram and the band-pass filtered mean square (MS) trace of an earthquake of magnitude (M) 6.1, respectively. We note that the mean square wave envelope, which is the running mean of the squared trace with characteristic time of a few times the center period, is proportional to the time trace of the wave energy density. The coda wave oscillation lasts more than several hundreds of seconds. The duration of coda waves measured from the P-wave onset until when the coda amplitude decreases to the microseism’s level has been used as a quick measure of the earthquake magnitude from a single station observation since the 1960s. Having a motivation to extract the source spectrum of a large earthquake from clipped seismograms, Aki [1] first studied the characteristics of coda waves as scattered waves. Coda envelopes of a local earthquake have a smoothly decaying common curve with lapse time increasing irrespective of epicentral distances and the source radiation pattern. Aki and Chouet [4] interpreted coda waves as single back-scattered S-waves due to heterogeneities randomly distributed in the lithosphere. Their model based on the radar equation for the same location of a source and a receiver can be written as follows.

Point-like isotropic scatterers characterized by total scattering cross-section  $\sigma_0$  are randomly and homogeneously distributed with number density  $n$  in a 3-D medium with background wave velocity  $V_0$ . The scattering power per unit volume is characterized by the total scatter-



Seismic Waves in Heterogeneous Earth, Scattering of, Figure 1  
**a** Seismogram of a local earthquake of M 6.2 in northeastern Honshu, Japan recorded by F-net, NIED. **b** Bandpass-filtered MS envelope (Courtesy of T. Maeda)

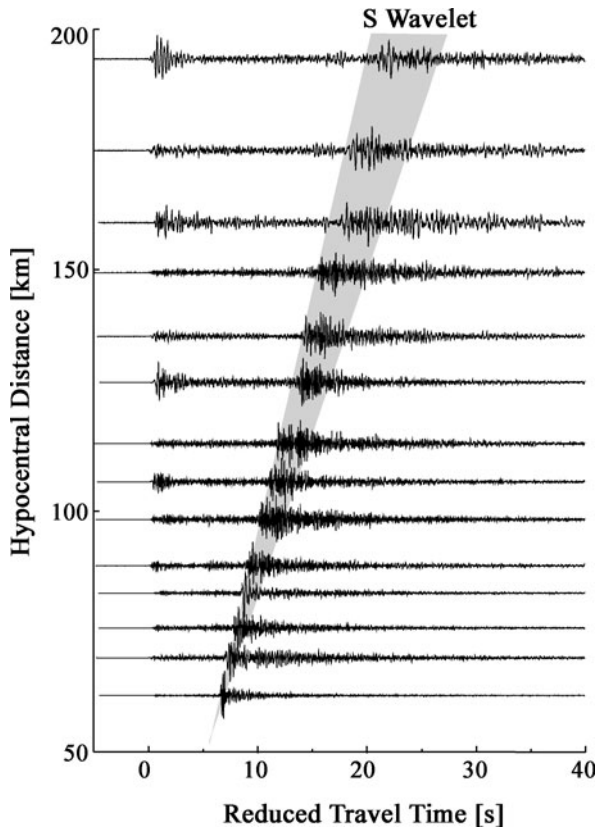
ing coefficient  $g_0 = n\sigma_0$ , of which the reciprocal gives the mean free path. When the total wave energy  $W$  is impulsively radiated from a point source at time  $t = 0$ , the wave energy density of singly back-scattered waves at the source location is written as

$$E^{\text{SB}}(t) \approx \frac{Wg_0}{2\pi V_0^2 t^2} e^{-Q_C^{-1} 2\pi ft} \quad (1)$$

since the interference of scattered waves can be neglected because of the random distribution of scatterers. The inverse square of lapse time means geometrical spreading in a 3-D space. Here, an exponential damping term with coda attenuation factor  $Q_C^{-1}$  is introduced to represent phenomenological attenuation effect. This simple formula has been widely used for measurements of  $g_0$  and  $Q_C^{-1}$  for S-waves in the world since the 1980s. Reported  $g_0$ -values are of the order of  $0.01 \text{ km}^{-1}$  for 1–30 Hz and  $Q_C^{-1}$  values are about  $10^{-2}$  at 1 Hz and decrease to about  $10^{-3}$  at 20 Hz in the lithosphere [74].

### Envelope Broadening of S-Seismogram

There is another evidence of scattering due to random inhomogeneity in high-frequency seismograms. Observed S-seismograms of a microearthquake have broadened en-



Seismic Waves in Heterogeneous Earth, Scattering of, Figure 2  
Envelope broadening shown in horizontal component seismograms of a microearthquake with M 4.0 in Japan recorded by Hi-net, NIED, where the abscissa is reduced travel time with move-out velocity 7 km/s (Courtesy of T. Takahashi)

velopes around their peaks after the direct arrivals. As shown by an example in Fig. 2, the apparent duration time of the S-seismogram just after the direct S-arrival increases with increasing travel distance. It is more than ten seconds at distances larger than 100 km, where the source duration time is less than one second for an earthquake of M 4.0. Sato [70] called this phenomenon observed in an island arc as “envelope broadening”, and Atkinson [6] reported similar phenomenon in a continent. For P-waves of teleseismic events, broadening of the vertical component envelope [35] and the excitation in the transverse component amplitude [52] have been used as a measure of lithospheric heterogeneity. These phenomena can be interpreted by multiple scattering within a narrow angle around the global ray direction due to random velocity inhomogeneities. When the wavelength is much shorter than the characteristic scale of the random velocity inhomogeneity, the scattering process of waves can be repre-

sented by successive ray bending processes, where scattering angles are statistically controlled by the spectrum of random velocity inhomogeneity. At a given distance from the source, a small number of rays with large scattering angles arrive long after the direct ray.

## Radiative Transfer Theory for a Scattering Medium

### Radiative Transfer Integral Equation for the Isotropic Scattering Process

Disregarding wave interference and focusing on wave power, the radiative transfer theory [8] treats the propagation of wave energy in a scattering medium. Wu [85] first introduced the radiative transfer theory for the stationary state in the synthesis of seismogram envelopes. The non-stationary multiple isotropic scattering process in 1-D was solved by Hemmer [24] and that in 2-D was solved by Shang and Gao [77]. Later, Zeng et al. [92] formulated the time-dependent multiple isotropic scattering process in 3-D as an extension of the single backscattering model [4] as follows.

In a 3-D scattering medium characterized by background velocity  $V_0$  and total scattering coefficient  $g_0$ , when the total wave energy  $W$  is impulsively radiated isotropically from a source at the origin, the multiple isotropic scattering process is written by the following integral equation for energy density:

$$E(\mathbf{x}, t) = WG_E(\mathbf{x}, t) + g_0 V_0 \cdot \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} G_E(\mathbf{x} - \mathbf{x}', t - t') E(\mathbf{x}', t') d\mathbf{x}' dt', \quad (2)$$

where the convolution integral in the second term means the propagation of energy from the last scattering point  $\mathbf{x}'$  to a receiver at  $\mathbf{x}$ . The first term is the ballistic term that means the direct propagation of energy from the source with scattering loss,  $G_E(\mathbf{x}, t) = \delta(t - r/V_0) \exp(-g_0 V_0 t) / (4\pi V_0 r^2)$ , where  $r = |\mathbf{x}|$ . The solution is written as [92]

$$E(\mathbf{x}, t) = WG_E(\mathbf{x}, t) + \frac{Wg_0 e^{-g_0 V_0 t}}{4\pi r^2} \frac{r}{V_0 t} \cdot \ln \left[ \frac{V_0 t + r}{V_0 t - r} \right] H \left( t - \frac{r}{V_0} \right) + Wg_0^2 V_0^2 \frac{1}{(2\pi)^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} d\omega dk e^{-i\omega t - ikr} \cdot \frac{ik}{2\pi r} \frac{\bar{G}_E(-k, -i\omega)^3}{1 - g_0 V_0 \bar{G}_E(-k, -i\omega)}, \quad (3)$$

where  $\bar{G}_E(k, s) = (1/kV_0) \tan^{-1} kV_0/(s + g_0V_0)$  is the Fourier–Laplace transform of  $G_E$  with respect to coordinate and time, respectively. The second term represents the single scattering process (see Sato [67]), which has a logarithmic divergence at the direct arrival  $t = r/V_0$  and decreases according to the inverse square of lapse time at long lapse times as  $Wg_0/(2\pi V_0^2 t^2)$ . The third term representing multiple scattering converges to a diffusion solution with lapse time increasing as

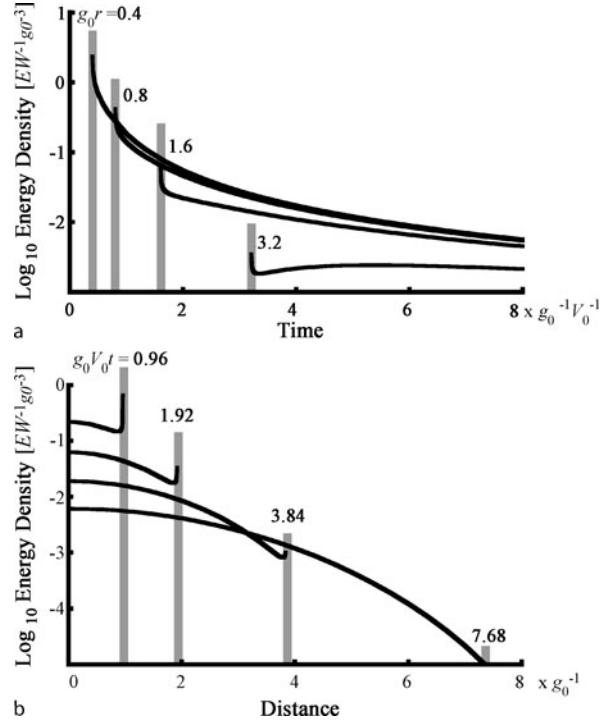
$$E^{\text{Dif.}}(\mathbf{x}, t) = W \left( \frac{3g_0}{4\pi V_0 t} \right)^{3/2} e^{-\frac{3g_0 r^2}{4V_0 t}} H(t), \quad (4)$$

where the factor  $V_0/(3g_0)$  is the diffusivity. Later, Paaschens [54] proposed an approximation as

$$E^{\text{Paa.}}(\mathbf{x}, t) \approx W G_E(\mathbf{x}, t) + \frac{W e^{-g_0 V_0 t}}{(4\pi V_0 t / (3g_0))^{3/2}} \cdot \left( 1 - \frac{r^2}{V_0^2 t^2} \right)^{\frac{1}{8}} M \left( g_0 V_0 t \left( 1 - \frac{r^2}{V_0^2 t^2} \right)^{\frac{3}{4}} \right) \cdot H \left( t - \frac{r}{V_0} \right), \quad (5)$$

where  $M(x) \approx e^x \sqrt{1 + 2.026/x}$ . The error of this approximation is of the order of 2% outside the ballistic peak and its tail for  $g_0 V_0 t < 6$  and  $2 < g_0 r < 4$ .

Figure 3 shows spatiotemporal variations in energy density in a scattering medium theoretically predicted by the approximation solution (5) for instantaneous spherical source radiation at the origin. Scattered energy density is shown by a black curve, where the ballistic term is shown by a vertical gray line. At a small distance from the source compared with the mean free path  $1/g_0$ , the energy density decreases rapidly after the direct arrival as predicted by the single scattering term; however, the decay rate becomes smaller due to multiple scattering according to the power of lapse time  $t^{-3/2}$  at long lapse times. At a long distance, for example at  $r = 3.2/g_0$ , the energy density has an additional diffusion peak. The spatial distribution of scattered energy density is uniform around the source at a short lapse time compared with the mean free time  $1/g_0 V_0$ ; however, it converges to a Gaussian curve at a long lapse time as theoretically predicted by the diffusion solution (4), for example at  $t = 7.48/g_0 V_0$ . There is no violation of causality since no signal exists beyond the ballistic peak. The smooth spatial distribution of scattered energy density around the source location gives the physical basis of the coda normalization method for measure-



Seismic Waves in Heterogeneous Earth, Scattering of, Figure 3 a Temporal change and b spatial variation of energy density in an isotropic-scattering medium for a point source radiation. Each black curve shows the scattering contribution predicted by the Paaschens approximation and each vertical gray line shows a ballistic term

ments of S-wave attenuation and site amplification factors (e. g. [3,56,89]).

Gusev and Abubakirov [21] and Hoshiba [27] numerically solved the radiative transfer equation for the isotropic scattering process by using the Monte Carlo method. Yoshimoto [88] numerically simulated envelopes in scattering media of which the background velocity decreases with depth. He found the concentration of scattered energy near the surface because of seismic ray bending.

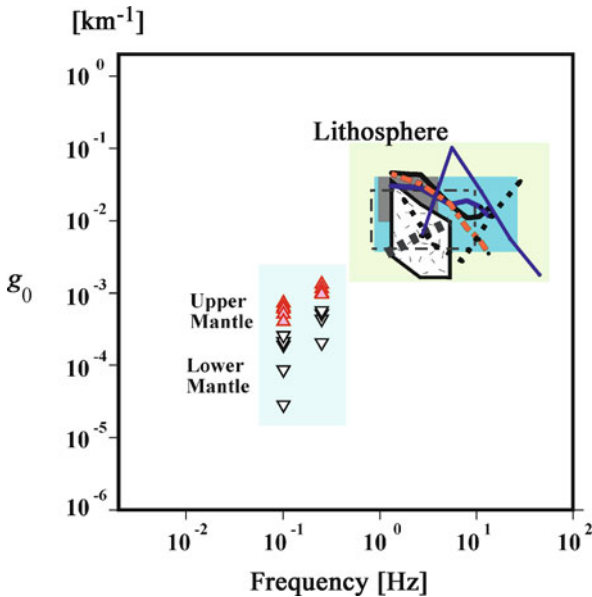
Nonisotropic radiation from a source can be easily introduced in the radiative transfer equation. Sato et al. [75] analytically solved the case of double couple source radiation: the energy density theoretically predicted faithfully reflects the source radiation pattern near the direct arrival; however, the azimuthal dependence diminishes with increasing lapse time. It qualitatively agrees with the observed radiation pattern independence of coda amplitudes at long lapse times. Their solution has been used in the envelope inversion of strong motion records for the spatial distribution of high-frequency wave energy radiation from an earthquake fault (e. g. [47]).

### Measurements of Total Scattering Coefficient and Attenuation

For the practical application of the radiative transfer theory to observed seismograms, it is necessary to introduce intrinsic absorption  $Q_{\text{Int}}^{-1}$  by multiplying an exponential temporal decay factor  $\exp[-Q_{\text{Int}}^{-1}2\pi ft]$  to the resultant energy density. By using the solution (3) of the radiative transfer theory for the isotropic scattering model, total scattering coefficient  $g_0$  and intrinsic absorption factor  $Q_{\text{Int}}^{-1}$  of the S-wave have been measured. Reported  $g_0$  values in the lithosphere are of the order of  $10^{-2} \text{ km}^{-1}$  for frequencies from 1 to 30 Hz as plotted in Fig. 4.

From the observed lapse time dependence of  $Q_{\text{C}}^{-1}$ , Gusev [20] quantitatively explained the decrease of  $g_0$  with depth. Lee et al. [40] analyzed coda envelopes of regional earthquakes before and after the ScS arrival around 900s in lapse time from the origin time based on the numerically simulated envelopes for the PREM model, which is characterized by depth-dependent background velocity and total attenuation for S-waves. They reported lower  $g_0$  values in 4s and 10s period bands in the upper and lower mantle compared with those in the lithosphere as illustrated in Fig. 4.

Hoshiba et al. [28] and Fehler et al. [10] developed a method to measure simultaneously  $g_0$  and  $Q_{\text{Int}}^{-1}$  values for the S-wave from the whole S-envelope analysis based on the synthetic envelope derived from the radiative



Seismic Waves in Heterogeneous Earth, Scattering of, Figure 4 Total scattering coefficient of S-waves in the Earth. Measurements in the mantle [40] are added to lithospheric inhomogeneity [74]

transfer theory. Their multiple lapse-time window analysis method has been widely used in the world. Estimated scattering loss  $g_0 V_0/\omega$  decreases with frequency; however, intrinsic absorption  $Q_{\text{Int}}^{-1}$  is rather insensitive to frequency. Estimated seismic albedo  $B \equiv g_0 V_0/(\omega Q_{\text{Int}}^{-1} + g_0 V_0)$ , the ratio of scattering loss to the total attenuation, of S-waves in the lithosphere widely distribute from 0.2 to 0.8 for 1–6 Hz, but they are limited between 0.2 and 0.5 for 6–20 Hz.

Total scattering cross-section and number density of scatterers appear jointly as the total scattering coefficient in theoretical models; however, Matsumoto [44] proposed a method to separate them from the temporal variation of the semblance coefficient of coda waves recorded by a seismic array. Analyzing data obtained in the aftershock area of the 2000 western Tottori earthquake, Japan, he estimated  $n = 0.03 \text{ km}^{-3}$  and  $g_0 = 0.001 \text{ km}^{-1}$  at 20 Hz.

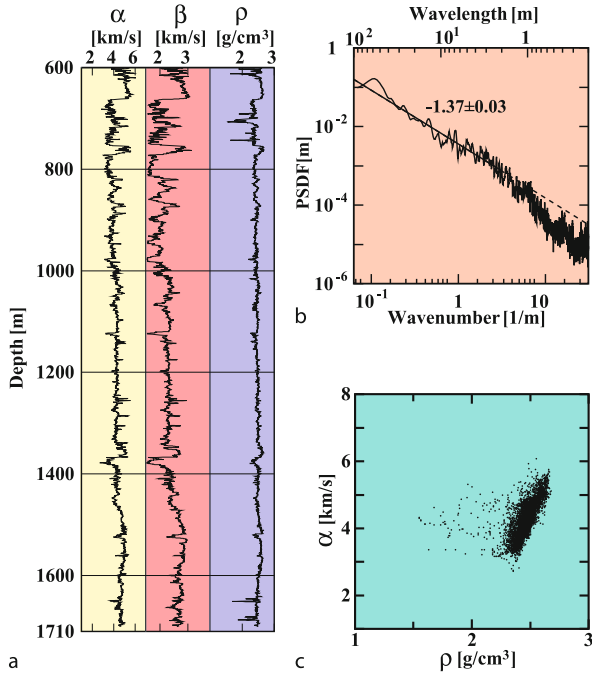
It should be noted that lunar seismograms have coda durations exceeding one hour. Applying the diffusion model for the explanation of spindle-like envelopes of lunar-quakes, Dainty and Toksöz [9] estimated  $g_0$  to be as large as  $0.05\text{--}0.5 \text{ km}^{-1}$  at 0.45 Hz.

### Wave Envelopes in Random Media and Statistical Characterization

#### Statistical Characterization of Random Media

As revealed from tomography analyses (e. g. [93]), velocity structure is three-dimensionally inhomogeneous especially in the lithosphere. Well log data show typical samples of the shallow crust. Figure 5a shows well log data of P- and S-wave velocities and mass density obtained in Kyushu, Japan [80], where wave velocities and mass density are measured in a borehole by using ultrasonic waves and gamma rays, respectively. These well log data clearly show random fluctuation with short wavelengths. As a natural consequence, we imagine random inhomogeneities widely distributed in the solid Earth medium.

By using computer power, wave propagation in random media has been numerically studied extensively. Using finite difference simulations of waves in random media, Frankel and Clayton [14] first examined the relation between coda excitation and scattering loss and the spectrum of random inhomogeneity. On the basis of numerical simulations, Frankel and Wennerberg [15] proposed the energy flux model that has a uniform distribution of scattered energy behind the direct waves for the analysis of high-frequency seismogram envelopes. Using a boundary integral method for the simulation of waves in a medium containing many cavities, Yomogida and Benites [87] examined a relation between coda attenuation and the distribution of cavities.



Seismic Waves in Heterogeneous Earth, Scattering of, Figure 5  
**a** Well log data at the YT-2 site, Kyushu, Japan. **b** Power spectral density function of the fractional fluctuation of P-wave velocity log. **c** Scattergram of P-wave velocity and mass density. Reproduced from [80]

There is an alternative approach to treat statistically randomly inhomogeneous media. The wave-velocity is written as  $V(\mathbf{x}) = V_0 \{1 + \xi(\mathbf{x})\}$ , where  $V_0$  is the average velocity and fractional fluctuation  $\xi(\mathbf{x})$  is a homogeneous and isotropic random function of space coordinate  $\mathbf{x}$ . We imagine an ensemble of random media  $\{\xi\}$ , which is statistically characterized by the autocorrelation function (ACF)  $R(\mathbf{x}) \equiv \langle \xi(\mathbf{x} + \mathbf{x}') \xi(\mathbf{x}') \rangle$ , where angular brackets mean the ensemble average. The MS fractional fluctuation  $\varepsilon^2 \equiv R(0)$  and the correlation distance  $a$  are key parameters. The Fourier transform of ACF gives the power spectral density function (PSDF)  $P$ . The PSDF of the P-wave velocity fractional fluctuation of well log data shows a power-law characteristic at large wavenumbers as illustrated in Fig. 5b. P-wave velocity and mass density show a good correlation as shown in Fig. 5c. The statistical view is useful for representing geological data, too (e. g. [18,26]).

### Scattering Coefficient Based on the Born Approximation

When the medium inhomogeneity is small  $|\xi| \ll 1$ , scalar wave  $\phi$  is governed by the following wave equation:

$$\left(\Delta - \frac{1}{V_0^2} \partial_t^2\right) \phi + \frac{2}{V_0^2} \xi(\mathbf{x}) \partial_t^2 \phi = 0. \quad (6)$$

Velocity inhomogeneity is supposed to localize in a volume around the origin, of which the dimension is chosen to be much larger than  $a$ . For the incidence of a plane wave of unit amplitude at angular frequency  $\omega$  as  $e^{i(k_0 \mathbf{e}_z x - \omega t)}$ , where  $\mathbf{e}_z$  is the unit vector to the  $z$  direction, we calculate the spherically outgoing scattered waves due to a localized inhomogeneity by using the Born approximation as  $\phi^1(\mathbf{x}, t) = -k_0^2 e^{i(k_0 r - \omega t)} \tilde{\xi}(k_0 \mathbf{e}_r - k_0 \mathbf{e}_z) / (2\pi r)$ , where the tilde means the Fourier transform with respect to coordinates in 3-D space and  $\mathbf{e}_r$  is a radial unit vector (e. g. [74]). According to Aki and Chouet [4], the scattering coefficient defined as the scattering power in a unit solid angle at certain direction by a unit volume of random inhomogeneous media for the incidence of unit energy flux density is statistically written by using its PSDF as

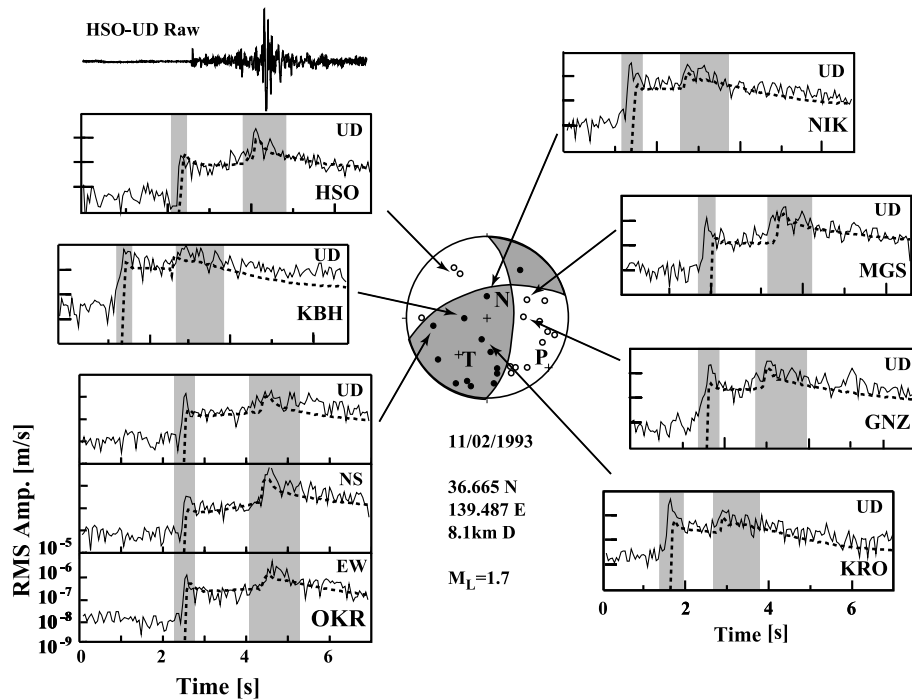
$$g(\psi; \omega) = \frac{k_0^4}{\pi} P(k_0 \mathbf{e}_r - k_0 \mathbf{e}_z) = \frac{k_0^4}{\pi} P\left(2k_0 \sin \frac{\psi}{2}\right), \quad (7)$$

where  $\psi$  is the scattering angle measured from the  $z$  direction. This functional form means anisotropic scattering depending on frequency. In general, scattering near around the forward direction becomes larger with increasing wavenumber in random media.

### Radiative Transfer Theory with Scattering Coefficients Calculated by the Born Approximation

Extending the above formulation to vector wave propagation in random elastic media, we can define scattering coefficients for different scattering modes as PP, PS, SP, and SS. For the case of random elastic media characterized by an exponential ACF with  $\varepsilon = 10\%$  and  $a = 2$  km, Sato [68] synthesized three-component seismogram envelopes of a microearthquake of M 3 as a superposition of polarized scattered waves' power at a finite distance from a point shear dislocation source. That is the single scattering approximation of the radiative transfer theory with scattering coefficients calculated by the Born approximation. The SS scattering mode dominates in S coda, and pseudo P and S waves are produced even at a receiver on the null direction of the source radiation. By using a von Kármán-type ACF for describing random elastic media, Sato [71] estimated parameters  $\kappa = 0.35$ ,  $\varepsilon = 8.4\%$  and  $a = 2.1$  km from observed frequency dependence of S-wave attenuation and  $g_0$ , where the parameter  $\kappa$  controls the role-off of PSDF at large wavenumbers. Extending the above vector wave envelope synthesis to include mode conversions at the free surface, Yoshimoto et al. [90]





Seismic Waves in Heterogeneous Earth, Scattering of, Figure 6

RMS envelopes of a microearthquake in the shallow crust, Nikko in northern Kanto, Japan. *Fine curves* and *broken curves* are observed and best-fit synthesized envelopes, respectively, where *shades* show time windows used for the estimation of the source radiation energy. The trace on the *top left* shows a raw seismogram. Reproduced from [90]

analyzed three-component seismogram envelopes of microearthquakes in the shallow crust in Nikko, northern Kanto, Japan. A raw seismogram is shown at the top-left of Fig. 6 as an example. Fine curves in Fig. 6 are logarithmic plots of observed root mean square (RMS) envelopes in the 2–16 Hz band. Broken curves are best-fit theoretical envelopes for random elastic media characterized by an exponential ACF with  $\varepsilon = 5.7\%$  and  $a = 400$  m. We find that the fitness is good not only for S coda but also for P coda.

Wave theory in random media predicts that the scattering coefficient has a large lobe in the forward direction in higher frequencies. Gusev and Abubakirov [22] used the Monte Carlo method to simulate envelopes for the multiple nonisotropic scattering process. There have been mathematical developments to derive the radiative transfer equation for multiple nonisotropic scattering from the stochastic averaging of the wave equation in random media (e. g. [13,29,41,64]). Przybilla et al. [58] showed an excellent coincidence of vector envelopes calculated from finite difference simulation in 2-D random elastic media and those synthesized by the radiative transfer theory with scattering amplitudes derived from the Born approximation and the wandering effect of travel time.

### Interference of Scattered Waves

The interference effect is neglected in conventional studies of wave scattering in random media; however, it becomes important for a specific case even in random media. When randomness is strong enough to produce multiple scattering, coda wave intensity at a receiver near a source is enhanced compared to the prediction of conventional radiative transfer theory. Margerin et al. [42] showed that a spot of backscattering enhancement stabilizes in a sphere of radius half a wavelength centered at the source after a transient regime. The enhancement persists in time and should be observable as long as a coda is measurable. From field experiment of seismic waves in a shallow volcanic structure Larose et al. [39] reported the existence of weak localization, where the size of enhancement spot was one wavelength and the estimated mean-free path was 200 m for seismic waves around 20 Hz.

### Envelope Broadening of a High-Frequency Seismogram

#### Markov Approximation for Parabolic Wave Equation

For the study of light propagation through the upper atmosphere and/or acoustic sound propagation through in-

ternal waves in oceans, various stochastic methods have been developed in the fields of physics. One of the most attractive methods for explaining the wave envelope around the direct arrival is the Markov approximation for the parabolic wave equation, which is an extension of the phase screen method or the split step Fourier method (e. g. [29,63]). This method is found to be applicable to seismogram envelopes. We imagine an elastic medium composed of a homogeneous half space  $z < 0$  and an inhomogeneous half space  $z > 0$ , where the inhomogeneity is supposed to be small ( $\varepsilon^2 \ll 1$ ) and the randomness is statistically homogeneous and isotropic. When the wavelength is smaller than the correlation distance  $a$ , we may neglect conversion scattering between P and S waves. Then, we can describe the principal characteristics of vector wave propagation by using potentials. For the incidence of plane P-wavelet to the  $z$  direction from the homogeneous zone, scalar potential is written as a superposition of harmonic waves of angular frequency  $\omega$  as  $\phi = \int_{-\infty}^{\infty} (2\pi i k_0)^{-1} U(\mathbf{x}_{\perp}, z, \omega) e^{ik_0 z - i\omega t} d\omega$  for  $z > 0$ , where  $\mathbf{x}_{\perp} = (x, y)$  on the transverse plane. Neglecting the second derivative with respect to  $z$ , we have the parabolic-type equation for  $U$  as

$$2ik_0 \partial_z U + (\partial_x^2 + \partial_y^2) U - 2k_0^2 \xi(\mathbf{x}) U = 0. \tag{8}$$

We define the two-frequency mutual coherence function (TFMCF) of field  $U$  between two different locations on the transverse plane at a distance  $z$  and different angular frequencies at  $\omega'$  and  $\omega''$  as  $\Gamma_2(\mathbf{x}_{\perp c}, \mathbf{x}_{\perp d}, z, \omega_c, \omega_d) \equiv \langle U(\mathbf{x}'_{\perp}, z, \omega') U(\mathbf{x}''_{\perp}, z, \omega'')^* \rangle$ , where  $\omega_c$  and  $\omega_d$  are center-of-mass and difference angular frequencies, respectively. In the case of quasi-monochromatic waves  $|\omega_d| \ll |\omega_c|$ , using causality and neglecting back scattering, we derive the master equation for TFMCF. This derivation is called the Markov approximation. For the  $i$ th component, the intensity is defined as the ensemble average of the square of displacement  $\langle \partial_i \phi \partial_i \phi^* \rangle = 1/(2\pi) \int_{-\infty}^{\infty} \widehat{I}_i^P d\omega_c$ . The integrand is the intensity spectral density (ISD)  $\widehat{I}_i^P$ , which means the time trace of MS amplitude in a band having the central angular frequency  $\omega_c$ .

**Vector Envelopes for a Gaussian ACF**

The case of a Gaussian ACF  $R(\mathbf{x}) = \text{Exp}(-r^2/a^2)$  is mathematically tractable. For the initial condition  $\widehat{I}_x^P = \widehat{I}_y^P = 0$  and  $\widehat{I}_z^P = \delta(t - z/V_0)$  at  $z = 0$ , ISDs are analyti-

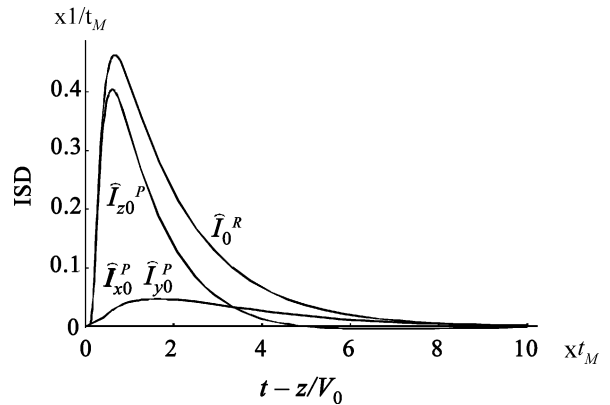
cally written as [72]

$$\begin{aligned} \widehat{I}_{x0}^P(z, t, \omega_c) &= \widehat{I}_{y0}^P(z, t, \omega_c) \\ &= 2(V_0/z)(t - z/V_0) \cdot \widehat{I}_0^R(z, t, \omega_c) \\ \widehat{I}_{z0}^P(z, t, \omega_c) &= [1 - 4(V_0/z)(t - z/V_0)] \\ &\quad \cdot \widehat{I}_0^R(z, t, \omega_c), \end{aligned} \tag{9}$$

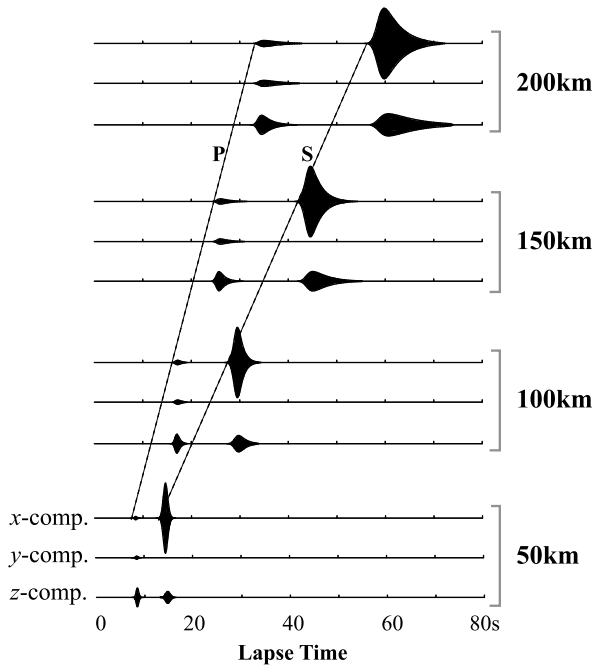
where subscript “0” means ISD without the wandering effect. The reference ISD is a solution for scalar waves for the initial condition  $\Gamma_2(\mathbf{x}_{\perp}, z = 0) = 1$  [82]:

$$\widehat{I}_0^R(z, t, \omega_c) = \frac{1}{t_M} \frac{\pi}{8} \vartheta_1' \left( 0, e^{-\frac{\pi^2}{4} \frac{(t-z/V_0)^2}{t_M}} \right) \cdot H \left( t - \frac{z}{V_0} \right), \tag{10}$$

where  $t_M = \sqrt{\pi} \varepsilon^2 z^2 / (2V_0 a)$  is the characteristic time and function  $\vartheta_1'$  is the derivative of the elliptic theta function of the first kind. Function  $\widehat{I}_0^R$  shows a broadened envelope having a delayed peak and a smoothly decaying tail as illustrated by a chained curve in Fig. 7, where solid and broken curves show three-component ISDs  $\widehat{I}_{0z}^P$  and  $\widehat{I}_{x0}^P (= \widehat{I}_{y0}^P)$ , respectively, for  $V_0 t_M / z = 0.05$  as an example. All the three component envelopes have broadened traces; however, the peak height of the transverse component is smaller than that of the longitudinal component and the peak delay of the transverse component is larger than that



Seismic Waves in Heterogeneous Earth, Scattering of, Figure 7 Chained curve shows the reference ISD without the wandering effect in 3-D random elastic media characterized by a Gaussian ACF for the incidence of a plane P-wavelet. Solid and broken curves show three-component ISDs without the wandering effect for  $V_0 t_M / z = 0.05$ . Reproduced from [72]



Seismic Waves in Heterogeneous Earth, Scattering of, Figure 8  
Synthesized vector envelopes in random media characterized by a Gaussian ACF for radiation of P wavelet and S wavelet with a polarization to the  $x$ -axis from a point source. Reproduced from [73]

of the longitudinal component. When  $\varepsilon^2 z/a \ll 1$ , the peak height of  $\widehat{I}_{z0}^P$  approximately decays according to the square of travel distance and the peak ratio of the transverse component to longitudinal component is proportional to  $\varepsilon^2 z/a$ . ISDs  $\widehat{I}_x^P$ ,  $\widehat{I}_y^P$ , and  $\widehat{I}_z^P$  can be calculated by using the convolution of (9) with the travel-time wandering effect  $\exp[-(V_0 t - z_0)^2 / 2\sqrt{\pi}\varepsilon^2 a z] V_0 / \sqrt{2\pi\sqrt{\pi}\varepsilon^2 a z}$  in time domain. For 2-D cases, the validity of the Markov approximation was numerically confirmed by a comparison with the finite difference simulations [11,36].

The above synthesis can be simply extended to S wave envelopes. Extension from plane wave incidence to impulsive radiation from a point source is also possible [73]. Figure 8 shows simulated three-component RMS envelopes along the  $z$  axis for a point source radiation, where random media are characterized by average P and S wave velocities, 6 km/s and 3.46 km/s, respectively, and a Gaussian ACF with  $\varepsilon = 5\%$  and  $a = 5$  km. We assume that P-wavelet radiation is isotropic and S-wavelet radiation is axially symmetric around the  $y$  axis with polarization to the  $x$  axis, where the ratio of S to P-wave source energy is chosen to be 23.3. Envelope broadening is common to both P and S waves in the syntheses. Excitation of the

transverse component for P-waves and that of the radial component for S-waves are prominent. The appearance of scattered S-waves having long duration in synthesized envelopes at large travel distances qualitatively well explains observed characteristics shown in Fig. 2.

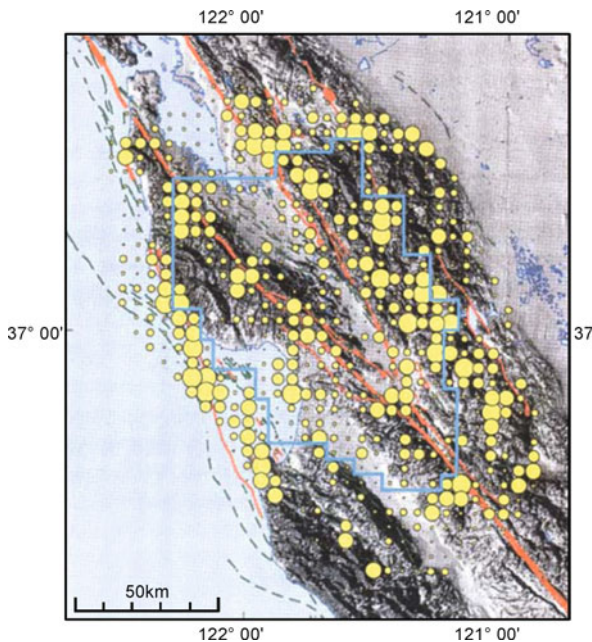
### Randomness in the Lithosphere

Applying the envelope broadening model to S-wave seismograms recorded in Kanto, Japan, Sato [70] estimated the ratio  $\varepsilon^2/a \approx 10^{-3} \text{ km}^{-1}$  with the assumption of a Gaussian ACF and S-wave attenuation  $Q^{-1} = 0.014 f^{-1}$ . Saito et al. [66] studied the case of a von Kármán-type random media having a power-law spectrum at large wavenumbers, which are more appropriate for the real Earth inhomogeneity. The resultant envelope shows frequency dependence, which is controlled by the roll-off of the PSDF. Analyzing the hypocentral-distance dependence and frequency dependence of S-wave seismogram envelopes in northern Honshu, Japan for 2–32 Hz, Saito et al. [66] estimated parameters of von Kármán-type ACF as  $\kappa = 0.6$  and  $\varepsilon^{2.2}/a \approx 10^{-3.6} \text{ km}^{-1}$  with  $Q^{-1} = 0.009 f^{-1}$ . It means the PSDF decreases as wavenumber to the power of  $-4.2$ . Petukhin and Gusev [55] averaged S-wave seismogram envelopes of small earthquakes recorded in Kamchatka and compared the shapes with those numerically calculated for various types of random media. They concluded that random media whose PSDF decreases as the wavenumber to the power of  $-3.5$  to  $-4$  are appropriate.

### Spatial Variation of Scattering Characteristics

#### Scattering Coefficient and Active Faults

Precisely examining coda envelopes of local earthquakes against lapse time measured from the origin time, we find temporal fluctuations around the smoothly decaying master curve. We may interpret that swellings and dips around the master curve are caused by stronger and weaker scatterers, respectively, distributed in the subsurface. By using a single isotropic scattering model, Nishigami [48] proposed an inversion scheme from coda envelopes of local earthquakes recorded at multiple stations for estimating the spatial variation of the scattering coefficient. Applying this inversion scheme to coda records obtained in central California, Nishigami [49] mapped the distribution of relative scattering coefficient in the shallow crust as in Fig. 9. A good correlation is found between sub-parallel active faults and relatively stronger scattering zones marked by larger circles, where some large circles are caused by topographic roughness. He also suggested that segment bound-



Seismic Waves in Heterogeneous Earth, Scattering of, Figure 9  
**Distribution of relative scattering coefficient at a depth of 0–5 km in central California revealed from the coda envelope inversion. Circles with larger diameter indicate stronger scattering and solid lines represent active faults. Reproduced with permission from [49]**

aries of the San Andreas Fault are characterized by relatively stronger scattering.

Stacking forward scattered energy in the coda of teleseismic P waves observed by a local seismographic network, Revenaugh [59] proposed a Kirchhoff coda migration method, which puts a focus on small-angle scattering from the forward direction. He made a map of P-wave scatterers in the upper mantle beneath southern California. Between depths of 50 km and 200 km, the southern flank of the slab subducting beneath the Transverse Ranges was marked by strong scattering. Using the same method, Revenaugh [60,61] estimated geographic variation of the statistical significance of scattering potential in the upper crust in California, where the scattering potential is a measure of the likelihood that scattering strength locally exceeds the regional mean. In the region surrounding the 1992 Landers earthquake of M 7.3, he found a noticeable tendency for aftershocks to cluster in regions of strong scattering potential.

There were more precise mappings of scattering coefficient. Slant-stacking records of 12 explosions in the Awaji island, Japan registered by a dense seismic array for a 6–10 Hz band, Matsumoto et al. [45] mapped the spatial distribution of PP signal scatterers. The resultant distribution

of scatterers shows higher strengths beneath the initiation point of the mainshock rupture and in the southwestern part of the fault plane of the 1995 Kobe earthquake (M 7.2). Analyzing precisely aftershock records of the 2000 western Tottori earthquake (M 7.3), Japan registered by a dense seismic network, Asano and Hasegawa [5] found strong scattering along and around the fault zone of 20 km in length.

### Coda Attenuation and Deformation Zone

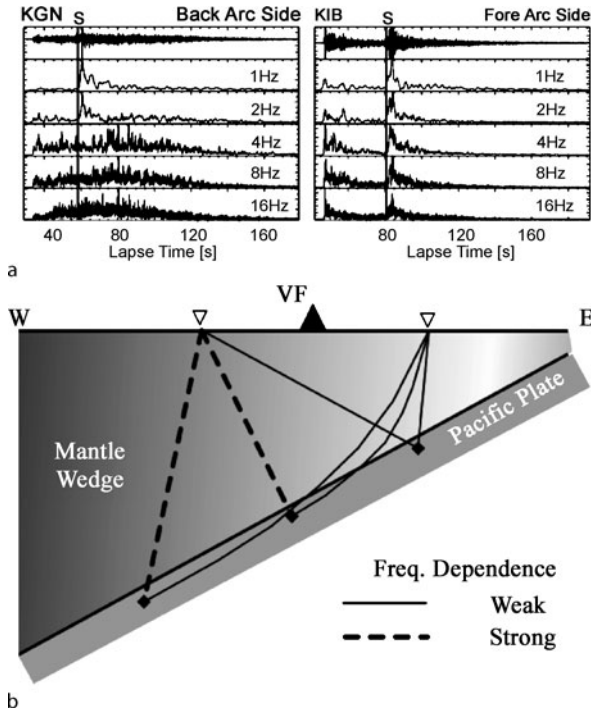
Regional variation of coda attenuation  $Q_c^{-1}$  has been measured from the decay gradient of coda amplitude envelopes of small earthquakes in various areas in the world. Jin and Aki [31] made a map of  $Q_c^{-1}$  at 1 Hz in China. They reported that  $Q_c^{-1}$  is as large as 0.01 in Tibet at the active continental collision. They found that large historical earthquakes took place in large  $Q_c^{-1}$  regions. Jin and Aki [33] made precise analysis of  $Q_c^{-1}$  for 1–32 Hz in Japan. They found significant spatial variation up to a factor of 3 for the lower frequency bands, as well as its strong frequency dependence. They found conspicuous large  $Q_c^{-1}$  for 1–4 Hz in a narrow belt from Niigata towards the south-west to the Biwa lake along the Japan Sea coast, which coincides with a high-deformation rate zone revealed from the GPS observation. For frequency bands 4–16 Hz (2–4 Hz in Kyushu), large  $Q_c^{-1}$  areas agree with volcanic and geothermal areas.

### Attenuation and Volcanoes

Yoshimoto et al. [91] studied the spatial variation of MS amplitude of S-coda at a fixed lapse time across the volcanic front in northeastern Honshu, Japan: S-coda energy is uniformly distributed in the fore-arc side, whereas an exponential decrease with horizontal offset to the west from the volcanic front was found in the back-arc side. The decay rate increases with increasing frequency. They interpreted this variation by a diffusion-absorption model, where the intrinsic absorption factor of S-wave  $Q_{\text{Int}}^{-1} = 0.02$  at a frequency of 10 Hz beneath the back-arc side, which is about twice as large as those reported for the fore-arc side.

### Scattering and Volcanoes

Medium heterogeneity is strong beneath volcanoes. Applying the diffusion model to seismogram envelopes beneath Merapi volcanoes, Friedrich and Wegler [16] estimated the total scattering coefficient as large as  $5 \text{ km}^{-1}$  as shown in Fig. 4. Nishimura et al. [50] applied an envelope inversion method based on the isotropic scatter-



Seismic Waves in Heterogeneous Earth, Scattering of, Figure 10 **a** Seismogram envelopes observed in the back-arc side and fore-arc side in Kanto-Tokai, Japan and **b** a schematic illustration of seismic rays. Reproduced from [53]

ing model for PP and PS scattering to artificial explosion records obtained in Jemez volcanic field, New Mexico. They found that the mid-crust under most of the region is fairly transparent but that the lower crust is heterogeneous. The strongest scattering occurs at shallow depths beneath the center of the caldera, where the medium is highly heterogeneous.

Obara and Sato [53] analyzed S-wave envelopes of microearthquakes in Kanto-Tokai, Japan, where the Pacific plate is subducting from east to west, to examine regional differences in their envelope broadening. As shown by examples in Fig. 10a, envelope broadening is typically stronger for higher frequencies in records at stations on the back-arc side of the volcanic front but weaker and frequency independent in records on the fore-arc side. These regional differences in the envelope broadening mean that PSDF of velocity inhomogeneity is rich in short-wavelength components in the mantle wedge on the back-arc side and poor on the fore-arc side as schematically illustrated in Fig. 10b. Takahashi et al. [83] precisely examined how the peak delay from the S-wave onset depends on the ray path in northern Japan. They found that peak delays observed in the back-arc side of the volcanic front

are larger for rays which propagate beneath Quaternary volcanoes (see Fig. 11b and d); however, peak delays for rays which propagate between Quaternary volcanoes are as short as those observed in the fore-arc side (see Fig. 11a, c, and e). Large peak delay suggests strong scattering due to medium inhomogeneity. That is, the structure beneath Quaternary volcanoes is not only characterized by low velocity and large intrinsic absorption revealed from tomography studies but also by strong inhomogeneity.

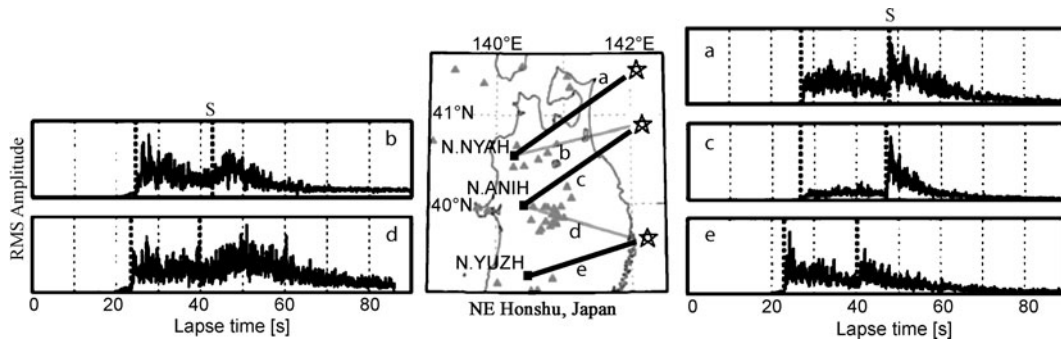
### Nonisotropic Random Medium Oceanic Slab

If random media are statistically nonisotropic, scattering contribution depends on the propagation direction. Saito [65] studied the envelope broadening in nonisotropic random media based on the Markov approximation. His simulations show that the envelope of scalar wavelet propagating in parallel to the longer correlation direction has longer duration compared to that with the shorter correlation direction. The effective envelope broadening in the elongated direction shows the wave trap phenomenon of nonisotropic random media.

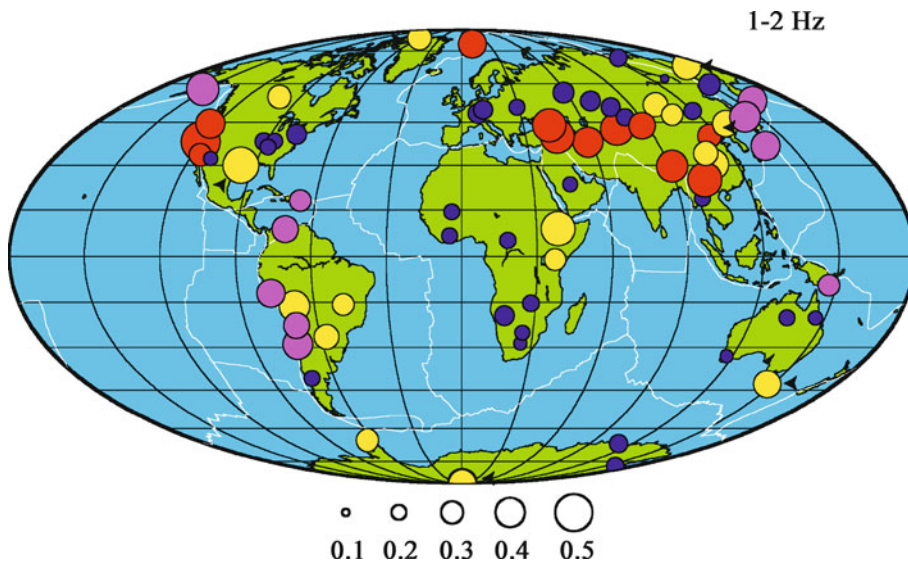
An intensity anomaly is observed on the eastern seaboard of northern Japan for deep earthquakes. The waveform records in the region of high intensity show a low-frequency ( $f < 0.25$  Hz) onset for both P and S waves, followed by large-amplitude high-frequency ( $f > 2$  Hz) later arrivals with long coda. A simple subduction zone model comprising a high-velocity plate with low attenuation cannot explain quantitatively these observed facts. Furumura and Kennett [17] proposed a scattering slab model that the nonisotropic random structure in the Pacific plate works as a wave-guide for high-frequency seismic waves. Their preferred random medium is characterized by a von Kármán-type ACF with elongated correlation distance of about 10 km parallel to the plate margin and much shorter correlation length of about 0.5 km in thickness and  $\varepsilon$  of about 2%. They clearly demonstrated the scattering waveguide effects and frequency selectivity for seismic waves traveling through the Pacific plate by using 3-D numerical simulations.

### Lateral Variation of Lithospheric Heterogeneity

Korn [34] developed the energy flux model [15] appropriate for the wave front of teleseismic P and P-coda waves propagating through a scattering layer. During the propagation the primary wave loses energy due to scattering and intrinsic absorption, then the scattered energy appears as coda energy behind the wave front. From the analysis of vertical component trace envelopes observed in the world, Korn [35] found strong scattering at island arcs



Seismic Waves in Heterogeneous Earth, Scattering of, Figure 11  
 Path dependence of RMS envelopes (16–32 Hz) in northeastern Honshu, Japan, where stars and triangles indicate earthquake epicenters and Quaternary volcanoes, respectively. Ray paths b and d travel beneath Quaternary volcanoes (triangles), and a, c, and e travel between Quaternary volcanoes. Reproduced from [83]



Seismic Waves in Heterogeneous Earth, Scattering of, Figure 12  
 Plots of the square root of the relative partition of wave energy into the transverse component (1–2 Hz) by a diameter of the circle revealed from the teleseismic P-waves. Reproduced from [37]

and smaller scattering on stable continental areas like Australia. Nishimura et al. [52] analyzed the energy partition of teleseismic P and P-coda into the transverse component to evaluate the lithospheric heterogeneity in the western Pacific region. They showed the presence of strong heterogeneity in and around the tectonically active regions. Kubanza et al. [37] systematically characterized the medium heterogeneity of the lithosphere by analyzing the partition of P-wave energy into the transverse component for 0.5–4 Hz. They found significant regional differences as shown in Fig. 12. The energy partition to the transverse component is small at stations on stable continents while

the partition is large at stations in tectonically active regions such as island arcs or collision zones.

**Random Inhomogeneity in the Lithosphere and Mantle**

Records of earthquakes registered by arrays of seismographs are useful for the statistical measurement of the Earth inhomogeneity. Aki [2] first analyzed teleseismic P-waves centered on about 0.6 Hz registered by a seismic array in Montana for the quantification of the lithospheric inhomogeneity. Measuring transverse correlation functions of teleseismic P-waves arriving from near ver-

tical incidence, he found a positive correlation between log-amplitude and phase fluctuations as theoretically predicted for a Gaussian ACF. From plots of the ratio of RMS log-amplitude to RMS phase fluctuations against the correlation between log-amplitude and phase fluctuations, he estimated the thickness of the inhomogeneous lithosphere to be 60 km,  $a = 10$  km, and  $\varepsilon = 4\%$ . Flatté and Wu [12] measured the transverse correlation of log-amplitude and phase fluctuations of teleseismic P-wave beams with 2 Hz center frequency recorded at NORSAR. They also introduced the new concept of angular correlation functions, which are based on measurements of two rays with different incident angles. They proposed a model for lithospheric and asthenospheric inhomogeneities that consists of two overlapping layers: the upper layer extending from the surface to about the 200 km depth has a white PSDF, however, the lower layer extending from 15 to 250 km has poor amplitude in the short wavelength spectrum. There are more small-scale inhomogeneities near the surface compared with the deeper portions.

The radiative transfer theory with scattering coefficients calculated from the Born approximation was also used for the study of mantle inhomogeneity. Analyzing stacked P and P coda envelopes of teleseismic ( $>10^\circ$ ) events at 1 Hz, Shearer and Earle [79] concluded that most scattering occurs in the lithosphere and upper mantle, but that some lower mantle scattering is likely required. They estimated  $\varepsilon$  to be 3–4% and  $a = 4$  km in the upper mantle and 0.5% and 8 km in the lower mantle. Analyzing envelopes of precursors to PKP, Margerin and Nolet [43] found that inhomogeneity can not be restricted to the D'' layer and a small inhomogeneity spread over the whole lower mantle. They proposed a von Kármán ACF of  $\kappa = 0$  for random media, which has a power law PSDF rich in short wavelengths compared with an exponential ACF. They mentioned that  $\varepsilon$  of 0.1–0.2% in the whole lower mantle is enough to explain the observation even though correlation distance is irresolvable because of the limited range of observations.

### Imaging of Subsurface Heterogeneity

There have been developments in deterministic imaging of medium inhomogeneity from the analysis of array records of P-coda waves. When the structures of interest are characterized by laterally variable stratification, the receiver-function technique [38] is useful since it is based on the deconvolution of the horizontal component trace in the radial direction by the vertical component trace for measuring the Ps conversion depth. On the other hand, scattering from localized volume inhomogeneity is most read-

ily treated by using the Born approximation. Analyzing array records of teleseismic P coda in central Oregon by using the Born approximation for both forward scattered waves and backscattered free-surface reflected waves, Rondenay et al. [62] successfully imaged the precise structure of the Cascadia subduction zone, which is consistent with the consequences of prograde metamorphic reactions occurring within the oceanic crust. Analyzing array records of P coda waves of regional earthquakes at Izu-Oshima volcano, which erupted in 1986, Mikada et al. [46] deterministically imaged PP and PS scatterers on the basis of diffraction tomography. They interpreted a cloud of scatterers centered at about 10 km depth beneath the volcano crater as a primary magma reservoir and smaller and shallower patches of high scattering strength with sub-magma reservoirs.

### Temporal Change in the Earth Medium Structure

There were reports on the temporal changes in the Earth medium structure revealed from the analyses of scattered waves. One is coda amplitude envelope analysis, which gives information about the change in intrinsic absorption and scattering strength of the crustal heterogeneity. Another is coda phase interferometry, which offers information about the change in background velocity.

### Change in Coda Characteristics

Monitoring coda envelopes of local earthquakes, Gusev and Lemzikov [23] reported temporal change in  $Q_c^{-1}$  before and after the 1971 Ust-Kamchatsk earthquake (M 7.8), and Jin and Aki [30] reported temporal change in  $Q_c^{-1}$  associated with the 1976 Tangshan earthquake (M 7.8) in China. Their observation attracted the interest of geophysicists to the temporal variation of coda characteristics because of a potential for monitoring the stress accumulation process preceding an earthquake occurrence. Analyzing high-frequency seismograms recorded at Riverside, California for 55 years, Jin and Aki [32] found a temporal variability in  $Q_c^{-1}$  at about 1.6 Hz having a positive correlation with the seismic  $b$ -value calculated for  $M > 3$  earthquakes within a 180 km radius. The seismic  $b$ -value is a measure of the ratio between the numbers of small to large earthquakes; smaller  $b$ -values mean that there are relatively fewer small earthquakes compared to the number of larger ones. They interpreted these changes in  $Q_c^{-1}$  and the  $b$ -value by creep fractures in the ductile part of the lithosphere. Hiramatsu [25] precisely examined the temporal variation in  $Q_c^{-1}$  and  $b$ -value for 10 years before and after the 1995 Hyogo-ken Nanbu earthquake (M 7.2) in Japan. At frequencies between 1.5 and 4.0 Hz the tempo-

ral variation in  $Q_c^{-1}$  increased after the mainshock occurrence, where the variation in  $b$ -value was opposite.

Sato [69] analyzed the relation between coda duration time and earthquake magnitude of small earthquakes before and after an M 6.8 earthquake in central Japan. He found that coda durations were anomalously longer than usual for 16 months before the earthquake occurrence. From 24-year observation of coda at 0.5 Hz in Kamchatka Gusev [19] reported prominent anomalies in coda level residual from the mean coda excitation level at 100s lapse time associated with two M 8 earthquakes and a volcanic eruption.

Sawazaki et al. [76] measured the temporal variation of the spectral ratio of coda waves registered on the ground surface to that at the bottom of a borehole of 100 m depth in Japan, which experienced strong ground motion of several hundred gals. They reported a sudden drop of the site amplification factor caused by earthquake strong motion and gradual recovery for a few years approaching to the original ratio. They suggested crack formation and ground water movement for explaining the site factor weakening observed.

### Coda Interferometry

Pairs of earthquakes with almost identical focal mechanisms are called earthquake doublets. The cross-correlation function of earthquake doublet records allow us to detect differences in the background velocity of the Earth medium that took place in between the pair of earthquakes. Applying the phase spectral analysis to coda wave records of earthquake doublets before and after the 1979 Coyote earthquake of M 5.9 in California, Poupinet et al. [57] found that the coda wave arrivals for some stations are progressively delayed for the second earthquake in the doublet. They interpreted systematic variation along the coda as a decrease of background S-wave velocity by 0.2% in an oblong region 5–10 km in radius at the south end of the aftershock zone. Applying the phase spectral analysis to records of repeated artificial explosions, Nishimura et al. [51] found that the average seismic velocity of the crust in the frequency range of 3–6 Hz decreased by about 1% around the focal region of an M 6.1 earthquake at Iwate volcano in northeastern Honshu, Japan in 1998. They interpreted this velocity drop by the dilatation caused by the M 6.1 earthquake with stress sensitivity of the velocity change  $(\delta V/V)/\delta\sigma$  of the order of  $0.1 \text{ MPa}^{-1}$ . From the set of successive artificial explosion experiments, they observed gradual recovery of the seismic velocity towards its original value over the next four years. While interferometry detected a change in velocity of the order of

1%, it was unidentifiable from travel time analysis of first arrivals. Using coda waves is superior to direct waves since coda waves volumetrically sample the Earth medium.

Snieder et al. [81] demonstrated detection of the non-linear dependence of the seismic velocity in granite on temperature and the associated acoustic emissions from the interference measurement of coda waves in rock samples as a laboratory experiment. They named this method “coda interferometry” and proposed to use it for detecting the presence of temporal changes in the medium, or in diagnostic mode. There is an idea to retrieve the Green function from the stacked cross-correlation function (CCF) of multiple scattered waves or microseisms at a pair of stations on condition that the propagation directions of those waves are randomly isotropic. Stacking CCFs of coda waves at several pairs of stations for regional earthquakes in Mexico, Campillo and Paul [7] estimated the surface wave velocity between each station pair from the peak delay. The idea was extended for monitoring the temporal change in the crustal structure. Wegler and Sens-Schönfelder [84] computed the ACF of microseisms recorded at a site in the vicinity of the source region of the 2004 Mid-Niigata earthquake (M 6.6) in Japan for three months. They detected a sudden decrease of relative seismic velocity in the crust of 0.6% at the occurrence of the earthquake from the temporal variation of stacked ACFs.

### Future Directions

In addition to classic parameterization as a layered structure with sharp edges and smooth velocity perturbation, we introduced new approaches using scattered waves that reflect solid Earth heterogeneity. For high-frequency seismograms of earthquakes, envelope characteristics such as the excitation level and the decay gradient of coda envelopes and the envelope broadening of the direct wavelet are useful for the study of small-scale inhomogeneities. The lithospheric inhomogeneity is phenomenologically well characterized by the scattering coefficient and coda attenuation factor as a function of frequency. Furthermore, the power spectral density function of random velocity inhomogeneity is estimated from the frequency dependence of high-frequency seismogram envelopes of local earthquakes or the array analysis of teleseismic waves. The radiative transfer theory with scattering coefficients calculated from the Born approximation and the Markov approximation for the parabolic wave equation are useful mathematical tools for the analyses.

Scattering characteristics are found to vary spatially reflecting seismotectonic settings. It will be necessary for us to make a classification of seismogram-envelope patterns



in various regions in the world under different tectonic conditions. It is interesting to model how such a variation of medium inhomogeneity was created through the geodynamic process. Compared to the lithospheric inhomogeneity, there were insufficient numbers of studies on the mantle inhomogeneity. It will be necessary to map the distribution of inhomogeneities deep in the mantle, which is useful for the study of the evolution of the planet Earth.

For mathematical simplicity; however, most approaches assume homogeneity and isotropy of randomness and a constant background velocity, which are somewhat different from reality. It will be necessary to mathematically develop the envelope synthesis in inhomogeneous media that are a superposition of small-scale random inhomogeneities and a gradually varying background velocity. As revealed from the ray path dependence of S-wave envelope broadening, randomness varies from place to place. It is also necessary to develop the envelope synthesis for random media having spatially varying statistical parameters. In addition, it is important to examine how conversion scattering between P and S waves contributes to form spindle-like envelopes in highly scattering media as shown in high-frequency seismograms observed in volcanoes and on the Moon.

For further understanding, there are monographs that treat the discussed subjects as follows: Sato and Fehler [74] review seismological observation and mathematical models; Shapiro and Hubral [78] put special focus on wave propagation through stratified random media; Goff and Holliger [18] summarize the crustal heterogeneity; Wu and Maupin [86] compile recent developments in mathematical modeling of wave propagation in inhomogeneous media; Chandrasekhar [8] is a classic text for radiative transfer theory; Ishimaru [29] and Rytov et al. [63] offer advanced mathematical tools for the study of wave propagation in random media.

### Acknowledgments

The author thanks Michael Korn, Heiner Igel, and an anonymous reviewer. Their comments were helpful for improving the readability. The author is grateful to NIED, Japan for providing digital seismic data.

### Bibliography

1. Aki K (1969) Analysis of seismic coda of local earthquakes as scattered waves. *J Geophys Res* 74:615–631
2. Aki K (1973) Scattering of P waves under the Montana LASA. *J Geophys Res* 78:1334–1346
3. Aki K (1980) Attenuation of shear-waves in the lithosphere for frequencies from 0.05 to 25 Hz. *Phys Earth Planet Inter* 21:50–60
4. Aki K, Chouet B (1975) Origin of coda waves: Source, attenuation and scattering effects. *J Geophys Res* 80:3322–3342
5. Asano Y, Hasegawa A (2004) Imaging the fault zones of the 2000 western Tottori earthquake by a new inversion method to estimate three-dimensional distribution of the scattering coefficient. *J Geophys Res* 109:B06306. doi:10.1029/2003JB002761
6. Atkinson GM (1993) Notes on ground motion parameters for Eastern North America: Duration and H/V ratio. *Bull Seismol Soc Am* 83:587–596
7. Campillo M, Paul A (2003) Long-Range Correlations in the Diffuse Seismic Coda. *Science* 299:547–549. doi:10.1126/science.1078551
8. Chandrasekhar S (1960) *Radiative Transfer*. Dover, New York
9. Dainty AM, Toksöz MN (1981) Seismic codas on the earth and the moon: A comparison. *Phys Earth Planet Inter* 26:250–260
10. Fehler M, Hoshiya M, Sato H, Obara K (1992) Separation of scattering and intrinsic attenuation for the Kanto-Tokai region, Japan, using measurements of S-wave energy versus hypocentral distance. *Geophys J Int* 108:787–800
11. Fehler M, Sato H, Huang LJ (2000) Envelope broadening of outgoing waves in 2-D random media: A comparison between the Markov approximation and numerical simulations. *Bull Seismol Soc Amer* 90:914–928
12. Flatté SM, Wu RS (1988) Small-scale structure in the lithosphere and asthenosphere deduced from arrival time and amplitude fluctuations at NORSAR. *J Geophys Res* 93:6601–6614
13. Foldy LL (1945) The multiple scattering of waves-I General theory of isotropic scattering by randomly distributed scatterers. *Phys Rev* 67:107–119
14. Frankel A, Clayton RW (1986) Finite difference simulations of seismic scattering: Implications for the propagation of short-period seismic waves in the crust and models of crustal heterogeneity. *J Geophys Res* 91:6465–6489
15. Frankel A, Wennerberg L (1987) Energy-flux model of seismic coda: Separation of scattering and intrinsic attenuation. *Bull Seismol Soc Am* 77:1223–1251
16. Friedrich C, Wegler U (2005) Localization of seismic coda at Merapi volcano (Indonesia). *Geophys Res Lett* 32:L14312. doi:10.1029/2005GL023111
17. Furumura T, Kennett BLN (2005) Subduction zone guided waves and the heterogeneity structure of the subducted plate: intensity anomalies in northern Japan. *J Geophys Res* 110:B10302. doi:10.1029/2004JB003486
18. Goff JA, Holliger K (2002) *Heterogeneity in the Crust and Upper Mantle – Nature, Scaling and Seismic Properties*. Kluwer Academic/Plenum Publishers, Dordrecht, pp 1–358
19. Gusev AA (1995) Baylike and continuous variations of the relative level of the late coda during 24 years of observation on Kamchatka. *J Geophys Res* 100:20311–20319
20. Gusev AA (1995) Vertical profile of turbidity and coda Q. *Geophys J Int* 123:665–672
21. Gusev AA, Abubakirov IR (1987) Monte-Carlo simulation of record envelope of a near earthquake. *Phys Earth Planet Inter* 49:30–36
22. Gusev AA, Abubakirov IR (1996) Simulated envelopes of non-isotropically scattered body waves as compared to observed ones: Another manifestation of fractal heterogeneity. *Geophys J Int* 127:49–60

23. Gusev AA, Lemzikov VK (1985) Properties of scattered elastic waves in the lithosphere of Kamchatka: Parameters and temporal variations. *Tectonophysics* 112:137–153
24. Hemmer PC (1961) On a generalization of Smoluchowski's diffusion equation. *Physica A* 27:79–82
25. Hiramatsu Y, Hayashi N, Furumoto M (2000) Temporal changes in coda Q21 and *b* value due to the static stress change associated with the 1995 Hyogo-ken Nanbu earthquake. *J Geophys Res* 105:6141–6151
26. Holliger K, Levander A (1992) A stochastic view of lower crustal fabric based on evidence from the Ivrea zone. *Geophys Res Lett* 19:1153–1156
27. Hoshiaba M (1991) Simulation of multiple-scattered coda wave excitation based on the energy conservation law. *Phys Earth Planet Inter* 67:123–136
28. Hoshiaba M, Sato H, Fehler M (1991) Numerical basis of the separation of scattering and intrinsic absorption from full seismogram envelope – A Monte-Carlo simulation of multiple isotropic scattering. *Pa Meteorol Geophys, Meteorol Res Inst* 42:65–91
29. Ishimaru A (1978) *Wave Propagation and Scattering in Random Media*, vol 1 and 2. Academic, San Diego
30. Jin A, Aki K (1986) Temporal change in coda *Q* before the Tangshan earthquake of 1976 and the Haicheng earthquake of 1975. *J Geophys Res* 91:665–673
31. Jin A, Aki K (1988) Spatial and temporal correlation between coda *Q* and seismicity in China. *Bull Seismol Soc Am* 78:741–769
32. Jin A, Aki K (1989) Spatial and temporal correlation between coda  $Q^{-1}$  and seismicity and its physical mechanism. *J Geophys Res* 94:14041–14059
33. Jin A, Aki K (2005) High-resolution maps of Coda *Q* in Japan and their interpretation by the brittle-ductile interaction hypothesis. *Earth Planets Space* 57:403–409
34. Korn M (1990) A modified energy flux model for lithospheric scattering of teleseismic body waves. *Geophys J Int* 102:165–175
35. Korn M (1993) Determination of site-dependent scattering *Q* from P-wave coda analysis with an energy-flux model. *Geophys J Int* 113:54–72
36. Korn M, Sato H (2005) Synthesis of plane vector-wave envelopes in 2-D random elastic media based on the Markov approximation and comparison with finite difference simulations. *Geophys J Int* 161:839–848
37. Kubanza M, Nishimura T, Sato H (2006) Spatial variation of lithospheric heterogeneity on the globe as revealed from transverse amplitudes of short-period teleseismic P-waves. *Earth Planets Space* 58:45–e48
38. Langston CA (1979) Structure under Mount Rainer, Washington, inferred from teleseismic body waves. *J Geophys Res* 84:4749–4762
39. Larose E, Margerin L, van Tiggelen BA, Campillo M (2004) Weak Localization of Seismic Waves. *Phys Rev Lett* 93:048501-4. doi:10.1103/PhysRevLett.93.048501
40. Lee WS, Sato H, Lee KW (2003) Estimation of S-wave scattering coefficient in the mantle from envelope characteristics before and after the ScS arrival. *Geophys Res Lett* 30:2248. doi:10.1029/2003GL018413
41. Margerin L (2005) Introduction to radiative transfer of seismic waves. In: Levander A, Nolet G (eds) *Seismic Earth: Array Analysis of Broad-band Seismograms*, Geophysical Monograph Series, vol 157, chap 14. AGU, Washington, pp 229–252
42. Margerin L, Campillo M, van Tiggelen BA (2001) Coherent backscattering of acoustic waves in the near field. *Geophys J Int* 145:593–603
43. Margerin L, Nolet G (2003) Multiple scattering of high-frequency seismic waves in the deep Earth: PKP precursor analysis and inversion for mantle granularity. *J Geophys Res* 108, B11:2514. doi:10.1029/2003JB002455
44. Matsumoto S (2005) Scatterer density estimation in the crust by seismic array processing. *Geophys J Int* 163:622–628
45. Matsumoto S, Obara K, Hasegawa A (1998) Imaging P-wave scatterer distribution in the focal area of the 1995 M7.2 Hyogo-ken Nanbu (Kobe) Earthquake. *Geophys Res Lett* 25:1439–1442
46. Mikada H, Watanabe H, Sakashita S (1997) Evidence for subsurface magma bodies beneath Izu-Oshima volcano inferred from a seismic scattering analysis and possible interpretation of the magma plumbing system of the 1986 eruptive activity. *Phys Earth Planet Inter* 104:257–269
47. Nakahara H, Nishimura T, Sato H, Ohtake M (1998) Seismogram envelope inversion for the spatial distribution of high-frequency energy radiation from the earthquake fault: Application to the 1994 far east off Sanriku earthquake, Japan. *J Geophys Res* 103:855–867
48. Nishigami K (1991) A new inversion method of coda waveforms to determine spatial distribution of coda scatterers in the crust and uppermost mantle. *Geophys Res Lett* 18:2225–2228
49. Nishigami K (2000) Deep crustal heterogeneity along and around the San Andreas fault system in central California and its relation to the segmentation. *J Geophys Res* 105:7983–7998
50. Nishimura T, Fehler M, Baldrige WS, Roberts P, Steck L (1997) Heterogeneous structure around the Jemez Volcanic Field, New Mexico, USA, as inferred from the envelope inversion of active-experiment seismic data. *Geophys J Int* 131:667–681
51. Nishimura T, Tanaka S, Yamawaki T, Yamamoto H, Sano T, Sato M, Nakahara H, Uchida N, Hori S, Sato H (2005) Temporal changes in seismic velocity of the crust around Iwate volcano, Japan, as inferred from analyses of repeated active seismic experiment data from 1998 to 2003. *Earth Planets Space* 57:491–505
52. Nishimura T, Yoshimoto K, Ohtaki T, Kanjo K, Purwana I (2002) Spatial distribution of lateral heterogeneity in the upper mantle around the western Pacific region as inferred from analysis of transverse components of teleseismic P-coda. *Geophys Res Lett* 29:2089–2137. doi:10.1029/2002GL015606
53. Obara K, Sato H (1995) Regional differences of random inhomogeneities around the volcanic front in the Kanto-Tokai area, Japan, revealed from the broadening of S wave seismogram envelopes. *J Geophys Res* 100:2103–2121
54. Paaschens JCJ (1997) Solution of the time-dependent Boltzmann equation. *Phys Rev E* 56:1135–1141
55. Petukhin AG, Gusev AA (2003) The Duration-distance Relationship and Average Envelope Shapes of Small Kamchatka Earthquakes. *Pure Appl Geophys* 160:1717–1743
56. Phillips WS, Aki K (1986) Amplification of coda waves from local earthquakes in Central California. *Bull Seismol Soc* 76:627–648
57. Poupinet G, Ellsworth WL, Frechet J (1984) Monitoring velocity variations in the crust using earthquake doublets: An application to the Calaveras fault, California. *J Geophys Res* 89:5719–5731

58. Przybilla J, Korn M, Wegler U (2006) Radiative transfer of elastic waves versus finite difference simulations in two-dimensional random media. *J Geophys Res* 111:B04305. doi:10.1029/2005JB003952
59. Revenaugh J (1995) A scattered-wave image of subduction beneath the Transverse Ranges. *Science* 268:1888–1892
60. Revenaugh J (1995) Relationship of the 1992 Landers, California, earthquake sequence to seismic scattering. *Science* 270:1344–1347
61. Revenaugh J (1999) Geologic Applications of Seismic Scattering. *Annu Rev Earth Planet Sci* 27:55–73
62. Rondenay S, Bostock MG, Shragge J (2001) Multiparameter two-dimensional inversion of scattered teleseismic body waves 3. Application to the Cascadia 1993 data set. *J Geophys Res* 106:30795–30807
63. Rytov SM, Kravtsov YA, Tatarskii VI (1987) Principles of Statistical Radio Physics, vol 4, Wave Propagation Through Random Media. Springer, Berlin
64. Ryzhik LV, Papanicolaou GC, Keller JB (1996) Transport equations for elastic and other waves in random media. *Wave Motion* 24:327–370
65. Saito T (2006) Synthesis of scalar-wave envelopes in two-dimensional weakly anisotropic random media by using the Markov approximation. *Geophys J Int* 165:501–515. doi:10.1111/j.1365-246X.2006.02896.x
66. Saito T, Sato H, Ohtake M (2002) Envelope broadening of spherically outgoing waves in three-dimensional random media having power-law spectra. *J Geophys Res* 107:2089. doi:10.1029/2001JB000264
67. Sato H (1977) Single isotropic scattering model including wave conversions: Simple theoretical model of the short period body wave propagation. *J Phys Earth* 25:163–176
68. Sato H (1984) Attenuation and envelope formation of three-component seismograms of small local earthquakes in randomly inhomogeneous lithosphere. *J Geophys Res* 89:1221–1241
69. Sato H (1987) A precursor-like change in coda excitation before the western Nagano earthquake ( $M_s = 6.8$ ) of 1984 in central Japan. *J Geophys Res* 92:1356–1360
70. Sato H (1989) Broadening of seismogram envelopes in the randomly inhomogeneous lithosphere based on the parabolic approximation: Southeastern Honshu, Japan. *J Geophys Res* 94:17735–17747
71. Sato H (1990) Unified approach to amplitude attenuation and coda excitation in the randomly inhomogeneous lithosphere. *Pure Appl Geophys* 132:93–121
72. Sato H (2006) Synthesis of vector wave envelopes in three-dimensional random elastic media characterized by a Gaussian autocorrelation function based on the Markov approximation: Plane wave case. *J Geophys Res* 111:B06306. doi:10.1029/2005JB004036
73. Sato H (2007) Synthesis of vector-wave envelopes in 3-D random elastic media characterized by a Gaussian autocorrelation function based on the Markov approximation: Spherical wave case. *J Geophys Res Solid Earth* 112:B01301. doi:10.1029/2006JB004437
74. Sato H, Fehler M (1998) *Seismic Wave Propagation and Scattering in the Heterogeneous Earth*. AIP Press/Springer, New York
75. Sato H, Nakahara H, Ohtake M (1997) Synthesis of scattered energy density for non-spherical radiation from a point shear dislocation source based on the radiative transfer theory. *Phys Earth Planet Inter* 104:1–281
76. Sawazaki K, Sato H, Nakahara H, Nishimura T (2006) Temporal Change in Site Response Caused by Earthquake Strong Motion as Revealed from Coda Spectral Ratio Measurement. *Geophys Res Lett* 33:L21303. doi:10.1029/2006GL027938
77. Shang T, Gao L (1988) Transportation theory of multiple scattering and its application to seismic coda waves of impulsive source. *Sci Sin* 31B:1503–1514
78. Shapiro SA, Hubral P (1999) *Elastic Waves in Random Media – Fundamentals of Seismic Stratigraphic Filtering*. Springer, Berlin
79. Shearer PM, Earle PS (2004) The global short-period wavefield modeled with a Monte Carlo seismic phonon method. *Geophys J Int* 158:1103–1117
80. Shiomi K, Sato H, Ohtake M (1997) Broad-band power-law spectra of well-log data in Japan. *Geophys J Int* 130:57–64
81. Snieder R, Gret A, Douma A, Scales J (2002) Coda wave interferometry for estimating nonlinear behavior in seismic velocity. *Science* 295:2253–2255
82. Sreenivasiah I, Ishimaru A, Hong ST (1976) Two-frequency mutual coherence function and pulse propagation in a random medium: An analytic solution to the plane wave case. *Radio Sci* 11:775–778
83. Takahashi T, Sato H, Nishimura T, Obara K (2006) Strong inhomogeneity beneath Quaternary volcanoes revealed from the peak delay analysis of S-wave seismograms of microearthquakes in northeastern, Japan. *Geophys J Int* 168:90–99. doi:10.1111/j.1365-246X.2006.03197.x
84. Wegler U, Sens-Schönfelder C (2007) Fault zone monitoring with passive image interferometry. *Geophys J Int* 168:1029–1033. doi:10.1111/j.1365-246X.2006.03284.x
85. Wu RS (1985) Multiple scattering and energy transfer of seismic waves – separation of scattering effect from intrinsic attenuation – I Theoretical modeling. *Geophys J R Astron Soc* 82:57–80
86. Wu RS, Maupin V (eds) (2007) *Advances in Wave Propagation in Heterogeneous Earth*. In: Dmowska R (ed) *Advanced in Geophysics*, vol 48. Academic Press, San Diego, pp 561–596
87. Yomogida K, Benites R (1995) Relation between direct wave Q and coda Q: A numerical approach. *Geophys J Int* 123:471–483
88. Yoshimoto K (2000) Monte-Carlo simulation of seismogram envelope in scattering media. *J Geophys Res* 105:6153–6161
89. Yoshimoto K, Sato H, Ohtake M (1993) Frequency-dependent attenuation of P and S waves in the Kanto area, Japan, based on the coda-normalization method. *Geophys J Int* 114:165–174
90. Yoshimoto K, Sato H, Ohtake M (1997) Short-wavelength crustal inhomogeneities in the Nikko area, central Japan, revealed from the three-component seismogram envelope analysis. *Phys Earth Planet Inter* 104:63–73
91. Yoshimoto K, Wegler U, Korn M (2006) A volcanic front as a boundary of seismic attenuation structures in northeastern Honshu, Japan. *Bull Seismol Soc Am* 96:637–646
92. Zeng Y, Su F, Aki K (1991) Scattering wave energy propagation in a random isotropic scattering medium I Theory. *J Geophys Res* 96:607–619
93. Zhao D, Hasegawa A, Horiuchi S (1992) Tomographic imaging of P and S wave velocity structure beneath Northeastern Japan. *J Geophys Res* 97:19909–19928

## Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space

GERT ZÖLLER<sup>1</sup>, SEBASTIAN HAINZL<sup>2</sup>,  
YEHUDA BEN-ZION<sup>3</sup>, MATTHIAS HOLSCHNEIDER<sup>1</sup>

<sup>1</sup> Institute of Mathematics and Centre for Dynamics of  
Complex Systems, University of Potsdam,  
Potsdam, Germany

<sup>2</sup> GFZ German Research Centre for Geosciences,  
Potsdam, Germany

<sup>3</sup> Department of Earth Sciences, University of Southern  
California, Los Angeles, USA

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Modeling Seismicity in Real Fault Regions](#)

[Results](#)

[Summary and Conclusions](#)

[Future Directions](#)

[Acknowledgments](#)

[Bibliography](#)

### Glossary

**Bayesian analysis** A model estimation technique that accounts for incomplete knowledge. Bayes' theorem is a mathematical formulation of how an a priori estimate of the probability of an event can be updated, if a new information becomes available.

**Critical earthquake concept** The occurrence of large earthquakes may be described in terms of statistical physics and thermodynamics. In this view, an earthquake can be interpreted as a critical phase transition in a system with many degrees of freedom. The preparatory process is characterized by acceleration of the seismic moment release and growth of the spatial correlation length as in the percolation model. This interpretation of earthquake occurrence is referred to as the critical earthquake process.

**Earthquake forecast/prediction** The forecast or prediction of an earthquake is a statement about time, hypocenter location, magnitude, and probability of occurrence of an individual future event within reasonable error ranges.

**Fault model** A fault model calculates the evolution of slip, stress, and related quantities on a fault segment or a fault region. The range of fault models varies from

conceptual models of cellular automaton or slider-block type to detailed models for particular faults.

**Probability** A quantitative measure of the likelihood for an outcome of a random process. In the case of repeating a random experiment a large number of times (e. g. flipping a coin), the probability is the relative frequency of a possible outcome (e. g. head). A different view of probability is used in the → Bayesian analysis.

**Seismic hazard** The probability that a given magnitude (or peak ground acceleration) is exceeded in a seismic source zone within a pre-defined time interval, e. g. 50 years, is denoted as the seismic hazard.

### Self-organized criticality

Self-organized criticality (SOC) as introduced by Bak [2] is the ability of a system to organize itself in the vicinity of a critical point independently of values of physical parameters of the system and initial conditions. Self-organized critical systems are characterized by various power law distributions. Examples include models of sandpiles and forest-fires.

### Definition of the Subject

The most fundamental question in earthquake science is whether earthquake prediction is possible. Related issues include the following: Can a prediction of earthquakes solely based on the emergence of seismicity patterns be reliable? In other words, is there a single or several “magic” parameters, which become anomalous prior to a large earthquake? Are pure observational methods without specific physical understanding, like the pattern recognition approach of Keilis–Borok and co-workers [41], sufficient? Taking into account that earthquakes are monitored continuously only for about 100 years and the best available data sets (“earthquake catalogs”) cover only a few decades, it seems questionable to forecast earthquakes solely on the basis of observed seismicity patterns. This is because large earthquakes have recurrence periods of decades to centuries; consequently, data sets for most regions include less than ten large events making a reliable statistical testing questionable.

In the studies discussed here, the goal is not to forecast individual earthquakes. Instead, we aim at developing a combined approach based on numerical modeling and data analysis in order to understand seismicity and the emergence of patterns in the occurrence of earthquakes. The discussion and interpretation of seismicity in terms of statistical physics leads to the concept of “critical states”, i. e. states in the seismic cycle with an increased probability for abrupt changes involving large earthquakes. A more general goal of this work is to provide perspectives for

the understanding of the relevant mechanisms and to give outlines for developments related to time-dependent seismic hazard.

## Introduction

Several empirical relationships for the occurrence of seismicity are well-known. The most common one is probably the Gutenberg–Richter law [30] for the relation between frequency and magnitude of earthquakes in a large seismically active region,

$$\log N = a - bM, \quad (1)$$

where  $N$  is the frequency of earthquakes with magnitude equal to or greater than  $M$ ;  $a$  is a measure of the overall seismicity level in the region and the  $b$  value determines the relation between large and small earthquakes. The Gutenberg–Richter law provides an important constraint for the design of physical models and serves as a key ingredient for seismic hazard estimations. Statistical relations for the temporal occurrence of large events are less well known, because the corresponding data records are too short.

Several additional problems exist in the understanding and interpretation of observed seismicity patterns. First, it is important to decide whether an observed pattern has a physical origin or is an artifact, arising for example from inhomogeneous reporting or from man-made seismicity like quarry blasts or explosions [69]. Second, the non-artificial events have to be analyzed with respect to their underlying mechanisms. This leads to an inverse problem with a non-unique solution, which can be illustrated for the most pronounced observed temporal pattern associated with aftershocks. It is empirically known that the earthquake rate  $\dot{N}$  after a large event at time  $t_M$  follows the Omori–Utsu law [49,67]

$$\dot{N} = \frac{K}{(c + t - t_M)^p}, \quad (2)$$

where  $t$  is the time,  $K$  and  $c$  are constants, and the Omori exponent  $p$  is close to unity. In particular, aftershocks are an almost universal phenomenon; that is, they are observed nearly after each mainshock. The underlying mechanisms leading to aftershocks are, however, unknown. Various physical models have been designed to explain aftershock occurrence following Eq. (2). These models include viscoelasticity [32], pore fluid flow [46], damage rheology [9,57], and rate-state friction [24]. The question which mechanism or combination of mechanisms is relevant in a given fault zone remains open. Detailed comparisons of observed and modeled seismicity with respect to

the aftershock rate, the duration of aftershock sequences, the dependence on the mainshock size, and other features are necessary to address this problem. Additionally, results from lab experiments on rupture dynamics and satellite observations of deformation provide important information for the design of such models.

Apart from aftershock activity, other seismicity patterns are occasionally associated with observations, including foreshocks [39], seismic quiescence [34,72,78], and accelerating moment release [17,38]. These patterns have been documented in several cases before large earthquakes. They occur, however, far less frequently than aftershocks. For example, foreshocks are known to precede only 20–30% of large earthquakes [71]. Therefore, their predictive power is questionable. Moreover, it is not clear whether these patterns can be attributed to physical processes or to random fluctuations in the highly sparse and noisy earthquake catalogs. This problem can be addressed by using fault models which simulate long and complete earthquake sequences over thousands of years. If the models capture the main features of the underlying physics, the occurrence of seismicity patterns can be studied with reasonable statistics. The main ingredients of such models are the geometry of a fault region, empirically known constitutive laws, spatial heterogeneities, and stress and displacement functions following dislocation theory [20,47]. In order to allow for detailed studies of the relations between the imposed mechanisms and the observed seismicity functions, it is important that the number of adjustable parameters is limited.

It is emphasized that these models do not aim to reproduce an observed earthquake catalog in detail. Instead, the main goal is to address questions like: Why is the Parkfield segment of the San Andreas fault characterized by relatively regular occurrence of earthquakes with magnitude  $M \approx 6$ , while on the San Jacinto fault in California the properties of earthquake occurrence are more irregular? Basic models for seismicity are mainly based on one or more solid blocks, which are driven by a plate over a sliding surface. The plate and the blocks are connected with springs. This model can generate stick-slip events considered to represent earthquakes. The slider-block models can produce a wide range of complexity, beginning with a single block model leading to periodic occurrence of events of uniform size, and progressing to an array of connected blocks [18] leading to complex sequences of events with variable size. In order to reduce the computational effort cellular automata are commonly used [42,48]. Mathematically, these models include maps instead of differential equations; physically, this corresponds to instantaneously occurring slip events, neglecting inertia effects.

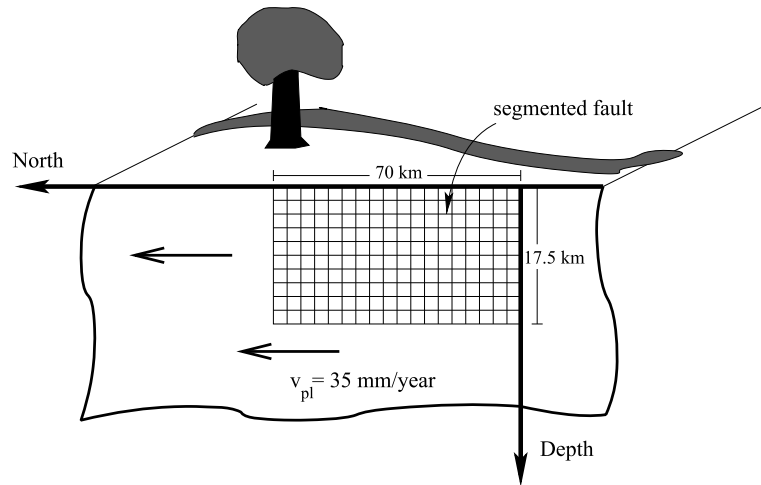
The main ingredients of slider-block and cellular automaton models are (1) external driving (plate motion), and (2) sudden local change of system parameters (stress), when a critical value (material strength) is reached, followed by an avalanche of block slips (stress drop and coseismic stress transfer during an earthquake). While the first process lasts for years to several hundred years, the second occurs on a time scale of a few seconds. The simplest model including these features has been formulated by Reid [52] and is known as *Reid's elastic rebound theory*; in terms of slider-block models, this corresponds to a single block model with constant plate velocity. Accounting for spatial heterogeneity and fault segmentation, many interacting blocks, or fault segments, have to be considered. This leads to a spatiotemporal stress field instead of a single stress value. In general, the material strength will also become space-dependent. Such a model framework can be treated with the methodology of statistical physics similar to the Ising model or percolation models [43]. In this context, large earthquakes are associated with second-order phase transitions [2,59,64]. The view of earthquakes as phase transitions in a system with many degrees of freedom and an underlying critical point, is hereinafter referred to as the “critical point concept”. The period before such a phase transition is expected to be characterized by a preparation process including development of power laws and growing spatial correlation length [14]. However, depending on the parameters of a model, different situations are conceivable: the system trajectory can enter the critical state and the critical point frequently (“supercritical”) or it may never become critical (“subcritical”). A case of special interest is the class of models [2] showing *self-organized criticality* (SOC), which have their origin in a simple cellular automaton model for a sandpile [3]. In this case the system drives itself permanently to the vicinity of the critical point with almost scale-free characteristics. Consequently, each small event can grow into a large earthquake with some probability [28].

Long simulations of earthquake activity can be used to calculate statistical features like the recurrence time distribution of large earthquakes and the frequency-size distribution with high precision. Despite the scaling behavior (Eq. (1)) in the earthquake magnitudes for small and intermediate earthquakes, which is observed for many sets of model parameters, clear deviations become visible for large magnitudes. Such deviations are known from real catalogs, but their statistical significance is not clear in all cases. The model simulations suggest that deviations from scaling for strong earthquakes can be attributed to physical properties. One important property is the spatial disorder of brittle parameters of the fault. The presence

of strong heterogeneities suppresses system-wide events with some probability, whereas such events can evolve more easily on smooth faults. The degree of quenched (time-independent) spatial heterogeneity turns out to be a key parameter for statistical and dynamical properties of seismicity [5,12,80]. This includes the temporal regularity of mainshock occurrence, various properties of the stress and displacement fields, and a spontaneous mode-switching between different dynamical regimes without changing parameters. The degree of heterogeneity can act as a tuning parameter that allows for a continuous change of the model dynamics between the end-member cases of supercritical and subcritical behavior. Such a dependence, which is observed also for other parameters, can be visualized in phase diagrams similar to the phase diagram for the different aggregate states of water [22,79,80]. For increasing complexity of a model, the number of axes of the phase diagram, representing the relevant model parameters, will increase. The above mentioned question of distinguishing different faults like the Parkfield segment and the San Jacinto fault can be rephrased as the problem of assigning the faults to different regions in such a diagram. An important step in this direction is the physical modeling of observed seismicity patterns, including universal patterns like aftershocks (Eq. (2)), common fluctuations like foreshocks and the acceleration of seismic energy release before large earthquakes. The latter phenomenon which sometimes occurs over large regions including more than one fault, can be interpreted in terms of the approach towards a critical point. This view is supported by an observational study of the growth of the spatial correlation length which is a different aspect of the same underlying physics [73,75,76,77].

The establishment of relationships between model parameters and observational features may be used to tune the model towards a specific fault zone, and use the tuned fault model for practical applications of seismic hazard estimations. Toward this end the recurrence time distribution of large earthquakes is needed. Since observational data records are often short and noisy, the use of Bayesian probability theory is helpful for the estimation of uncertain model parameters, and the incorporation of various types of observational data in seismic hazard estimations. The Parkfield segment, as one of the best monitored seismically-active regions, serves as an excellent natural laboratory for such a case study. A discussed example illustrates how partially known parameters like the stress drop and the seismic hazard can be estimated by combining numerical models and observational data [74].

In Section “**Modeling Seismicity in Real Fault Regions**”, the physical fault model used for the discussed studies is described. Results from numerical simulations



Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 1  
Sketch of the fault model framework

are presented in Sect. “Results”. A summary is given in Sect. “Summary and Conclusions”.

### Modeling Seismicity in Real Fault Regions

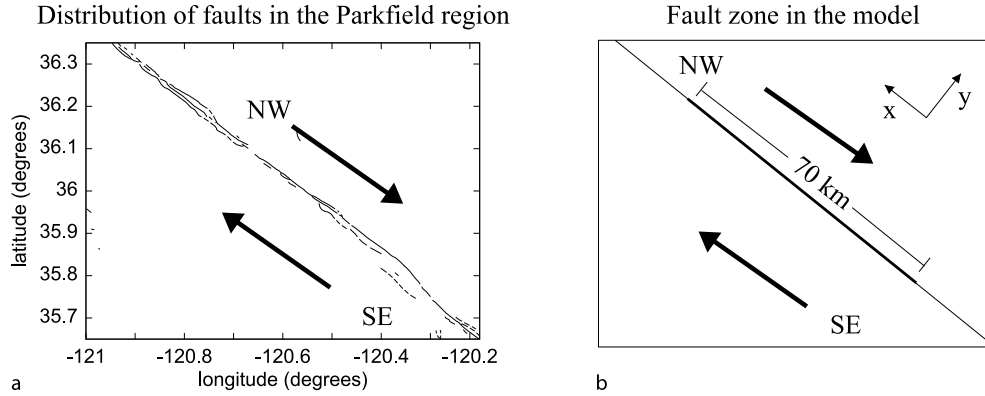
Numerous frameworks have been used to simulate seismicity (see e.g. [6,9,10,18,32,37,48] and references therein). These include slider-block models, cellular automata, “inherently discrete” fault models where the discreteness is an inherent feature of the imposed physics, and continuum models. In this section we illustrate how a fault model (Fig. 1) can be adjusted in order to simulate seismicity of a real fault region, e.g. the Parkfield segment of the San Andreas fault in California.

### Fault Geometry and Model Framework

A first constraint for a specific model is to represent the geometry of the fault segment. As shown in Fig. 2, the region of Parkfield is characterized by a distribution of fault segments, which have generally the same orientation. It is therefore reasonable to map these segments in the model on a plane intersecting the surface at a straight line from SE to NW. The dimensions of the fault segment for modeling (Fig. 1) are chosen to be 70 km in length and 17.5 km in depth. As discussed in [10], this geometry corresponds approximately to the San Andreas fault near Parkfield. The entire fault is an infinite half-plane, but the brittle processes are calculated on the above rectangular section referred to below as the computational grid. The computational grid is discretized to  $128 \times 32$  cells of uniform size, where stress and slip are calculated. The size of the cells is

not related here to observations; rather it depends on the magnitude range under consideration and the computational effort. The failure of a single cell defines the lowest magnitude. A higher resolution of the grid with same overall dimensions increases the magnitude range, because the magnitude is calculated from the slip of all cells during an earthquake. Following Ben-Zion and Rice [10], the material surrounding the fault is assumed to be a homogeneous elastic half space, which is characterized by elastic parameters and a related Green’s function:

1. The elastic properties are expressed by the Lamé constants  $\lambda$  and  $\mu$ , which connect stress and strain in Hooke’s law. For many rocks, these constants are almost equal; therefore we use  $\lambda = \mu$  with  $\mu$  being the rigidity. An elastic solid with this property is called a *Poisson solid*. Because the strain is dimensionless,  $\mu$  has the same dimension as the stress. In the present study, we use  $\mu = 30$  GPa.
2. The (static) Green’s function  $G(\vec{y}_1, \vec{y}_2)$  defines the static response of the half space at a position  $\vec{y}_1$  to a displacement at  $\vec{y}_2$ , which may arise from (coseismic) slip or (aseismic) creep motion. Due to the discretization of the fault plane into computational cells, we use the Green’s function for static dislocations on rectangular fault patches of width  $dx$  and height  $dz$ , which is given in [20] and [47]. The main difference between this Green’s function and the nearest-neighbor interaction of most slider-block models and cellular automata is the infinite-range interaction following a decay according to  $1/r^3$ , where  $r$  is the distance between source cell and receiver point.



Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 2  
**a** Distribution of faults in the Parkfield (California) region; **b** fault region in the model

### Interseismic Processes

The motion of the tectonic plates, indicated in Fig. 2, is responsible for the build-up of stress in the fault zone. Geodetic measurements of surface displacements provide estimates of the velocity of the plates. For the San Andreas fault, a value of  $v_{pl} = 35$  mm/year as a long-term average [55] is widely accepted and is adopted for the model. The displacement  $du$  in the regions surrounding the grid during a time period  $\Delta t$  is simply  $du = v_{pl} \cdot \Delta t$ . While the average slip rate  $\dot{u}$  is independent of the location of a cell, the stress rate  $\dot{\tau}$  depends on space. The assumption that the computational grid is embedded in a half-plane which undergoes constant creep, implies that cells at the boundaries of the grid experience higher load than cells in the center of the grid. The Green's function  $G(i, j; k, l)$  defines the interaction of points  $(i, j)$  and  $(k, l)$  in the medium. In particular, the stress response at a position  $(i, j)$  to a static change of the displacement field  $du(k, l)$  is given by

$$d\tau(i, j) = - \sum_{(k, l) \in \text{half space}} G(i, j; k, l) \cdot du(k, l), \quad (3)$$

where the minus sign stems from the fact that forward (right-lateral) slip of regions around a locked fault segment is equivalent to back (left-lateral) slip of the locked fault segment. Taking into account that

$$\sum_{(k, l) \in \text{half space}} G(i, j; k, l) = 0, \quad (4)$$

Eq. (3) can be written as

$$\tau(i, j; t) = - \sum_{(k, l) \in \text{half space}} G(i, j; k, l) \cdot [u(k, l; t) - v_{pl}t], \quad (5)$$

where  $u(k, l; t)$  is the total displacement at position  $(k, l)$  and time  $t$  since the start of the simulation. Because the surrounding regions sustain stable sliding,  $u(k, l; t) = v_{pl}t$  for  $(k, l) \notin \text{grid}$ , the slip deficit outside the fault region vanishes and it is sufficient to perform the summation on the computational grid:

$$\tau(i, j; t) = \sum_{(k, l) \in \text{grid}} G(i, j; k, l) \cdot [v_{pl}t - u(k, l; t)]. \quad (6)$$

Equation (6) can be decomposed to a part for the tectonic loading and a residual part for slip on the computational grid. The tectonic loading follows the formula

$$\tau_{\text{load}}(i, j; t) = \gamma(i, j) \cdot t \quad (7)$$

with a space-dependent but time-independent loading rate

$$\gamma(i, j) = v_{pl} \cdot \sum_{(k, l) \in \text{grid}} G(i, j; k, l). \quad (8)$$

The build-up of stress may be reduced by aseismic creep motion, which is implemented by a local constitutive law corresponding to lab-based dislocation creep [5]:

$$\dot{u}_{\text{creep}}(i, j; t) = c(i, j) \cdot \tau^3(i, j; t) \quad (9)$$

with space dependent but time-independent creep coefficients  $c(i, j)$ .

### Friction and Coseismic Stress Transfer; Quasidynamic Approach

It is widely accepted that earthquakes on large faults are due to frictional processes on pre-existing structures. The friction is therefore an important empirical ingredient of a fault model [56]. Numerous laboratory experiments



have been carried out to characterize frictional behavior of different materials (see e.g. [19]). An important finding is that the friction coefficient defined as the ratio of shear stress  $\tau_{\text{shear}}$  and compressional normal stress  $\tau_{\text{normal}}$ ,  $\mu_f = \tau_{\text{shear}}/\tau_{\text{normal}}$  at the initiation of slip is approximately constant for many materials; the value of  $\mu_f$  lies between 0.6 and 0.85. This observation, known as *Byerlee's law*, is related to the Coulomb failure criterion [16] for the Coulomb stress CS,

$$CS = \tau_{\text{shear}} - \mu_f \tau_{\text{normal}}. \quad (10)$$

The Coulomb stress depends on a plane where shear stress and normal stress are calculated. Neglecting cohesion, the Coulomb criterion for brittle failure is

$$CS \geq \tau_0, \quad (11)$$

which for  $CS = 0$  is Byerlee's law.

The North-American plate and the Pacific plate move in opposite directions along the fault plane having strike-slip motion. The absence of normal and thrust faulting reduces the problem to a one-dimensional motion: all parts of the fault move along the fault direction. The stress state of the fault is fully determined by the shear stress  $\tau_{xy}$  in the coordinates given in Fig. 2b. Slip is initiated if  $\tau_{xy}$  exceeds  $\mu_f \tau_{yy}$ . This quantity, which is called the static strength  $\tau_s$  is constant in time if  $\mu_f$  is assumed to be constant. Note that the normal stress on a planar fault in a homogeneous solid does not change [1]. The shear stress  $\tau_{xy}$  will be denoted simply by  $\tau$ . In this notation, the failure criterion Eq. (11) reduces to

$$\tau \geq \tau_s. \quad (12)$$

When a cell  $(k, l)$  fails, the stress drops in this cell to the arrest stress  $\tau_a$ :

$$\tau(k, l) \rightarrow \tau_a, \quad (13)$$

where the value  $\tau_a$  maybe space-dependent. The stress change produces a corresponding slip

$$du(k, l) = \frac{\tau(k, l) - \tau_a}{G(k, l; k, l)} \quad (14)$$

with the self-stiffness  $G(k, l; k, l)$  of cell  $(k, l)$ .

The observational effect of dynamic weakening includes also a strength drop from the static strength to a lower dynamic strength:

$$\tau_s \rightarrow \tau_d. \quad (15)$$

In particular, slipping material becomes weaker during rupture and recovers to the static level at the end of the

rupture. This approximation of the strength evolution is known as static-kinetic friction.

The values  $\tau_s$ ,  $\tau_d$ , and  $\tau_a$  are connected by the dynamic overshoot coefficient  $D$ :

$$D = \frac{\tau_s - \tau_a}{\tau_s - \tau_d}, \quad (16)$$

or alternatively by the dynamic weakening coefficient  $\varepsilon$ :

$$\varepsilon = 1 - \frac{\tau_d}{\tau_s}. \quad (17)$$

Following [10] we use in most simulations  $D = 1.25$ .

The redistribution of the stress release  $\Delta\tau(k, l) = \tau(k, l) - \tau_a$  from cell  $(k, l)$  to a point  $(i, j)$  at time  $t$  is

$$d\tau(i, j; t) = G(i, j; k, l) \cdot \delta\left(t - \frac{r(i, j; k, l)}{v_s}\right) \cdot \frac{\Delta\tau(k, l)}{G(k, l; k, l)}, \quad (18)$$

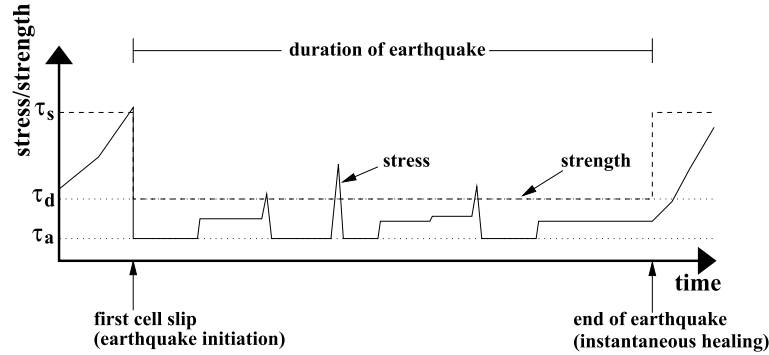
where  $\delta(x)$  denotes the  $\delta$ -function, which is 1 for  $x = 0$  and 0 else;  $v_s$  is the shear-wave velocity, and  $r(i, j; k, l)$  is the distance between source cell  $(k, l)$  and receiver position  $(i, j)$ . That is, regions far from the slipping cell receive their stress portion later than regions close to the slipping cell. The value of  $v_s$  is assumed to be constant. Each "stress transfer event" associated with Eq. (18) gives a transfer of a stress  $d\tau$  from a source cell  $(k, l)$  to a receiver cell  $(i, j)$  at time  $t$ . This time-dependent stress transfer is called the *quasidynamic* approach in contrast to the *quasistatic* approach used in most similar models.

The evolution of stress and strength in a failing cell is shown schematically in Fig. 3. When the slip is initiated, both the stress and the strength drop. Due to coseismic stress transfer during the event, the cell may slip several times before the earthquake is terminated and instantaneous healing takes place in all cells. The piecewise constant failure envelope (dashed line) indicates static-kinetic friction. A model version with gradual healing was employed by [79]. A review of analytical results associated with the basic model in the context of a large universality class is given in ► [Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of.](#)

We note that the Green's function leads to an infinite interaction range. Using open boundary conditions with respect to the computational grid, the stress release from a slipping cell is not conserved on the grid, but on the infinite half plane.

## Data

The model produces two types of data, earthquake catalogs and histories of stress and displacement fields. As demon-



Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 3  
Pictorial evolution of stress (solid line) and strength (dashed line) of a hypocenter cell in the quasidynamic approach

strated below, all parameters of the model have physical dimensions and can therefore be compared directly with real data, where they are available. This is in contrast to most of the slider-block and cellular automaton models.

Earthquake catalogs include values of the earthquake time, hypocenter, and size. The time of an earthquake is the time of the first slip; the hypocenter is determined by the position of the corresponding cell along strike and depth. The size of an event can be described by different measures: The rupture area  $A$  is the total area, which slipped during an earthquake. The potency

$$P = \int u(A)dA \quad (19)$$

measures [7] the total slip during the event and is related to the seismic moment  $m_0$  by the rigidity:  $m_0 = \mu P$ . The (moment) magnitude  $M$  can be calculated from the potency [10] using

$$M = (2/3) \log_{10}(P) + 3.6, \quad (20)$$

where  $P$  is given in  $\text{cm} \cdot \text{km}^2$ .

## Results

Numerous simulations of the model described in the previous section have been performed. Firstly, simulations have been examined with respect to the spatiotemporal propagation of stress during single earthquakes (“rupture histories”). Then, long deformation histories have been simulated in order to search in a large fraction of the parameter space for relationships between input parameters and observed seismicity features. In this section, a selection of key results is presented and discussed in relation to critical states of seismicity.

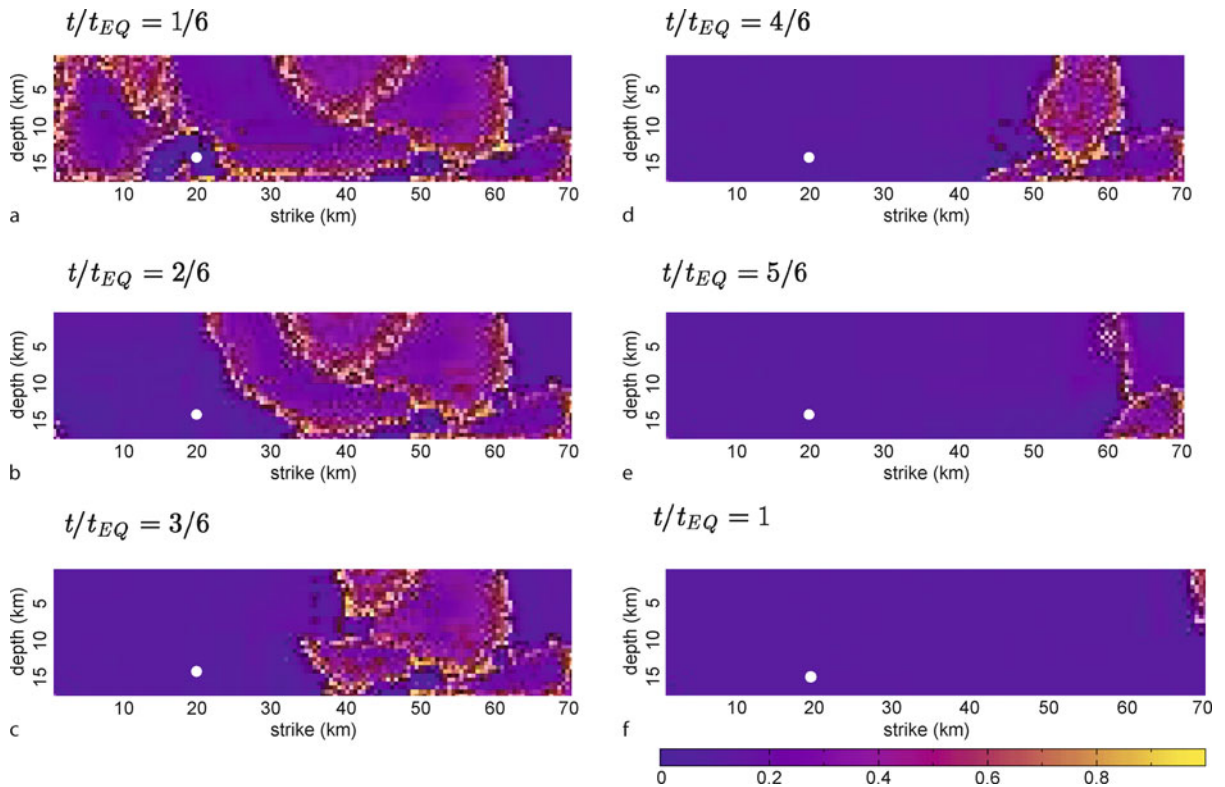
## Rupture Histories

We compare qualitatively rupture histories of large earthquakes for three end-member cases in parameter space:

- (1) a smooth fault,
- (2) a rough fault, and
- (3) a fault without dynamic weakening ( $\tau_d = \tau_s$  or  $D \rightarrow \infty$  in Eq. (16)).

Following [5], we vary the degree of quenched spatial disorder for a particular realization by introducing barriers of high stress drop  $\tau_s - \tau_a$  in an environment of low stress drop.

The observation that smooth faults show a more regular earthquake occurrence than rough faults, can be explained by the ability of the stress field to synchronize on certain fault patches. On a disordered fault, this type of synchronization is unlikely. Figure 4a shows the stress field (normalized between 0 and 1) immediately before a large earthquake on a smooth fault. The most striking feature is the emergence of clearly defined patches with highly loaded boundaries. During rupture evolution, these patches rupture almost in series until the fault is nearly unloaded (see Fig. 4b–f). A different situation is shown in Fig. 5, corresponding to a rough fault with creep coefficients  $c(i, j)$  (Eq. (9)) that increase with depth leading to a brittle-ductile transition zone as in [5] and [80]. Here the stress field in the brittle regime is irregular without obvious pattern formation. Similar behavior is found in a case where dynamic weakening is switched off ( $D \rightarrow \infty$ ); in other words, the material heals instantaneously. Figure 6 shows the stress field in this case. In analytical studies, it has been shown that this corresponds exactly to a critical point in a phase diagram spanned by the stress dissipation and dynamic weakening [22,27]. Observational results indicate [68] that irregular slip histories and power



Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 4

Snapshots of rupture evolution for a system-wide event on a smooth fault without creep motion.  $t$  denotes the time after the rupture initiation (given in units of the total earthquake duration  $t_{EQ}$ ). The white circle is the hypocenter of the event. The figure shows the dimensionless stress state  $\hat{\tau} = \frac{\tau - \tau_a}{\tau_s - \tau_a}$  of the cells. **a**  $t/t_{EQ} = 1/6$ ; **b**  $t/t_{EQ} = 2/6$ ; **c**  $t/t_{EQ} = 3/6$ ; **d**  $t/t_{EQ} = 4/6$ ; **e**  $t/t_{EQ} = 5/6$ ; **f**  $t/t_{EQ} = 1$

law frequency-size distributions are associated with geometrically disordered fault structures, while characteristic earthquake statistics and overall regular ruptures are found on mature fault with large total displacements.

Although the stress field shows a complex evolution during a simulation, the presence or absence of characteristic length scales indicating the relation to a critical point is easily detected. From an observational point of view, the stress field is not accessible. The evolution of the displacement field may be estimated, e. g. from seismic and geodetic data using slip inversion techniques. Because of the high uncertainties in the calculated slip histories, a quantitative comparison of the simulated data with “natural” data is questionable. However, general features of the quasidynamic ruptures are quite realistic, e. g. the irregular patterns in Fig. 5 resemble the rupture of the Chi–Chi (Taiwan) earthquake on September 21, 1999 ( $M_w = 7.6$ ) [58].

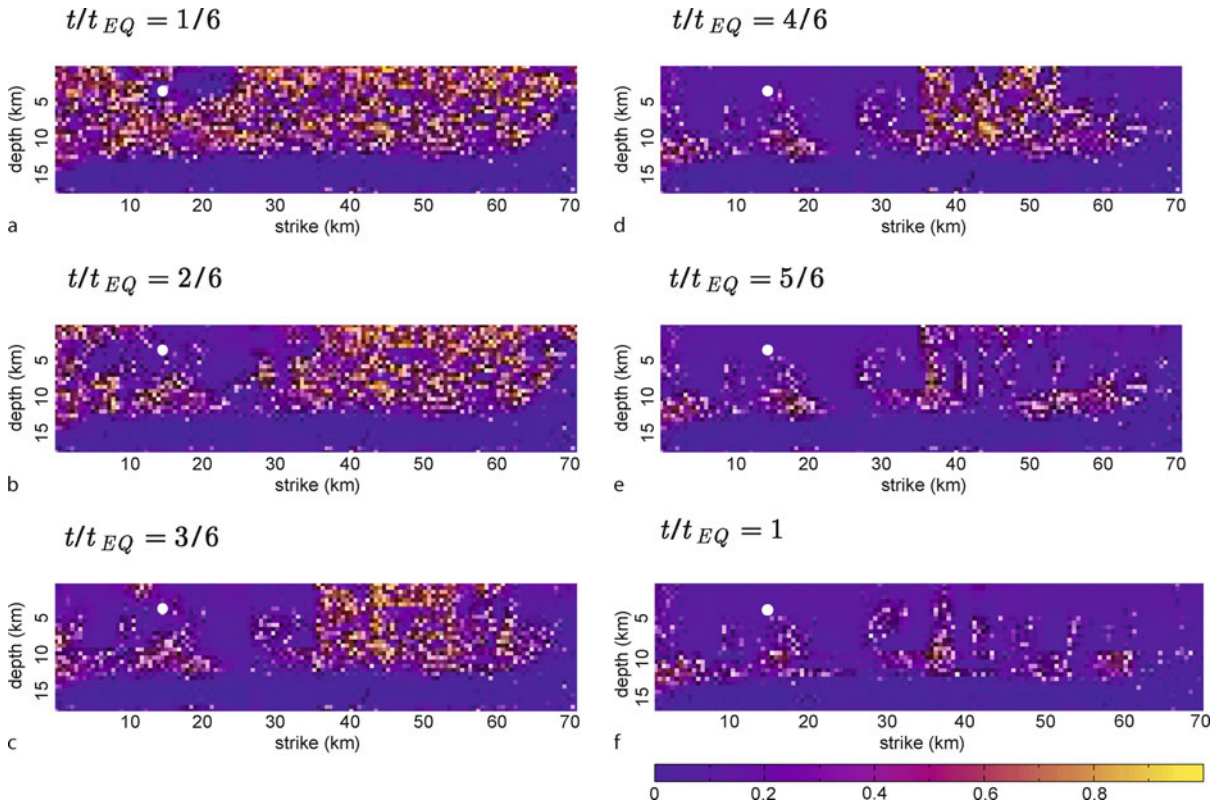
Later we will show that the frequency-size distribution of earthquakes can serve to some extent as a proxy for the degree of disorder of the stress field. Ben-Zion et al. [12]

discuss additional seismicity functions that may be used as surrogate variables for the stress.

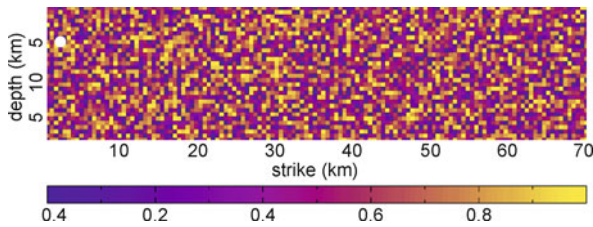
### Frequency–Size Distributions

The frequency-size (FS) distribution is one of the most important characteristics of observed seismicity. For worldwide seismicity, this distribution is given by the Gutenberg–Richter law (Eq. (1)). Figure 7 shows the FS distribution of California seismicity from 1970 to 2004. Here we use the non-cumulative version of Eq. (1), where  $N$  is the number of earthquakes with magnitude between  $M$  and  $M + dM$  with a magnitude bin  $dM = 0.1$ .

For individual fault zones the FS distribution can deviate from Eq. (1), especially for high magnitudes. Examples are given in Fig. 8, which shows the FS distribution of the Parkfield segment (Fig. 8a) and for the San Jacinto fault (Fig. 8b) in California for a time span of 45 years. The distribution of the Parkfield segment consists of two parts: A scaling regime for  $2.2 \leq M \leq 4.5$  and a “bump” for  $4.5 < M \leq 6.0$ . For the San Jacinto fault, the scaling range



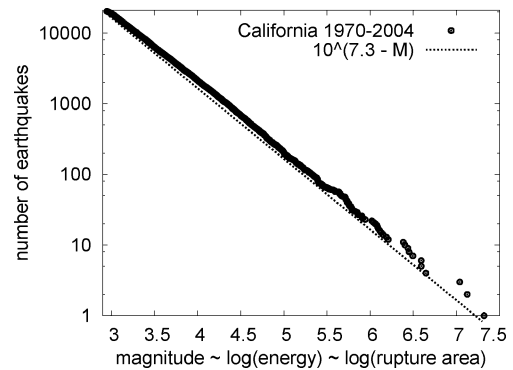
Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 5  
 Same as Fig. 4 for a strongly disordered fault with a brittle-ductile transition at about 15 km depth. **a**  $t/t_{EQ} = 1/6$ ; **b**  $t/t_{EQ} = 2/6$ ; **c**  $t/t_{EQ} = 3/6$ ; **d**  $t/t_{EQ} = 4/6$ ; **e**  $t/t_{EQ} = 5/6$ ; **f**  $t/t_{EQ} = 1$



Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 6  
 Same as Fig. 4a for a fault without dynamic weakening ( $\tau_d = \tau_s$ ) corresponding to a dynamic overshoot coefficient  $D \rightarrow \infty$  (Eq. (16))

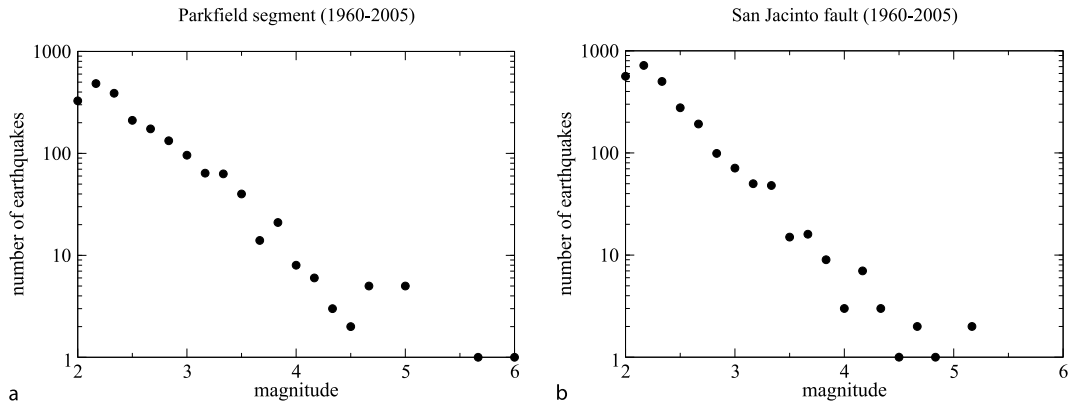
is observed for almost all events ( $2.2 \leq M \leq 5.0$ ). The decrease for  $M \approx 2$  in both plots is probably due to catalog incompleteness.

The FS distribution as shown in Fig. 8a is called the characteristic earthquake distribution, because of the increased probability for the occurrence of large (“characteristic”) events compared to the prediction of the Gutenberg–Richter relation. The latter is an exponential dis-

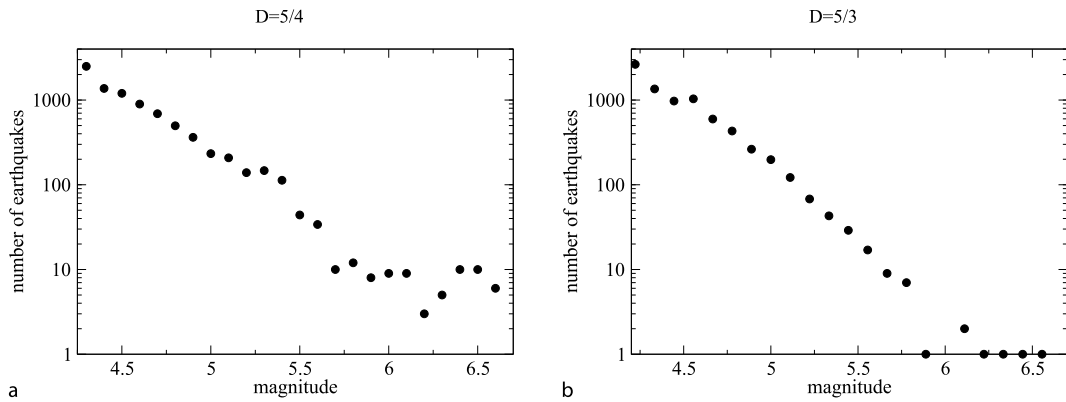


Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 7  
 Frequency-size distribution for California from 1970 to 2004; the dashed line denotes a power law fit to the data

tribution for the earthquake frequency as a function of magnitude, or a power law distribution for the earthquake frequency as a function of potency (Eqs. (19), (20)), moment, energy, or rupture area, over a broad range of mag-



Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 8  
Frequency-size distribution for two faults in California: **a** the Parkfield segment, and **b** the San Jacinto fault



Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 9  
Frequency-size distribution for model realizations with different dynamic overshoot coefficients (Eq. (16)): **a**  $D = 5/4$ , **b**  $D = 5/3$

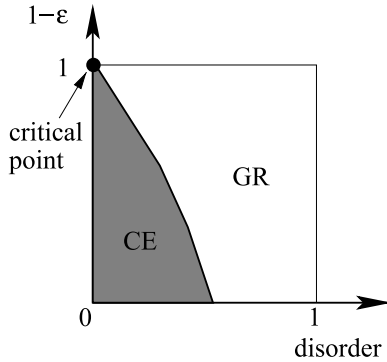
nitudes [66]. The Gutenberg–Richter relation is “scale-free” because a power law distribution indicates the absence of a characteristic scale of the earthquake size [64]. In terms of critical point processes, the absence of a characteristic length scale indicates that the system is close to the critical point. In this state, earthquakes of all magnitudes can occur, or each small rupture can grow into a large one. Therefore, the frequency-size distribution can serve as a proxy for the current state of a system in relation to a critical point.

The FS distribution in a model can be tuned by varying the mean stress  $\langle \tau \rangle$  on the fault, where  $\langle \rangle$  denotes the spatial average. This can be achieved, for instance, by varying brittle properties, e.g. in terms of the dynamic overshoot coefficient  $D$  (Eq. (16)), or by introducing dissipation [31,79]. Figure 9 shows FS distributions for two different values of  $D$ :  $D = 5/4$  (Fig. 9a) and a higher value  $D = 5/3$  (Fig. 9b). While Fig. 9a follows a characteristic

earthquake behavior similar to the Parkfield case (Fig. 8a), Fig. 9b resembles the shape of the FS distribution of the San Jacinto fault (Fig. 8b).

As an outcome, three cases can be distinguished by means of a critical mean stress  $\tau_{\text{crit}}$ :

1. subcritical fault ( $\langle \tau \rangle < \tau_{\text{crit}}$ ): the mean stress on the fault is too small to produce large events. The system is always far from the critical point. The FS distribution is a truncated Gutenberg–Richter law.
2. supercritical fault ( $\langle \tau \rangle > \tau_{\text{crit}}$ ): the mean stress is high and produces frequently large events. After a large earthquake (critical point), the stress level is low (system is far from the critical point) and recovers slowly (approaches the critical point). The FS distribution is a characteristic earthquake distribution.
3. critical fault ( $\langle \tau \rangle \approx \tau_{\text{crit}}$ ): the system is always close to the critical point with scale-free characteristics. The FS

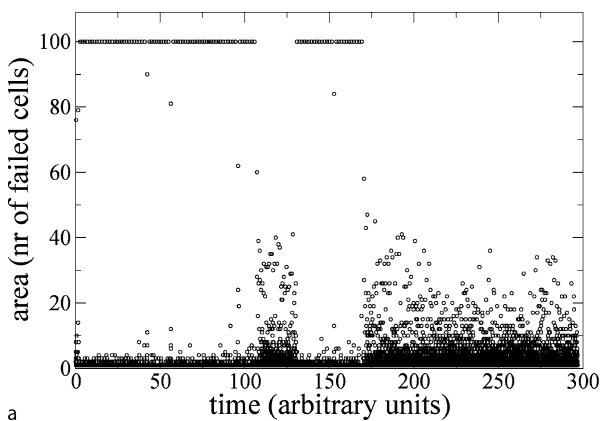


Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 10

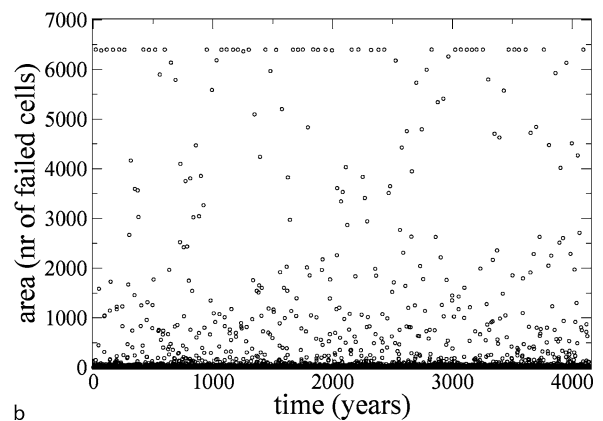
Phase diagram for the frequency-size distribution (GR=Gutenberg-Richter distribution, CE=characteristic earthquake distribution) spanned by the degree of quenched spatial disorder and the dynamic weakening represented by  $\varepsilon$ . The upper left corner corresponds exactly to a critical point [22,27] and results in scale-free characteristics as shown in Fig. 6

distribution is a Gutenberg-Richter law with a scaling range over all magnitudes.

If the FS distribution is plotted as a function of the model parameters, the result can be visualized by a phase diagram [22,31,79,80]. An example is provided in Fig. 10, which shows schematically the phase diagram spanned by the degree of quenched spatial disorder and  $1 - \varepsilon$  with the dynamic weakening coefficient  $\varepsilon$  (Eq. (17)). The phase diagram summarizes results from various studies, which demonstrate that the degree of spatial disorder of the stress drop acts as a tuning parameter for the FS distribution [5,36,80].

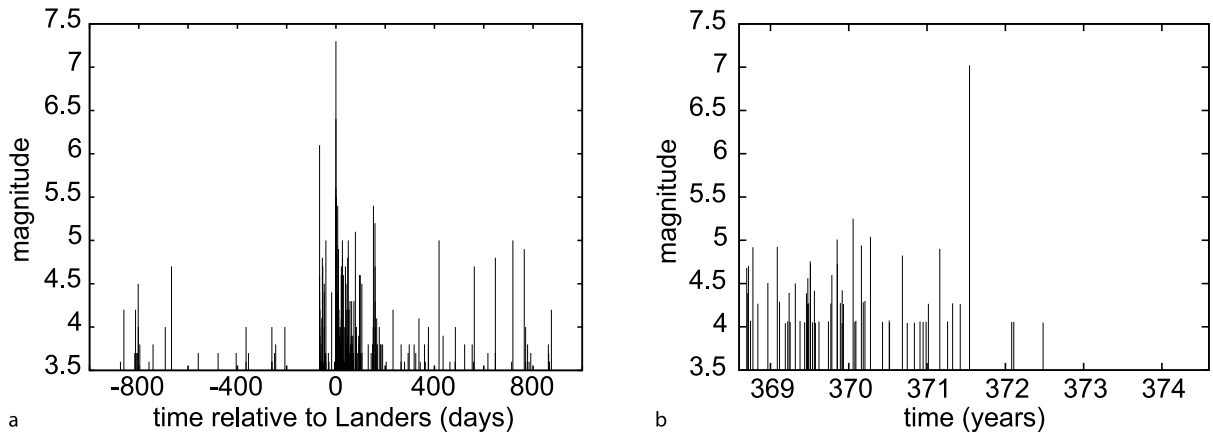


If the model is in the transition regime between Gutenberg-Richter statistics and characteristic earthquake behavior, the ability of the stress field to synchronize on parts of the fault can have additional impact on the dynamics of seismicity: for a model with small cells and high stress fluctuations along the cell boundaries arising from a high degree of spatial disorder, the system can undergo a spontaneous transition from an ordered state and characteristic behavior to a disordered state following Gutenberg-Richter statistics. Due to the high fluctuations in the stress field, there is some probability that a certain number of cells synchronize by chance, leading to an ordered behavior for some seismic cycles, until the order is destroyed, again resulting from stress fluctuations. This type of mode-switching behavior has been observed earlier in a mean-field model and a damage rheology model [11,22]. Figure 11a gives a corresponding earthquake sequence with spontaneous mode-switching behavior. Figure 11b shows a sequence calculated with a higher grid resolution ( $128 \times 50$  cells). The tendency to mode-switching is less pronounced, but still visible. In [79] it is shown that the emergence of such mode-switching depends both on the spatial range of interaction (given as the decay of the Green's function) and the discretization of the computational grid. In the less realistic model of [22], where the stress redistribution is governed by a constant (space-independent) Green's function, analytical expressions for persistence times have been calculated [27]. In [11] some evidence for mode-switching behavior in long seismic records based on paleoseismic and geologic data from the Dead Sea fault and other regions are discussed. However, the relevance of mode-switching to natural seismicity remains unclear due to the general lack of very long data records.



Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 11

Earthquake area (measured as the number of failed cells) as a function of time a in the mean-field model of Dahmen et al. [22] for a fault with 100 cells and b in the elastic model with  $128 \times 50$  cells



Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 12  
 Earthquakes before and after a mainshock: **a** the  $M7.3$  Landers (California) earthquake; **b**  $M7.3$  earthquake in the basic version of the model

### Aftershocks and Foreshocks

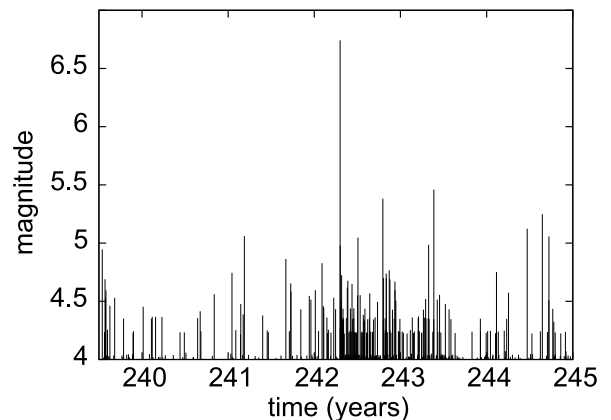
The most pronounced temporal pattern in observed seismicity is the emergence of strongly clustered aftershock activity following a large earthquake. Apart from the Omori–Utsu law (Eq. (2)), it is widely accepted that aftershocks are characterized by the following properties:

1. The aftershock rate scales with the mainshock size [51].
2. Aftershocks occur predominantly around the edges of the ruptured fault segments [66].
3. Båth's law [4]: The magnitude of the largest aftershock is usually about one unit smaller than the mainshock magnitude.

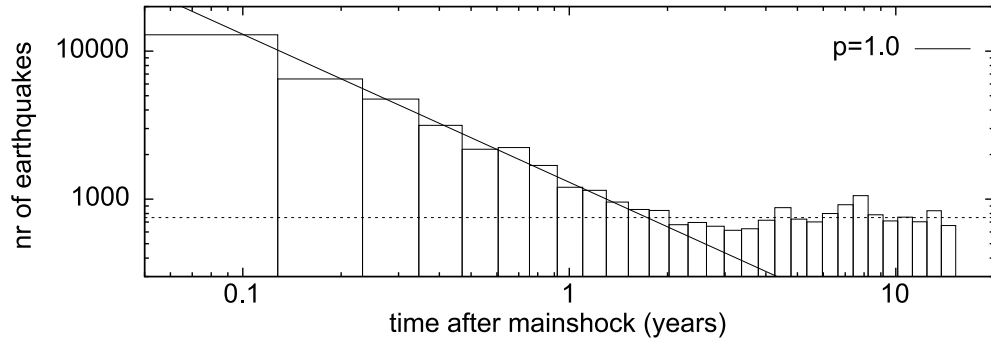
Deviations from the Omori–Utsu law, especially for rough faults, are discussed in [45]. While aftershocks are observed after almost all large earthquakes, foreshocks occur less frequent [71]. As a consequence, much less is known about the properties of these events. Kagan and Knopoff [40] and Jones and Molnar [39] propose a power law increase of activity according to an “inverse” Omori–Utsu law.

Figure 12a shows an example for the aftershock sequence following the  $M7.3$  Landers earthquake in California on June 28, 1992. An earthquake of similar size generated by the model is given in Fig. 12b. The absence of aftershocks in the simulation is clearly visible. The reason for the lack of aftershocks is the unloading of the fault resulting from the mainshock: When a large fraction of the fault has ruptured, the stress in this region will be close to the arrest stress after the event. Consequently, the seismicity rate will be almost zero until the stress field has recovered to a moderate level.

Discussions for likely mechanisms of aftershocks are given in [9] and [81]. A common feature is the presence of rapid postseismic stress which generates aftershock activity. In [32], for instance, postseismic stress has been attributed to a viscoelastic relaxation process following the mainshock. In the work discussed here, continuous creep displacement following Eq. (9) is assumed. Additionally, the computational grid is divided by aseismic barriers from the free surface to depth into several seismically active fault segments. As shown in [81], this modification results in a concentration of stress in the aseismic regions during rupture and, subsequently in a release of stress after the event according to the coupled creep



Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 13  
 Earthquakes before and after a mainshock with  $M = 6.8$  in the modified model



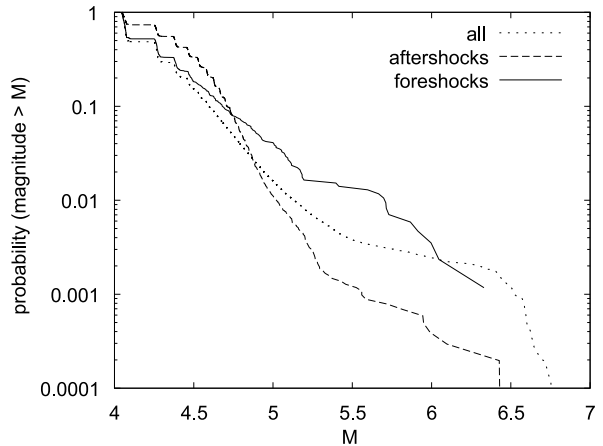
Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 14

Earthquake rate as a function of time for the model with seismic and aseismic regions. The calculation is based on a simulation with 200,000 earthquakes covering about 5000 years; the earthquake rates are averaged over about 300 mainshocks. A fit of the Omori-Utsu law (Eq. (2)) with  $p = 1$  is denoted as a solid line. The dashed line gives the estimated background level of seismicity

process. This stress release triggers aftershock sequences obeying the Omori-Utsu law (Eq. (2)). A typical aftershock sequence after a  $M_{6.8}$  event is shown in Fig. 13. In agreement with Båth's law, the strongest aftershock has the magnitude  $M = 5.5$  in this sequence. The sequence shows also the effect of secondary aftershocks, namely aftershocks of aftershocks [61]. The stacked earthquake rate as a function of the time after the mainshock is given in Fig. 14. In this case, where the barriers are characterized by creep coefficients higher by a factor of  $10^5$  than in the other patches, a realistic Omori exponent of  $p = 1$  is found.

Aftershock sequences like in Fig. 13 emerge after all large events in the extended model. In contrast, there is generally no clear foreshock signal visible in single sequences. However, stacking many sequences together unveils a slight increase of the earthquakes rate prior to a mainshock supporting the observation of accelerating foreshock activity. An explanation of these events can be given in the following way: Between two mainshocks the stress field organizes itself towards a critical state, where the next large earthquake can occur. This critical state is characterized by a disordered stress field and the absence of a typical length scale, where earthquakes of all sizes can occur [12]. The mainshock may occur immediately or after some small to moderate events. The latter case can be considered as a single earthquake, which is interrupted in the beginning. This phenomenon of delayed rupture propagation has also provided a successful explanation of foreshocks and aftershocks in a cellular automaton model [33,35].

The hypothesis that foreshocks occur in the critical point and belong, in principle, to the mainshock, can be verified by means of the findings from Subsect. "Frequency-Size Distributions". In particular, the frequency-size distribution in the critical point (or close to the



Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 15

Frequency-magnitude distributions of all earthquakes, foreshocks and aftershocks. Foreshocks and aftershocks are defined as earthquakes occurring within one month before and after an earthquake with  $M \geq 6$

critical point) is expected to show scale-free statistics. If an overall smooth fault model following characteristic earthquake statistics is studied over a long time period, the approach of the critical point should be seen in terms of a change of the frequency-size distribution towards Gutenberg-Richter behavior [12]. This change of frequency-size statistics is indeed observed in the model (Fig. 15) and supports the validity of the critical point concept [82].

### Accelerating Moment Release

In the previous section, it has been argued that large earthquakes are associated with a critical point and the prepa-



ration process is characterized by increasing disorder of the stress field and increasing tendency for scale-free characteristics in the frequency-size distribution. Further support for critical point dynamics has been provided by the observational finding of [17] that the cumulative Benioff strain  $\Sigma \Omega(t)$  follows a power law time-to-failure relation prior to the *M*7.1 Loma Prieta earthquake on October 17, 1989:

$$\Sigma \Omega(t) = \sum_{i=1}^{N(t)} \sqrt{E_i} = A - B(t_f - t)^m \tag{21}$$

Here,  $E_i$  is the energy release of earthquake  $i$ ,  $N(t)$  is the number of earthquakes before time  $t$ ;  $t_f$  is the failure time and  $A$ ,  $B$ , and  $m > 0$  are constants. Similar studies for numerous seismically active regions followed (see [8,77] and references therein).

The time-to-failure relation Eq. (21) has been explained by [54] and [60] from the viewpoint of renormalization theory and by [8] and [65] in terms of damage rheology. Similar to the findings about foreshocks, the time-to-failure pattern is not universal. Therefore, a stacking procedure is adopted in order to obtain robust results on the validity of Eq. (21) in the model. This is not straightforward, since the interval of accelerating moment release is not known a priori and the duration of a whole seismic cycle, as an upper limit, is not constant. To normalize the time interval for the stacking, the potency release (Eq. (19)) is computed as a function of the (normalized) stress level (Fig. 16). Taking into account that the stress level increases almost linearly during a large fraction of

the seismic cycle, the stress level axis in Fig. 16 can effectively be replaced by the time axis leading to a power law dependence of the potency release on time. The best fit is provided with an exponent  $s = -1.5$ . Transforming the potency release to the cumulative Benioff strain (Eq. (21)), results in an exponent  $m = 0.25$  in Eq. (21). This finding is based on a simulation over 5000 years; the exponent is in good agreement with the theoretical work [53], that derives  $m = 0.25$  for a spinodal model, and the analytical result of  $m = 0.3$  in the damage mechanics model [8]. An observational study of California seismicity finds  $m$  between 0.1 and 0.55 [15].

### Interevent Times

In recent studies it has been shown that the distribution of interevent times can be described by a universal law. In particular, the distributions from different tectonic environments, different spatial scales (from worldwide to local seismicity) and different magnitude ranges collapse if the time  $\Delta t$  is rescaled with the rate  $R_{xy}$  of seismic occurrence in a region denoted by  $(x, y)$  [21]. Such rescaling leads to

$$D_{xy}(\Delta t) = R_{xy} \cdot f(R_{xy} \Delta t), \tag{22}$$

where  $D_{xy}$  is the probability density for the interevent time  $\Delta t$ , and  $f$  can be expressed by a generalized gamma distribution

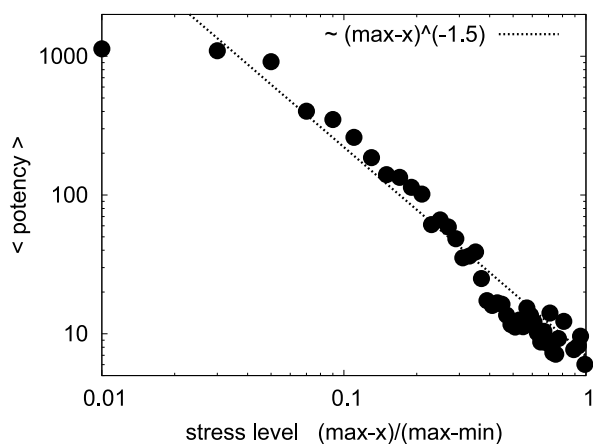
$$f(\theta) = C \frac{1}{\theta^{\gamma-1}} \exp\left(-\frac{\theta^\delta}{B}\right). \tag{23}$$

The parameters  $C$ ,  $\gamma$ ,  $\delta$ , and  $B$  have been determined by a fit to several observational catalogs [21].

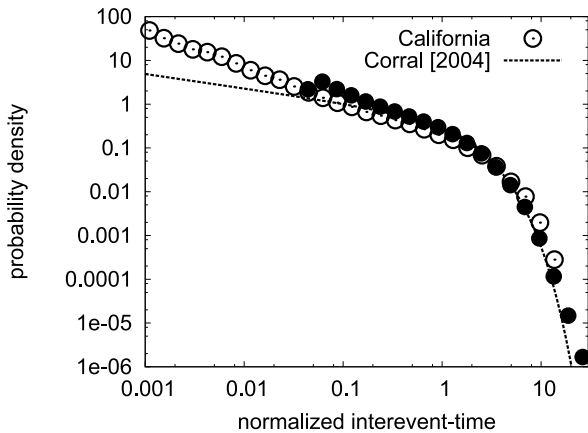
In Fig. 17 we compare  $D_{xy}(\Delta t)$  from Eq. (22) with two earthquake catalogs: (1) The ANSS catalog of California (catalog ranges are given in the caption), and (2) the model catalog. Due to the universality of Eq. (22) with respect to different spatial scales, the comparison of the model simulating a single fault of 70 km length with a region of hundreds of kilometers including several faults in California does not require coarse graining the ANSS catalog. In the region where the interevent times are calculated, we find a remarkable agreement of the three curves. For small values of  $\Delta t$ , Eq. (22) deviates from the California data; for high values the model has a slightly better correspondence with the observational data than Eq. (22). Thus the results generally support the recent findings of [21].

The degree of temporal clustering of earthquakes can be estimated by the coefficient of variation  $CV$  of the interevent time distribution.

$$CV = \sigma/\mu, \tag{24}$$



Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 16  
**Mean potency release (Eq. (19)) as a function of the stress level. The stress level is normalized to the maximum (max) and minimum (min) observed stress**

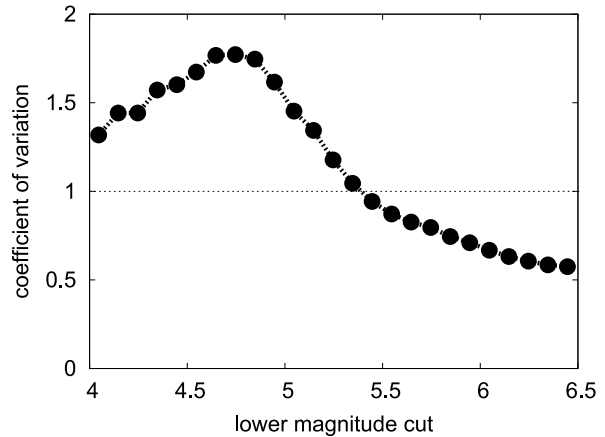


Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 17

The normalized interevent time distribution of the model simulations (black dots) compared with the result of [21] and the distribution of earthquakes in California (ANSS catalog of  $M \geq 3$  earthquakes occurred between 1970 and 2004 within  $29^\circ$  and  $43^\circ$  latitude and  $-113^\circ$  and  $-123^\circ$  longitude)

where  $\sigma$  is the standard deviation and  $\mu$  the mean value of the interevent time distribution. Values of  $CV > 1$  denote clustered activity, while  $CV < 1$  represents quasiperiodic occurrence of events. The case  $CV = 1$  corresponds to a random Poisson process. The studies of [5] and [80] have found that the clustering properties of the large events depend on the degree of quenched spatial disorder of the fault. Figure 18 shows that  $CV$  as a function of the lower magnitude cutoff has a characteristic shape. The values of  $CV$  are higher than 1 (clustered) for small and intermediate earthquakes ( $M \leq 5.4$ ) and smaller than 1 (quasiperiodic) for larger earthquakes. This corresponds to the case of a low degree of disorder in [80], because the brittle cells which participate in an earthquake have no significant spatial disorder. We note that this behavior resembles the seismicity on the Parkfield segment of the San Andreas fault, which is characterized by a quasiperiodic occurrence of mainshocks. Based on the analysis of 37 earthquake sequences, an estimation of  $CV \approx 0.5$  has been found for multiple tectonic environments [26].

A different behavior is observed on the San Jacinto fault in California, where the largest events occur less regularly and have overall smaller magnitudes. As discussed in [5] and [80], this can be modeled by imposing higher degrees of disorder leading to a broader range of spatial size scales, e. g. by using a higher number of near-vertical barriers. While barriers provide a simple and physically motivated way to tune the degree of disorder, other types of heterogeneities may work as well, as long as they are



Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 18

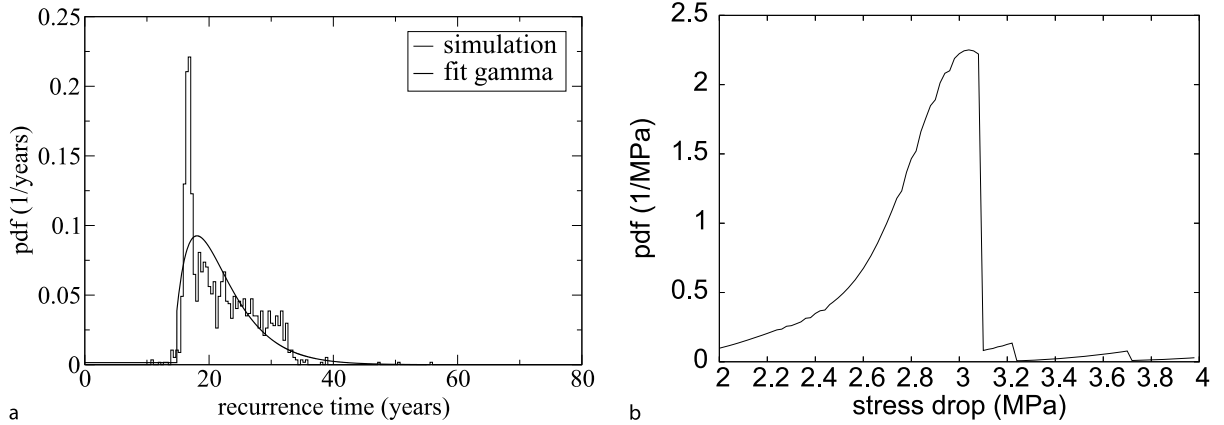
The temporal earthquake occurrence quantified by the coefficient of variation as a function of the lower magnitude cutoff. Values larger than 1 indicate clustering, whereas lower values point to quasiperiodic behavior

able to produce strong enough fluctuations of the stress field. As an example, we mention fractal distributions of the stress drop, which can be tuned easily by changing the fractal dimension [63,80].

### Recurrence Times of Large Earthquakes

While interevent times are waiting times between successive earthquakes in a given catalog, recurrence times are defined as waiting times between two successive *large* events, typically in the magnitude range  $6 \leq M \leq 9$ , depending on the region. For example, on the Parkfield segment of the San Andreas fault seven  $\sim M6$  earthquakes occurred between 1857 and 2004 with recurrence times  $T_1 = 24$  years,  $T_2 = 20$  years,  $T_3 = 21$  years,  $T_4 = 12$  years,  $T_5 = 32$  years, and  $T_6 = 38$  years.

The distribution of recurrence times of large earthquakes is crucial for the calculation of seismic hazard. Due to a lack of observational data, this distribution is unknown for real fault systems. Commonly used distributions are based on extreme value statistics and on models for catastrophic failure. These include the lognormal distribution [50], the Brownian passage time distribution [44], and the Gumbel distribution [29]. All distributions are characterized by a maximum for a certain recurrence time followed by an asymptotic decay. The Brownian passage time distribution and the lognormal distribution have been used by the Working Group on California Earthquake Probabilities [70], e. g. for calculating earthquake probabilities in the San Francisco Bay area.



Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 19

**a** Approximated probability density function of the recurrence time distribution of large earthquakes ( $M > 6.2$ ) for a simulated earthquake catalog and fit with a truncated Gamma distribution; **b** A posteriori distribution  $P(\Delta\tau|T_1, \dots, T_N)$  of stress drop  $\Delta\tau$  calculated with Bayes' theorem (Eq. (25))

Figure 19a shows the probability density function (pdf) of the recurrence times of earthquakes with magnitude  $M > 6.2$  in a realization of the numerical model for the Parkfield region [74]. Since we focus on long time-scales, we use here a minimal model without aseismic creep and strong spatial heterogeneities. This model leads to characteristic earthquake statistics and quasiperiodic occurrence of large events, and can therefore serve as a model framework for large earthquakes on the Parkfield segment. However, quantities which are only poorly known from empirical data, e.g. the stress drop, have to be chosen in order to perform a numerical simulation. Starting with an imposed uniform a priori distribution  $P(\Delta\tau)$  of stress drops between a lower bound  $\Delta\tau_{\min}$  and an upper bound  $\Delta\tau_{\max}$ , an a posteriori distribution  $P(\Delta\tau|T_1, \dots, T_N)$  can be estimated using observational recurrence times  $T_1, \dots, T_N$  from Parkfield and Bayes' theorem [13],

$$P(\Delta\tau|T_1, \dots, T_N) = \frac{P(T_1, \dots, T_N|\Delta\tau)P(\Delta\tau)}{\sum_{s=\Delta\tau_{\min}}^{\Delta\tau_{\max}} P(T_1, \dots, T_N|s)P(s)}, \quad (25)$$

with the likelihood function

$$P(T_1, \dots, T_N|\Delta\tau) = \prod_{i=1}^N f(T_i|\Delta\tau). \quad (26)$$

The function  $f(T_i|\Delta\tau)$  denotes the pdf of recurrence times simulated with a model stress drop  $\Delta\tau$ . To get an analytic expression of this function, it is fitted by a Gamma distribution  $f(t) = \beta^{-1}(\Gamma(\gamma))^{-1}(\frac{t-\mu}{\beta})^{\gamma-1} \exp(-\frac{t-\mu}{\beta})$  with

the location parameter  $\mu$ , the shape parameter  $\gamma \equiv 2.0$  and the scale parameter  $\beta$  (with  $x \geq \mu$ ;  $\gamma, \beta > 0$ ). For an example see Fig. 19a. In [74] it is shown that the mean value  $\mu_t$  and the standard deviation  $\sigma_t$  of the fits in this model are related to the average stress drop of a large earthquake  $\Delta\tau$  by the simple empirical relations

$$\begin{aligned} \mu_t(\Delta\tau) &= 9.7 \cdot \Delta\tau \\ \sigma_t(\Delta\tau) &= 1.8 \cdot \Delta\tau^2 - 6.8 \cdot \Delta\tau + 11.7 \end{aligned} \quad (27)$$

with  $\mu_t, \sigma_t$  in years and  $\Delta\tau$  in MPa. Using this approximation in combination with six observational recurrence times from  $\sim M6$  earthquakes on the Parkfield segment, we find the a posteriori distribution of stress drops shown in Fig. 19b. The position where this distribution reaches the maximum,  $\Delta\tau = (3.04 \pm 0.27)$  MPa, is the most representative value of the stress drop of  $\sim M6$  Parkfield events.

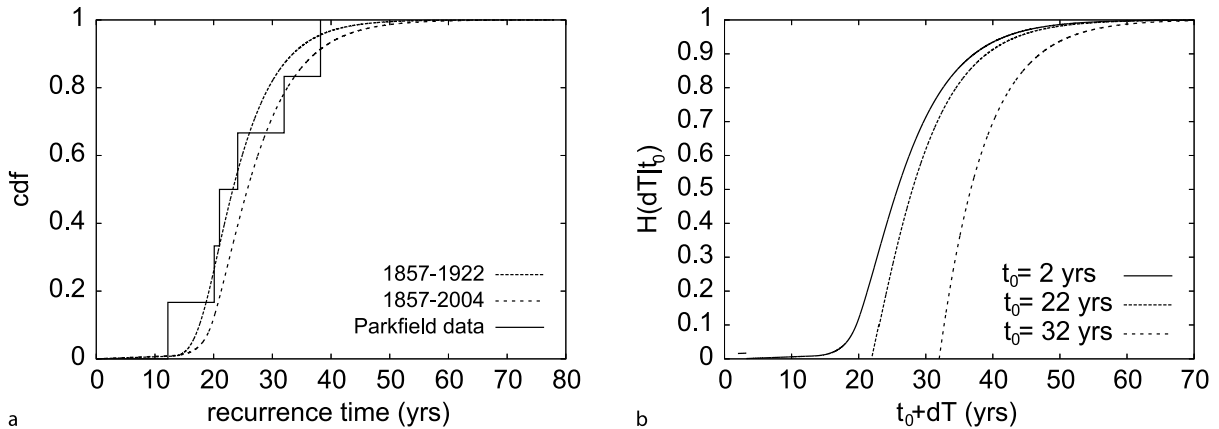
The cumulative probability density function (cdf) of recurrence times based on Eq. (26) and the observational data can now be calculated by

$$C(t) = \int_0^t \int_{\Delta\tau_{\min}}^{\Delta\tau_{\max}} f(t'|\Delta\tau)P(\Delta\tau|T_1, \dots, T_N)d\Delta\tau dt'. \quad (28)$$

The hazard function

$$H(\Delta t|t_0) = \frac{C(t_0 + \Delta t) - C(t_0)}{1 - C(t_0)} \quad (29)$$

is the conditional probability that the next large earthquake occurs in the interval  $[t_0; t_0 + \Delta t]$  given the time  $t_0$  since the last large event. Results for two choices of observational data (corresponding to two different observa-



Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 20

**a** Cumulative recurrence time distribution  $C(t)$  (Eq. (28)) for (1) the Bayesian approach with three data points: long-dashed line; (2) the Bayesian approach with six data points: short-dashed line. The solid line denotes the cdf of the six Parkfield recurrence times; **b** Hazard function  $H(t_0|\Delta T)$  (Eq. (29)) based on the six observational recurrence times between 1857 and 2004 a as a function of  $dT$  for different values of  $t_0$

tional periods) in comparison to the Parkfield cdf are given in Fig. 20a. The hazard function for three fixed values of  $t_0$  and varying  $\Delta t$  is given in Fig 20b.

This approach enables us to calculate the most likely occurrence time of the next (post 2004) Parkfield earthquake by picking the maximum of the (non-cumulative) recurrence time distribution (inner integral in Eq. (28)) after taking all Parkfield earthquakes (1857–2004) into account. Based on the analysis done so far, we may forecast the next  $\sim M6$  Parkfield earthquake to occur in May 2027. The error associated with one standard deviation of the pdf is 7.7 years. We note, however, that the probability for the occurrence of a  $\sim M6$  earthquake between May 2026 and May 2028 is only about 14%.

## Summary and Conclusions

The present review deals with the analysis, the understanding and the interpretation of seismicity patterns with a special focus on the critical point concept for large earthquakes. Both physical modeling and data analysis are discussed. This study aims at practical applications to model data from real fault zones. A point of particular interest is the detection of phenomena prior to large earthquakes and their relevance for a possible prediction of these events. Despite numerous reports on anomalous precursory seismicity changes [62], there is no precursor in sight which obeys a degree of universality that would make it practically useful. It is, therefore, important to study less frequent precursory phenomena by means of long model simulations.

Toward this goal, we discuss a numerical model which is on one hand reasonably physical, and on the other hand simple enough that it allows to obtain some analytical results and perform long simulations. The basic version of the model consists of a segmented two-dimensional strike-slip fault in a three-dimensional elastic half space and is inherently discrete because of the abrupt transition from static to kinetic friction [10]. This paper and ► [Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of](#) summarize a large body of analytical and numerical results associated with the model.

The results of the simulations indicate an overall good agreement of the synthetic seismicity with natural earthquake activity, with respect to frequency-size distributions and various features of earthquake sequences. The degree of spatial heterogeneity on the fault, which is implemented by means of space-dependent rheological properties, has important effects on the resulting catalogs. Smooth faults are associated with the characteristic earthquake statistics, regular occurrence of mainshocks and overall smooth stress fields. On the other hand, rough faults generate scale-free Gutenberg–Richter statistics, irregular mainshock occurrence, and overall rough stress fields. A closer look at the disorder of the stress field shows, however, that even on a smooth fault a gradual roughening takes place when the next large earthquake is approached [12,82]. This is reflected in the frequency-size distribution which evolves towards the Gutenberg–Richter law and other changes of seismicity. The results can be used to establish relations between the proximity of a state on a fault to a critical point, the (unobservable)

stress field, and the (observable) seismicity functions. Furthermore, it is demonstrated that the concept of “self-organized criticality” can be folded back to criticality associated with tuning parameters [12,31]. We note that phase diagrams with different dynamic regimes as functions of tuning parameters, in addition to criticality, provide a general and rich description of seismicity. Accelerating seismic release, growing spatial correlation length, changes of frequency-size statistics and evolution of other seismicity parameters may be used to track the approach to criticality [73,75,76,77].

### Future Directions

We have demonstrated that numerical fault models are valuable for understanding the underlying mechanisms of observed seismicity patterns, as well as for practical estimates of future seismic hazard. The latter requires model realizations that are tuned to a specific fault zone by assimilating available observational results and their uncertainties. In a case study, the seismic hazard in the Parkfield region has been estimated by combining such a tuned model with few observational data. The use of Bayesian analysis allows us to construct a flexible hazard model for this region which can, in general, incorporate statistical and non-statistical data (e. g. from paleoseismology and geodesy) to improve and update the estimations of the seismic hazard. This approach is particularly promising for less-well monitored regions, and especially for low-seismicity regions like those in central Europe.

Modification of the stress transfer calculations to account for a statistical preference of earthquake propagation direction on a given fault section, e. g. [6,25], can improve the estimates of seismic hazard associated with large faults. It is also possible to extend the discussed framework to other geohazards with even smaller amount of observational data, e. g. the occurrence of landslides. Since the fault model deals with coupled physical processes leading to interacting earthquakes, a challenging future direction will be the design of a more general model for interacting geohazards including earthquakes on different faults as well as landslides triggered by earthquakes, and perhaps tsunamis initiated by (submarine) earthquakes or landslides.

### Acknowledgments

We thank Jürgen Kurths, James R. Rice, Frank Scherbaum, Karin Dahmen, Donald L. Turcotte, and many others for useful discussions. GZ and MH acknowledge support from the collaborative research center “Complex Non-

linear Processes” (SFB 555) of the German Research Society (DFG). SH acknowledges support from the DFG-project SCHE280/14 and the EU-project SAFER. YBZ acknowledges support from the National Science Foundation, the United States Geological Survey, and the Southern California Earthquake Center. GZ and YBZ thank the Kavli Institute for Theoretical Physics, UC Santa Barbara, for hospitality during a 2005 program on Friction, Fracture and Earthquake Physics, and partial support based on NSF grant PHY99-0794. We thank James R. Holliday, Vladimir Keilis-Borok and Willie Lee for providing comments on the paper.

### Bibliography

1. Aki K, Richards PG (2002) Quantitative seismology. University Science Books, Sausalito
2. Bak P (1996) How Nature Works. The science of self-organised criticality. Springer, New York
3. Bak P, Tang C (1989) Earthquakes as a phenomenon of self-organised criticality. *J Geophys Res* 94:15635–156637
4. Båth M (1965) Lateral inhomogeneities in the upper mantle. *Tectonophysics* 2:483–514
5. Ben-Zion Y (1996) Stress, slip, and earthquakes in models of complex single-fault systems incorporating brittle and creep deformations. *J Geophys Res* 101:5677–5706
6. Ben-Zion Y (2001) Dynamic rupture in recent models of earthquake faults. *J Mech Phys Solids* 49:2209–2244
7. Ben-Zion Y (2003) Appendix 2, Key Formulas in Earthquake Seismology. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part B*. Academic Press, San Diego, pp 1857–1875
8. Ben-Zion Y, Lyakhovsky V (2002) Accelerated seismic release and related aspects of seismicity patterns on earthquake faults. *Pure Appl Geophys* 159:2385–2412
9. Ben-Zion Y, Lyakhovsky V (2006) Analysis of aftershocks in a lithospheric model with seismogenic zone governed by damage rheology. *J Geophys Int* 165:197–210; doi:10.1111/j.1365-246X.2006.02878.x
10. Ben-Zion Y, Rice JR (1993) Earthquake failure sequences along a cellular fault zone in a three-dimensional elastic solid containing asperity and nonasperity regions. *J Geophys Res* 98:14109–14131
11. Ben-Zion Y, Dahmen K, Lyakhovsky V, Ertas D, Agnon A (1999) Self-driven mode switching of earthquake activity on a fault system. *Earth Plan Sci Lett* 172:11–21
12. Ben-Zion Y, Eneva M, Liu Y (2003) Large Earthquake Cycles and Intermittent Criticality On Heterogeneous Faults Due To Evolving Stress and Seismicity. *J Geophys Res* 108:2307; doi:10.1029/2002JB002121
13. Bernardo JM, Smith AFM (1994) *Bayesian Theory*. Wiley, Chichester
14. Binney JJ, Dowrick NJ, Fisher AJ, Newman MEJ (1993) *The theory of critical phenomena*. Oxford University Press, Oxford
15. Bowman DD, Oullion G, Sammis CG, Sornette A, Sornette D (1998) An observational test of the critical earthquake concept. *J Geophys Res* 103:24359–24372

16. Brace WF (1960) An extension of the Griffith theory of fracture to rocks. *J Geophys Res* 65:3477–3480
17. Bufe CG, Varnes DJ (1993) Predictive modeling of the seismic cycle of the greater San Francisco Bay region. *J Geophys Res* 98:9871–9883
18. Burridge R, Knopoff L (1967) Model and theoretical seismicity. *Bull Seim Soc Am* 57:341–371
19. Byerlee JD (1978) Friction of rocks. *Pure Appl Geophys* 116:615–616
20. Chinnery M (1963) The stress changes that accompany strike-slip faulting. *Bull Seim Soc Am* 53:921–932
21. Corral Á (2004) Long-term clustering, scaling, and universality in the temporal occurrence of earthquakes. *Phys Rev Lett* 92:108501; doi:10.1103/PhysRevLett.92.108501
22. Dahmen K, Ertas D, Ben-Zion Y (1998) Gutenberg–Richter and characteristic earthquake behavior in simple mean-field models of heterogeneous faults. *Phys Rev E* 58:1494–1501
23. Daley DJ, Vere-Jones D (1988) *An Introduction to the Theory of Point Processes*, Springer Series: Probability and its Applications. Springer, Heidelberg
24. Dieterich JH (1994) A constitutive law for earthquake production and its application to earthquake clustering. *J Geophys Res* 99:2601–2618
25. Dor O, Rockwell TK, Ben-Zion Y (2006) Geologic observations of damage asymmetry in the structure of the San Jacinto, San Andreas and Punchbowl faults in southern California: A possible indicator for preferred rupture propagation direction. *Pure Appl Geophys* 163:301–349; doi:10.1007/s00024-005-0023-9
26. Ellsworth WL, Matthews MV, Nadeau RM, Nishenko SP, Reasenberg PA, Simpson RW (1999) A physically based earthquake recurrence model for estimation of long-term earthquake probabilities. *US Geol Surv Open-File Rept*, pp 99–522
27. Fisher DS, Dahmen K, Ramanathan S, Ben-Zion Y (1997) Statistics of earthquakes in simple models of heterogeneous faults. *Phys Rev Lett* 78:4885–4888
28. Geller RJ, Jackson DD, Kagan YY, Mulargia F (1997) Earthquakes cannot be predicted. *Science* 275:1616–1617
29. Gumbel EJ (1960) *Multivariate Extremal Distributions*. *Bull Inst Int Stat* 37:471–475
30. Gutenberg B, Richter CF (1956) Earthquake magnitude, intensity, energy and acceleration. *Bull Seismol Soc Am* 46:105–145
31. Hainzl S, Zöller G (2001) The role of disorder and stress concentration in nonconservative fault systems. *Phys A* 294:67–84
32. Hainzl S, Zöller G, Kurths J (1999) Similar power laws for fore- and aftershock sequences in a spring-block model for earthquakes. *J Geophys Res* 104:7243–7253
33. Hainzl S, Zöller G, Kurths J (2000) Self-organization of spatio-temporal earthquake clusters. *Nonlin Proc Geophys* 7:21–29
34. Hainzl S, Zöller G, Kurths J, Zschau J (2000) Seismic quiescence as an indicator for large earthquakes in a system of self-organized criticality. *Geophys Res Lett* 27:597–600
35. Hainzl S, Zöller G, Scherbaum F (2003) Earthquake clusters resulting from delayed rupture propagation in finite fault segments. *J Geophys Res* 108:2013; doi:10.1029/2001JB000610
36. Hillers G, Mai PM, Ben-Zion Y, Ampuero JP (2007) Statistical Properties of Seismicity Along Fault Zones at Different Evolutionary Stages. *J Geophys Int* 169:515–533; doi:10.1111/j.1365-246X.2006.03275.x
37. Huang J, Turcotte DL (1990) Are earthquakes an example of deterministic chaos? *Geophys Res Lett* 17:223–226
38. Jaumé SC, Sykes LR (1999) Evolving towards a critical point: A review of accelerating seismic moment/energy release prior to large and great earthquakes. *Pure Appl Geophys* 155:279–306
39. Jones LM, Molnar P (1979) Some characteristics of foreshocks and their possible relation to earthquake prediction and premonitory slip on faults. *J Geophys Res* 84:3596–3608
40. Kagan YY, Knopoff L (1978) Statistical study of the occurrence of shallow earthquakes. *J Geophys R Astron Soc* 55:67–86
41. Keilis-Borok VI, Soloviev AA (2003) *Nonlinear Dynamics of the Lithosphere and Earthquake Prediction*, Springer Series in Synergetics. Springer, Heidelberg
42. Lomnitz-Adler J (1999) Automaton models of seismic fracture: constraints imposed by the magnitude-frequency relation. *J Geophys Res* 98:17745–17756
43. Main IG, O'Brian G, Henderson JR (2000) Statistical physics of earthquakes: Comparison of distribution exponents for source area and potential energy and the dynamic emergence of log-periodic quanta. *J Geophys Res* 105:6105–6126
44. Matthews MV, Ellsworth WL, Reasenberg PA (2002) A Brownian model for recurrent earthquakes. *Bull Seim Soc Am* 92:2233–2250
45. Narteau C, Shebalin P, Hainzl S, Zöller G, Holschneider M (2003) Emergence of a band-limited power law in the aftershock decay rate of a slider-block model. *Geophys Res Lett* 30:1568; doi:10.1029/2003GL017110
46. Nur A, Booker JR (1972) Aftershocks caused by pore fluid flow? *Science* 175:885–887
47. Okada Y (1992) Internal deformation due to shear and tensile faults in a half space. *Bull Seim Soc Am* 82:1018–1040
48. Olami Z, Feder HS, Christensen K (1992) Self-organized criticality in a continuous, nonconservative cellular automaton modeling earthquakes. *Phys Rev Lett* 68:1244–1247
49. Omori F (1894) On the aftershocks of earthquakes. *J Coll Sci Imp Univ Tokyo* 7:111–200
50. Patel JK, Kapadia CH, Owen DB (1976) *Handbook of statistical distributions*. Marcel Dekker, New York
51. Reasenberg P (1985) Second-order moment of central California seismicity. *J Geophys Res* 90:5479–5495
52. Reid HF (1910) *The Mechanics of the Earthquake*, The California Earthquake of April 18, 1906. Report of the State Investigation Commission, vol 2. Carnegie Institution of Washington, Washington
53. Rundle JB, Klein W, Turcotte DL, Malamud BD (2000) Precursory seismic activation and critical point phenomena. *Pure Appl Geophys* 157:2165–2182
54. Saleur H, Sammis CG, Sornette D (1996) Discrete scale invariance, complex fractal dimensions, and log-periodic fluctuations in seismicity. *J Geophys Res* 101:17661–17677
55. Savage JC, Svarc JL, Prescott WH (1999) Geodetic estimates of fault slip rates in the San Francisco Bay area. *J Geophys Res* 104:4995–5002
56. Scholz CH (1998) Earthquakes and friction laws. *Nature* 391:37–42
57. Shcherbakov R, Turcotte DL (2004) A damage mechanics model for aftershocks. *Pure Appl Geophys* 161:2379; doi:10.1007/s00024-004-2570-x
58. Shin TC, Teng TL (2001) An overview of the 1999, Chichi, Taiwan, earthquake. *Bull Seismol Soc Am* 91:895–913

59. Sornette D (2004) *Self-organization and Disorder: Concepts & Tools*, Springer Series in Synergetics. Springer, Heidelberg
60. Sornette D, Sammis CG (1995) Complex critical exponents from renormalization group theory of earthquakes: Implication for earthquake predictions. *J Phys* 1(5):607–619
61. Sornette D, Sornette A (1999) Renormalization of earthquake aftershocks. *Geophys Res Lett* 6:1981–1984
62. Field EH et al. (2007) Special Issue: Regional Earthquake Likelihood Models. *Seismol Res Lett* 78:1
63. Steacy SJ, McCloskey J, Bean CJ, Ren JW (1996) Heterogeneity in a self-organized critical earthquake model. *Geophys Res Lett* 23:383–386
64. Turcotte DL (1997) *Fractals and chaos in geology and geophysics*. Cambridge University Press, New York
65. Turcotte DL, Newman WI, Shcherbakov R (2003) Micro and macroscopic models of rock fracture. *J Geophys Int* 152:718–728
66. Utsu T (2002) Statistical features of seismicity. In: *Int Assoc Seismol & Phys Earth's Interior* (ed) International handbook of earthquake and engineering seismology, vol 81A. Academic Press, San Diego, pp 719–732
67. Utsu T, Ogata Y, Matsu'ura RS (1995) The centenary of the Omori formula for a decay law of aftershock activity. *J Phys Earth* 43:1–33
68. Wesnousky SG (1994) The Gutenberg–Richter or characteristic earthquake distribution, which is it? *Bull Seismol Soc Am* 90:525–530; 84:1940–1959
69. Wiemer S, Baer M (2000) Mapping and removing quarry blast events from seismic catalogs: Examples from Alaska, the Western United States, and Japan. *Bull Seismol Soc Am* 90:525–530
70. Working Group on California Earthquake Probabilities (2003) *Earthquake probabilities in the San Francisco Bay region*. US Geol Survey Open File Report 03–214, US Geological Survey
71. Wyss M (1997) Cannot earthquakes be predicted? *Science* 278:487
72. Wyss M, Habermann RE (1988) Precursory seismic quiescence. *Pure Appl Geophys* 126:319–332
73. Zaliapin I, Liu Z, Zöller G, Keilis-Borok V, Turcotte DL (2002) On increase of earthquake correlation length prior to large earthquakes in California. *Comp Seismol* 33:141–161
74. Zöller G, Ben-Zion Y, Holschneider M, Hainzl S (2007) Estimating recurrence times and seismic hazard of large earthquakes on an individual fault. *J Geophys Int* 170:1300–1310; doi:10.1111/j.1365-246X.200703480.x
75. Zöller G, Hainzl S (2001) Detecting premonitory seismicity patterns based on critical point dynamics. *Nat Hazards Earth Syst Sci* 1:93–98
76. Zöller G, Hainzl S (2002) A systematic spatiotemporal test of the critical point hypothesis for large earthquakes. *Geophys Res Lett* 29:1558; doi:10.1029/2002GL014856
77. Zöller G, Hainzl S, Kurths J (2001) Observation of growing correlation length as an indicator for critical point behavior prior to large earthquakes. *J Geophys Res* 106:2167–2175
78. Zöller G, Hainzl S, Kurths J, Zschau J (2002) A systematic test on precursory seismic quiescence in Armenia. *Nat Hazards* 26:245–263
79. Zöller G, Holschneider M, Ben-Zion Y (2004) Quasi-static and quasi-dynamic modeling of earthquake failure at intermediate scales. *Pure Appl Geophys* 161:2103–2118; doi:10.1007/s00024-004-2551-0
80. Zöller G, Holschneider M, Ben-Zion Y (2005) The role of heterogeneities as a tuning parameter of earthquake dynamics. *Pure Appl Geophys* 162:1027; doi:10.1007/s00024-004-2660-9
81. Zöller G, Hainzl S, Holschneider M, Ben-Zion Y (2005) Aftershocks resulting from creeping sections in a heterogeneous fault. *Geophys Res Lett* 32:L03308; doi:10.1029/2004GL021871
82. Zöller G, Hainzl S, Ben-Zion Y, Holschneider M (2006) Earthquake activity related to seismic cycles in a model for a heterogeneous strike-slip fault. *Tectonophysics* 423:137–145; doi:10.1016/j.tecto.2006.03.007

# Seismicity, Statistical Physics Approaches to

DIDIER SORNETTE<sup>1</sup>, MAXIMILIAN J. WERNER<sup>2</sup>

<sup>1</sup> Department of Management, Technology and Economics, ETH Zurich, Switzerland

<sup>2</sup> Swiss Seismological Service, Institute of Geophysics, ETH Zurich, Switzerland

## Article Outline

Glossary

Definition of the Subject

Introduction

Concepts and Computational Tools

Competing Mechanisms and Models

Empirical Studies of Seismicity Inspired by Statistical Physics

Future Directions

Bibliography

## Glossary

**Chaos** Chaos occurs in dynamical systems with two ingredients: (i) nonlinear recurrent re-injection of the dynamics into a finite domain in phase space and (ii) exponential sensitivity of the trajectories in phase space to initial conditions.

**Continuous phase transitions** If there is a finite discontinuity in the first derivative of the thermodynamic potential, then the phase transition is termed first-order. During such a transition, a system either absorbs or releases a fixed amount of latent heat (e.g. the freezing/melting of water/ice). If the first derivative is continuous but higher derivatives are discontinuous or infinite, then the phase transition is called continuous, of the second kind, or critical. Examples include the critical point of the liquid–gas transition, the Curie point of the ferromagnetic transition, or the superfluid transition [127,235].

**Critical exponents** Near the critical point, various thermodynamic quantities diverge as power laws with associated critical exponents. In equilibrium systems, there are scaling relations that connect some of the critical exponents of different thermodynamic quantities [32,127,203,216,235].

**Critical phenomena** Phenomena observed in systems that undergo a continuous phase transition. They are characterized by scale invariance: the statistical properties of a system at one scale are related to those at another scale only through the ratio of the two scales

and not through any one of the two scales individually. The scale invariance is a result of fluctuations and correlations at all scales, which prevents the system from being separable in the large scale limit at the critical point [32,203,235].

**Declustering** In studies of seismicity, declustering traditionally refers to the deterministic identification of fore-, main- and aftershocks in sequences (or clusters) of earthquakes clustered in time and space. Recent, more sophisticated techniques, e.g. stochastic declustering, assign to earthquakes probabilities of being triggered or spontaneous.

## Dynamical scaling and exponents

Non-equilibrium critical phase transitions are also characterized by scale invariance, scaling functions and critical exponents. Furthermore, some evidence supports the claim that universality classes also exist for non-equilibrium phase transitions (e.g. the directed percolation and the Manna universality class in sandpile models), although a complete classification of classes is lacking and may in fact not exist at all. Much interest has recently focused on directed percolation, which, as the most common universality class of absorbing state phase transitions, is expected to occur in many physical, chemical and biological systems [85,135,203].

**Finite size scaling** If a thermodynamic or other quantity is investigated at the critical point under a change of the system size, the scaling behavior of the quantity with respect to the system size is known as finite size scaling [32]. The quantity may refer to a thermodynamic quantity such as the free energy or it may refer to an entire probability distribution function. At criticality, the sole length scale in a finite system is the upper cut-off  $s_c$ , which diverges in the thermodynamic limit  $L \rightarrow \infty$ . Assuming a lower cut-off  $s_0 \ll s_c$ , a finite size scaling ansatz for the distribution  $P(s; s_c)$  of the observable variable  $s$ , which depends on the upper cut-off  $s_c$  is then given by:

$$P(s; s_c) = a s^{-\tau} G(s/s_c) \quad \text{for } s, s_c \gg s_0, \quad (1)$$

where the parameter  $a$  is a non-universal metric factor,  $\tau$  is a universal (critical) exponent, and  $G$  is a universal scaling function that decays sufficiently fast for  $s \gg s_c$  [32,36]. Pruessner [163] provides a simple yet instructive and concise introduction to scaling theory and how to find associated exponents. System-specific corrections appear to sub-leading order.

**Fractal** A deterministic or stochastic mathematical object that is defined by its exact or statistical self-similar-



ity at all scales. Informally, it often refers to a rough or fragmented geometrical shape which can be subdivided into parts which look approximately the same as the original shape. A fractal is too irregular to be described by Euclidean geometry and has a fractal dimension that is larger than its topological dimension but less than the dimension of the space it occupies.

**Mean-Field** An effective or average interaction field designed to approximately replace the interactions from many bodies by one effective interaction which is constant in time and space, neglecting fluctuations.

**Mechanisms for power laws** Power laws may be the hallmark of critical phenomena, but there are a host of other mechanisms that can lead to power laws (see Chapter 14 of [203] for a list of power law mechanisms as well as [37,143]). Observations of scale invariant statistics therefore do not necessarily imply SOC, of course. Power laws express the existence of a symmetry (scale invariance) and there are many mechanisms by which a symmetry can be obtained or restored.

**Non-equilibrium phase transitions** In contrast to systems at equilibrium, non-equilibrium phase transitions involve dynamics, energy input and dissipation. Detailed balance is violated and no known equivalent of the partition function exists, from which all thermodynamic quantities of interest derive in equilibrium. Examples of non-equilibrium phase transitions include absorbing state phase transitions, reaction-diffusion models, and morphological transitions of growing surfaces [85,135].

**Phase transitions** In (equilibrium) statistical mechanics, a phase transition occurs when there is a singularity in the free energy or one of its derivatives. Examples include the freezing of water, the transition from ferromagnetic to paramagnetic behavior in magnets, and the transition from a normal conductor to a superconductor [127,235].

**Renormalization group theory** A mathematical theory built on the idea that the critical point can be mapped onto a fixed point of a suitably chosen transformation on the system's Hamiltonian. It provides a foundation for understanding scaling and universality and provides tools for calculating exponents and scaling functions. Renormalization group theory provides the basis for our understanding of critical phenomena [32,216,235]. It has been extended to non-Hamiltonian systems and provides a general framework for constructing theories of the macro-world from the microscopic description.

**Self-organized criticality (SOC)** Despite two decades of research since its inception by [13] and the ambitious

claim by [11] that, as a mechanism for the ubiquitous power laws in Nature, SOC was “How Nature Works”, a commonly accepted definition along with necessary and sufficient conditions for SOC is still lacking [93,163,203]. A less rigorous definition may be the following: Self-organized criticality refers to a non-equilibrium, critical and marginally stable steady-state, which is attained spontaneously and without (explicit) tuning of parameters. It is characterized by power law event distributions and fractal geometry (in some cases) and may be expected in slowly driven, interaction-dominated threshold systems [93]. Some authors additionally require that temporal and/or spatial correlations decay algebraically (e. g. [84], but see [163]). Definitions in the literature range from broad (simply the absence of characteristic length scales in non-equilibrium systems) to narrow (the criticality is due to an underlying continuous phase transition with all of its expected properties) (see, e. g., [162] for evidence that precipitation is an instance of the latter definition of SOC in which a non-linear feedback of the order parameter on the control parameter turns a critical phase transition into a self-organized one attracting the dynamics [198]).

**Spinodal decomposition** In contrast to the slow process of phase separation via nucleation and slow growth of a new phase in a material inside the metastable region near a first-order phase transition, spinodal decomposition is a non-equilibrium, rapid and critical-like dynamical process of phase separation that occurs quickly and throughout the material. It needs to be induced by rapidly quenching the material to reach a sub-area (sometimes a line) of the unstable region of the phase diagram which is characterized by a negative derivative of the free energy.

**Statistical physics** is the set of concepts and mathematical techniques allowing one to derive the large-scale laws of a physical system from the specification of the relevant microscopic elements and of their interactions.

**Turbulence** In fluid mechanics, turbulence refers to a regime in which the dynamics of the flow involves many interacting degrees of freedom, and is very complex with intermittent velocity bursts leading to anomalous scaling laws describing the energy transfer from injection at large scales to dissipation at small scales.

**Universality** In systems with little or no frozen disorder, equilibrium continuous phase transitions fall into a small set of universality classes that are characterized by the same critical exponents and by certain scal-

ing functions that become identical near the critical point. The class depends only on the dimension of the space and the dimension of the order parameter. For instance, the critical point of the liquid–gas transition falls into the same universality class as the 3D Ising model. Even some phase transitions occurring in high-energy physics are expected to belong to the Ising class. Universality justifies the development and study of extremely simplified models (caricatures) of Nature, since the behavior of the system at the critical point can nevertheless be captured (in some cases exactly). However, non-universal features remain even at the critical point but are less important, e. g. amplitudes of fluctuations or system-specific corrections to scaling that appear at sub-leading order [32,216,235,239].

### Definition of the Subject

A fundamental challenge in many scientific disciplines concerns upscaling, that is, of determining the regularities and laws of evolution at some large scale from those known at a lower scale: biology (from molecules to cells, from cells to organs); neurobiology (from neurons to brain function), psychology (from brain to emotions, from evolution to understanding), ecology (from species to the global web of ecological interactions), condensed matter physics (from atoms and molecules to organized phases such as solid, liquid, gas, and intermediate structures), social sciences (from individual humans to social groups and to society), economics (from producers and consumers to the whole economy), finance (from investors to the global financial markets), Internet (from e-pages to the world wide web 2.0), semantics (from letters and words to sentences and meaning), and so on. Earthquake physics is no exception, with the challenge of understanding the transition from the laboratory scale (or even the microscopic and atomic scale) to the scale of fault networks and large earthquakes.

Statistical physics has had a remarkably successful track record in addressing the upscaling problem in physics. While the macroscopic laws of thermodynamics have been established in the 19th century, their microscopic underpinning were elaborated in the early 20th century by Boltzmann and followers, building the magnificent edifice of statistical physics. Statistical physics can be defined as the set of concepts and mathematical techniques allowing one to derive the large-scale laws of a physical system from the specification of the relevant microscopic elements and of their interactions. Dealing with huge ensembles of elements (atoms, molecules) of the order of the Avogadro number ( $\simeq 6 \cdot 10^{23}$ ), statistical physics uses the

mathematical tools of probability theory combined with other relevant fields of physics to calculate the macroscopic properties of large populations.

One of the greatest achievement of statistical physics was the development of the renormalization group analysis, to construct a theory of interacting fields and of critical phase transitions. The renormalization group is a perfect example of how statistical physics addresses the micro-macro upscaling problem. It decomposes a problem of finding the macroscopic behavior of a large number of interacting parts into a succession of simpler problems with a decreasing number of interacting parts, whose effective properties vary with the scale of observation. The renormalization group thus follows the proverb “divide to conquer” by organizing the description of a system scale-by-scale. It is particularly adapted to critical phenomena and to systems close to being scale-invariant. The renormalization group translates into mathematical language the concept that the overall behavior of a system is the aggregation of an ensemble of arbitrarily defined sub-systems, with each sub-system defined by the aggregation of sub-subsystems, and so on [203].

It is important to stress that up to now the term “statistical” has different meanings in statistical physics and in statistical seismology, a field that has developed as a marriage between probability theory, statistics and the part of seismology concerned with empirical patterns of earthquake occurrences [225] (but not with physics). Statistical seismology uses stochastic models of seismicity, which are already effective large-scale representation of the dynamical organization. In contrast, a statistical physics approach to earthquake strives to derive these statistical models or other descriptions from the knowledge of the microscopic laws of friction, damage, rupture, rock-water interactions, mechano-chemistry and so on, at the microscopic scales [200,201]. In other words, what is often missing in statistical seismology is the physics to underpin the stochastic model on physically-based laws, e. g. rate-and-state friction [55].

The previously mentioned successes of statistical physics promote the hope that a similar program can be developed for other fields, including seismology. The successes have been more limited, due to the much more complex interplay between mechanisms, interactions and scales found in these out-of-equilibrium systems. This short essay provides a subjective entry to understand some of the different attempts, underlining the few successes, the problems and open questions. Rather than providing an exhaustive review, we mention what we believe to be important topics and have especially included recent work.

## Introduction

Much of the recent interest of the statistical physics community has focused on applying scaling techniques, which are common tools in the study of critical phenomena, to the statistics of inter-event recurrence times or waiting times [14,40,41,42,43,44,46,48,134]. However, the debate over the relevance of critical phenomena to earthquakes stretches back as far as 30 years [7,11,12,47,61,84,93,104,106,109,114,147,151,178,192,193,194,202,203,207,223], ► **Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of** and ► **Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space**. The current debate on recurrence statistics is thus the latest tack in an evolving string of arguments with a long history. As discussed below, many of the claims made in the recent articles on recurrence statistics have either been challenged, refuted or explained by previously known facts about earthquake statistics [132,133,145,176,177,230]. As will be discussed below, this debate in the literature is important because of the potential consequences for understanding earthquakes, but it needs to be pursued with rigorous scientific arguments accessible to both the seismological and the statistical physics communities.

The debate would almost certainly benefit significantly from testing hypotheses with simulations to establish null hypotheses and benchmarks: seismicity patterns are sufficiently stochastic and earthquake catalogs contain a sufficient amount of observational uncertainties so as to make inference difficult. It is often not straightforward to predict the signal of well-known statistical features such as clustering in new data analysis techniques. Therefore, testing the purported claims by realistic simulations of earthquake catalogs can provide a strong benchmark against which the claims can be evaluated. This view and the corresponding criticism of many studies has been put forward and defended for a long time by Kagan [111].

Such a model-dependent approach may be at odds with the philosophy of a so-called “model-free” analysis, which the community of statistical physicists claim to take in their analysis. For instance, network theory-based approaches, space-time window-based finite size scaling, box-covering methods and other techniques used in the study of critical and fractal phenomena are said to be “model-free” because no assumptions about seismicity are supposedly made at the outset. By using model-free analysis techniques, the often uncertain and sometimes clearly wrong assumptions of flawed models and resulting biased results are meant to be circumvented.

However, as is almost always the case in statistical hypothesis testing, the less assumptions are made about the test, the less powerful the test statistic. More importantly, seismicity is sufficiently stochastic so that well-known features may appear as novel in new analysis methods. Furthermore, to convince the seismological community of new data analysis techniques, the methods need to be tested on established knowledge and show the improvement over traditional methods. These types of initial tests are rarely performed by the statistical physics community.

In the next Sect. “**Concepts and Computational Tools**”, we present a summary of some of the concepts and computational tools that have been developed in attempts to apply statistical physics approaches to seismology. Then, Sect. “**Competing Mechanisms and Models**” summarizes the leading theoretical physical models of the space-time organization of earthquakes. Section “**Empirical Studies of Seismicity Inspired by Statistical Physics**” presents a general discussion and several examples of the new metrics proposed by statistical physicists, underlining their strengths and weaknesses. Section “**Future Directions**” briefly outlines expected developments.

## Concepts and Computational Tools

### Renormalization, Scaling and the Role of Small Earthquakes in Models of Triggered Seismicity

A common theme in many of the empirical relations in seismology (and in those employed in seismicity models) is the lack of a dominating scale. Many natural phenomena can be approached by the traditional reductionist approach to isolate a process at a particular scale. For example, the waves of an ocean can be described quite accurately by a theory that entirely ignores the fact that the liquid is made out of individual molecules. Indeed, the success of most practical theories in physics depends on isolating a scale [234], although since this recognition, much progress has been made in developing a holistic approach for processes that do not fall into this class. Given current observational evidence, earthquakes seem to belong to the set of processes characterized by a lack of one dominating length scale: fluctuations of many or perhaps a wide continuum of sizes seem to be important and are in no way diminished – even when one is interested solely in large-scale descriptions [206].

The traditional reductionist approach in seismology, which, for instance, attempted to separate large (main) shocks from small (fore- or after-) shocks, is slowly giving way to the holistic approach, in which all earthquakes are created equal and seismicity is characterized by fluctuations of all sizes. This gradual shift is supported, on

a conceptual and qualitative level, by the vision of critical phenomena. A particularly strong model of the interactions between earthquakes has emerged in the concept of triggering, which places all earthquakes on the same footing: each earthquake can trigger its own events, which in turn can trigger their own events, and so on, according to the same probability distributions, and the resulting seismicity can be viewed as the superposed cascades of triggered earthquakes that cluster in space and time [77,117,149,150].

From this point of view, it is natural that small earthquakes are important to the overall spatio-temporal patterns of seismicity. Indeed, the scaling of aftershock productivity with mainshock magnitude suggests that small earthquakes are cumulatively as important for the triggered seismicity budget as rarer but larger events [60,76,82]. The importance of small earthquakes has also been documented in, e.g., [73,138,141].

But earthquake catalogs do not contain information (by definition) about the smallest, unobserved events, which we know to exist from acoustic emission experiments and earthquakes recorded in mines. To guarantee a finite seismicity budget, Sornette and Werner argued [209] for the existence of a smallest triggering earthquake, akin to a “ultra-violet cut-off” in quantum field theory, below which earthquakes do not trigger other events. Introducing a formalism which distinguishes between the detection threshold and the smallest triggering earthquake, Sornette and Werner placed constraints on its size by using a simplified version of the popular Epidemic-Type Aftershock Sequence (ETAS) Model [149], a powerful model of triggered seismicity based on empirical statistics, and by using observed aftershock sequences. Sornette and Werner [210] also considered the branching structure of one complete cascade of triggered events, deriving an apparent branching ratio and the apparent number of untriggered events, which are observed when only the structure above the detection threshold is known. As a result of our inability to observe the entire branching structure, inferred clustering parameters are significantly biased and difficult to interpret in geophysical terms. Second, separating triggered from untriggered events, commonly known as declustering, also strongly depends on the threshold, so that it cannot even in theory constitute a physically sound method.

Sornette and Werner [210] also found that a simplified, averaged version of the ETAS model can be renormalized onto itself, with effective clustering parameters, under a change of the threshold. Saichev and Sornette [175] confirmed these results for the stochastic number statistics of the model using a rigorous approach in terms of generat-

ing probability functions, but also showed that the temporal statistics could not be renormalized. Furthermore, it can be shown (see Chapter 4 of [229]) that the conditional intensity function of the ETAS model, the mathematical object which uniquely defines the model, cannot be renormalized onto itself under a change of magnitude threshold. It is not a fixed-point of the renormalization process operating via magnitude coarse-graining. The functional form of the model must change under a change in the detection threshold [175]. In other words, if earthquakes occur according to an ETAS model above some cut-off  $m_0$ , then earthquakes above  $m_d$  cannot be described by the ETAS model in a mathematically exact way. Although in practice, the ETAS model provides an excellent fit.

The issue of how to deal with small earthquakes is thus reminiscent of the decades of efforts that have been invested in physics to deal with the famous ultra-violet cut-off problem, eventually solved by the so-called “renormalization” theory of Feynmann, Schwinger and Tomonaga. In the 1960s and 1970s, this method of renormalization was extended into the “renormalization group” (in fact a semi-group in the strict mathematical sense) for the theory of critical phenomena (see glossary), which we also mention in Sect. “Competing Mechanisms and Models”. It is fair to say that there has been limited success in developing a multi-scale description of the physics of earthquakes and, in particular, in addressing the upscaling problem and the impact of the many small earthquakes.

One tantalizing approach, not yet really understood in terms of all its consequences and predictions, is the variant of the ETAS model proposed by Vere-Jones [224], which has the remarkable property of being bi-scale invariant under a transformation involving time and magnitudes. One of the modifications brought in by [224] is to assume that the distribution of the daughter magnitudes is dependent on the mother magnitude  $m_i$  through a modification of the Gutenberg–Richter distribution of triggered earthquake magnitudes by a term of the form  $\exp(-\delta|m - m_i|)$ , where  $\delta > 0$  quantifies the distance to the standard Gutenberg–Richter distribution. Remarkably, Saichev and Sornette [174], who studied the Vere-Jones model, found that, due to the superposition of the many magnitude distributions of each earthquake in the cascades of triggered events, the resulting distribution of magnitudes over a stationary catalog is a pure Gutenberg–Richter law. Thus, there might be hidden characteristic scales in the physics of triggering that are not revealed by the standard observable one-point statistical distributions. Simulation and parameter estimation algorithms for the Vere-Jones model are not yet available. If and when these algorithms become available, the study of this bi-scale in-

variant branching model may be a strong alternative to the ETAS model, as this model is exactly scale invariant with neither ultra-violet nor infra-red cut-offs.

Nevertheless, being empirically based, these stochastic point process models lack a genuine microscopic physical foundation. The underlying physics is not explicitly addressed and only captured effectively by empirical statistics, even at the smallest scales. The physical processes and their renormalization are missing in this approach.

### Universality

Universality, as defined in the glossary, justifies the development and study of extremely simplified models (caricatures) of Nature, since the behavior of a studied system at the critical point can nevertheless be captured by toy models (in some cases exactly). For instance, the liquid–gas transition, the ferromagnetic to paramagnetic transition, and the behavior of binary alloys, all apparently different systems, can be described successfully by an extremely simplified picture of Nature (the Ising model) because of universality. The hope that a similar principle holds for earthquakes (and other non-equilibrium systems) underpins many of the models and tools inspired by statistical physics that have been applied to seismicity, to bring about the “much coveted revolution beyond reductionism” [17,67].

Speaking loosely, the appearance of power laws in many toy models is often interpreted as a kind of universal behavior. The strict interpretation of universality classes, however, requires that critical exponents along with scaling functions are identical for different systems. It is interesting to note that slight changes in the sand-pile model already induce new universality classes, so that even within a group of toy models, the promise of universality is not, strictly, fulfilled [13,98,136].

A lively debate in seismology concerns the universality of, on one hand, the frequency-size distribution of earthquake magnitudes (e.g. [47,61,91,108,111,232]), and, on the other hand, the universality of the exponent of the Gutenberg–Richter distribution (e.g. [25,184,212,233]). A spatio-temporally varying critical exponent is not traditionally part of the standard critical phenomena repertoire, although analytical and numerical results based on a simple earthquake model on a fault showed a possible spontaneous switching between Gutenberg–Richter and characteristic earthquake behavior associated with a non-equilibrium phase transition [24,47,61], ► [Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of](#) and ► [Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space](#). Another

possible mechanism for the coexistence of and intermittent shifts between different regimes (Gutenberg–Richter scaling and characteristic earthquakes) stems from the competition between several interacting faults which give rise to long intervals of activity in some regions followed by similarly long intervals of quiescence [129,213,214]. Both careful empirical investigations of seismicity parameters and theoretical progress on heterogeneous, spatially extended critical phenomena may help elucidate the controversy.

### Intermittent Periodicity and Chaos

As part of the conquest of chaos theory in the 70s and 80s [58,139], its concepts and methods were invariably also applied to seismicity. Huang and Turcotte [87,88] modeled the interaction between two faults by two sliding blocks that are driven by a plate through springs, coupled to one another via another spring and endowed with a velocity-weakening friction law. The dynamical evolution of the blocks showed chaotic behavior and period-doubling (the Feigenbaum route to chaos [58]). Huang and Turcotte [87,88] suggested that the interaction of the Parkfield segment with the southern San Andreas fault may be governed by the kind of chaotic behavior they observed in their model: the lighter block slipped quasi-periodically for several times until both slipped together. Their study explained that apparent quasi-periodicity of earthquakes on fault segments may be a result of chaotic interactions between many fault segments, thereby providing a warning that the extrapolations of quasi-periodic models are not to be trusted (e.g. [15,75] and references therein). However, it is doubtful that models with just a few degrees of freedom can go a long way towards providing deeper physical insights, or predictive tools. One needs to turn to models with a large number of degrees of freedom, for which turbulence appears as the leading paradigm of complexity.

### Turbulence

A drastically different approach has been favored by Yan Kagan [104,106,109,114], who described seismicity as the “turbulence of solids” – attesting to the far greater problems in earthquake seismology than the theory of critical phenomena promises to solve. While renormalization group methods and scaling theory have contributed immensely to the study of turbulence [63], the problem of turbulence involves significant additional complications. First, loosely speaking, renormalization group theory helps predict global behavior by coarse-graining over degrees of freedom, which is essentially a bottom–up ap-

proach. In turbulence, the enstrophy acts bottom-up, but the energy cascades top down. Secondly, there is a significant spatial and topological aspect to turbulence, for instance involving topological defects, such as filament structures, which are crucial to the dynamical evolution of the system. The existence of the two cascades (top-down and bottom-up) as well as the influence of the dissipation scale all the way within the so-called inertial range makes turbulence the most important problem still unsolved in classical Physics. The importance of addressing the issue of the interplay between the top-down and bottom-up cascades in earthquake toy models has been outlined by [65, 66, 236, 237].

The analogous problem for seismicity lies in the complex fault network, which constrains seismicity through its weak structures but also grows and evolves because of earthquakes redistributing stresses and rupturing fresh surfaces. The statistical description of this tensorial and dynamical problem is only at its beginning [64, 99, 100, 101, 102, 103, 107, 116, 118, 119, 142, 213]. But it is likely to be a key aspect to the dynamical evolution of faults and seismicity. New physics and approaches are required to tackle the tensorial nature of the stress and strain fields and the complex topological structures of defects, from dislocations to joints and faults, as well as the many different physical processes operating from the atomic scale to the macro-scale [200, 201].

### Self-Organized Criticality

Self-organized criticality (SOC) refers to the spontaneous organization of a system driven from outside into a globally dynamical statistical stationary state, which is characterized by self-similar distributions of event sizes and sometimes fractal geometrical properties. SOC applies to the class of phenomena occurring in driven out-of-equilibrium systems made of many interactive components, which possess the following fundamental properties: 1) a highly non-linear behavior, 2) a very slow driving rate, 3) a globally stationary regime, characterized by stationary statistical properties, and 4) power-law distributions of event sizes and fractal geometrical properties. The crust obeys these four conditions, as first suggested by [12, 193], who proposed to understand the spatio-temporal complexity of earthquakes in this perspective.

The appeal of placing the study of earthquakes in the framework of critical phenomena may be summarized as follows. Power law distributions can be understood as a result of an underlying continuous phase transition into which the crust has organized itself [197, 211]. Applying

the methods of renormalization group theory may help calculate exponents and scaling functions and rationalize the spatio-temporal organization of seismicity along with its highly correlated structures. For instance, Sornette and Virieux [208] provided a theoretical framework which links the scaling laws exhibited by earthquakes at short times and plate tectonic deformations at large times. Perhaps earthquakes fall into a universality class which can be solved exactly and/or investigated in toy models. Moreover, studying the detailed and highly complicated microphysics involved in earthquakes may not lead to insights about the spatio-temporal organization, because, as a critical phenomenon, the traditional approach of separating length scales to describe systems is inadequate. On the other hand, as mentioned above, there is the possibility of a hierarchy of physical processes and scales which are inter-related [156, 157], for which the simplifying approach in terms of critical phenomena is likely to be insufficient.

As another reason for the importance of the topic, interesting consequences for the predictability of earthquakes might be derived, for instance by mapping earthquakes to a genuine critical point (the accelerating moment release hypothesis, e.g. [202, 207]) or by mapping earthquakes to SOC (e.g. [70, 147]). The latter mapping had led some to argue that earthquakes are inherently unpredictable. In the sandpile paradigm [13], there is little difference between small and large avalanches, and this led similarly to the concept that “large earthquakes are small earthquakes that did not stop,” hence their supposed lack of predictability. More than ten years after this contentious proposal, a majority of researchers, including most of the authors of this “impossibility claim,” recognize that there is some degree of predictability [83, 97]. Actually, the clarifications came from investigators of SOC, who recognized that the long-term organization of sandpiles [51, 52] and of toy models of earthquakes and fault networks [142, 213, 214] is characterized by long-range spatial and temporal correlations. Thus, large events may indeed be preceded by subtle long-range organizational structures, an idea at the basis of the accelerating moment release hypothesis. This idea is also underlying the pattern recognition method introduced by Gelfand et al. [69] and developed extensively by V. Keilis-Borok and his collaborators for earthquake predictions [122]. In addition, Huang et al. [89] showed that avalanche dynamics occurring within hierarchical geometric structures are characterized by significant precursory activity before large events; this provides a clear proof of the possible coexistence between critical-like precursors of large events and a long-term self-organized critical dynamical state.

In summary, self-organized criticality provides a general conceptual framework to articulate the search for a physical understanding of the large-scale and long-time statistical properties of the seismogenic process and of the predictability of earthquakes. Beyond this, it is of little help as many different mechanisms have been documented at the origin of SOC (see, e.g., chapter 15 in [203]). SOC is not a theory, it does not provide any specific calculation tools; it is a concept offering a broad classification of the kinds of dynamics that certain systems, including the Earth crust, seem to spontaneously select.

### Competing Mechanisms and Models

It should be noted at this point that the statistical physics approach to earthquake science is not limited to SOC. Over the years, several groups have proposed to apply the concepts and tools of statistical physics to understanding the multiscale dynamics of earthquake and fault systems. Various mechanisms drawn conceptually from statistical mechanics but not necessary even limited to critical (phase transition) phenomena have been proposed and are being pursued. Such approaches include the concept of the critical point earthquake related to accelerated moment release, network theory, percolation and fiber models as models for fracture, and many more, some of which can be found in [84,203,220,221].

In this section, we outline some of the major model classes which underpin distinct views on what are the dominating mechanisms to understand earthquakes and their space-time organization.

### Roots of Complexity in Seismicity: Dynamics or Heterogeneity?

The 1990s were characterized by vigorous discussions at the frontier between seismology and statistical physics aimed at understanding the origin of the observed complexity of the spatio-temporal patterns of earthquakes. The debate was centered on the question of whether space-time complexity can occur on a single homogeneous fault, solely as a result of the nonlinear dynamics [23,38,39,128,186,187,188,189], associated with the slip and velocity dependent friction law validated empirically in particular by [53,54,55,56]. Or, is the presence of quenched heterogeneity necessary [21,22,126,168]?

The rediscovery of the multi-slider-block-spring model of [31,33] led to a flurry of investigations by physicists [34,128,170], finding an enticing entry to this difficult field, in the hope of capturing the main empirical statistical properties of seismicity. It is now understood that complexity in the stress field, in co-seismic slips and

in sequences of earthquakes can emerge purely from the nonlinear laws. However, heterogeneity is probably the most important factor dominating the multi-scale complex nature of earthquakes and faulting [156,157,181,182]. It is also known to control the appearance of self-organized critical behavior in a class of models relevant to the crust [191,214].

### Critical Earthquakes

This section gives a brief history of the “critical earthquake” concept.

We trace the ancestor of the critical earthquake concept to Vere-Jones [223], who used a branching model to illustrate that rupture can proceed through a cascade of damage events. Allègre et al. [7] proposed what is in essence a percolation model of damage/rupture describing the state of the crust before an earthquake. They formulated the model using the language of real-space renormalization group, in order to emphasize the multi-scale nature of the underlying physics, and the incipient rupture as the approach to a critical percolation point. Their approach is actually a reformulation in the language of earthquakes of the real-space renormalization group approach to percolation developed by [165]. Chelidze [35] independently developed similar ideas. In the same spirit, Smalley et al. [192] proposed a renormalization group treatment of a multi-slider-block-spring model. Sornette and Sornette [194] took seriously the concept put forward by [7] and proposed to test it empirically by searching for the predicted critical precursors. Voight [227,228] was probably the first author to introduce the idea of a time-to-failure analysis quantified by a second order nonlinear ordinary differential equation. For certain values of the parameters, the solution of [227,228]’s time-to-failure equation takes the form of a finite time singularity (see [180] for a review and [204] for a mechanism based on the ETAS model). He proposed and did use it later to predict volcanic eruptions. The concept that earthquakes are somehow associated with critical phenomena was also underlying the research efforts of a part of the Russian school [120,222].

The empirical seed for the critical earthquake concept were the repeated observations that large earthquakes are sometimes preceded by an increase in the number of intermediate size events [29,59,92,96,121,123,131,144,164,217]. The relation between these intermediate events and the subsequent main event took a long time to be recognized because the precursory events occur over such a large area. Sykes and Jaumé [217] proposed a specific law  $\sim \exp[t/\tau]$  quantifying the acceleration of seismicity prior to large

earthquakes. Bufe and Varnes [30] proposed that the finite-time singularity law

$$\epsilon_{\text{Benioff}} \sim 1/(t_c - t)^m \quad (2)$$

is a better empirical model than the exponential law. In (2),  $\epsilon_{\text{Benioff}}$  is the cumulative Benioff strain,  $t_c$  is critical time of the occurrence of the target earthquake and  $m$  is a positive exponent. The fit with this law of the empirical Benioff strain calculated by summing the contribution of earthquakes in a given space-time window is supposed to provide the time  $t_c$  of the earthquake and thus constitutes a prediction. This expression (2) was justified by a mechanical model of material damage. It is important to understand that the law (2) can emerge as a consequence of a variety of mechanisms, as reviewed by [180].

One of these mechanisms has been coined the “critical earthquake” concept, first formulated by Sornette and Sammis [207], who proposed to reinterpret the formula (2) proposed by [30] and previous related works by generalizing them within the statistical physics framework. This concept views a large earthquake as a genuine critical point. Using the insight of critical points in rupture phenomena, Sornette and Sammis [207] proposed to enrich Eq. (2), now interpreted as a kind of diverging susceptibility in the sense of critical phenomena, by considering complex exponents (i. e. log-periodic corrections to scaling). These structures accommodate the possible presence of a hierarchy of characteristic scales, coexisting with power laws expressing the scale invariance associated with a critical phenomenon [199]. This was followed by several extensions [89,94,95,178]. Sornette [202] reviewed the concept of critical “ruptures” and earthquakes with application to prediction. Ike and Sornette [90] presented a simple dynamical mechanism to obtain finite-time singularities (in rupture in particular) decorated by complex exponents (log-periodicity). Bowman et al. [28, 153,154,242,243] proposed empirical tests of the critical earthquake concept. The early tests of [28] have been criticized by [74], while [226] commented on the lack of a formal statistical basis of the accelerating moment release model. This stresses the need for rigorous tests in the spirit of [166,167]. The debate is wide open, especially in view of the recent developments to improve the determination of the relevant spatio-temporal domain that should be used to perform the analyzes [26,27,124,130] (see [62] for a review).

### Spinodal Decomposition

Klein, Rundle and their collaborators have suggested a mean-field approach to the multi-slider-block-spring

model justified by the long-range nature of the elastic interactions between faults. This has led them to propose that the fluctuations of the strain and stress field associated with earthquakes are technically those occurring close to a spinodal line of an underlying first-order phase transition (see [125,169,172] and references therein). This conceptual view has inspired them to develop the “Pattern Informatics” technique, an empirical seismicity forecasting method based on the idea that changes in the seismicity rate are proxies for changes in the underlying stress [86, 218].

The fluctuations associated with a spinodal line are very similar to those observed in critical phenomena. It is thus very difficult if not impossible in principle to falsify this hypothesis against the critical earthquake hypothesis, since both are expected to present similar if not identical signatures. Perhaps, the appeal of the spinodal decomposition proposal has to be found at the theoretical level, from the fact that first-order phase transitions are more generic and robust than critical phenomena, for systems where heterogeneity and quenched randomness are not too large.

### Dynamics, Stress Interaction and Thermal Fluctuation Effects

Fisher et al. [24], Dahmen et al. [47], Ben-Zion et al. [61] and co-workers (see the reviews by ► [Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of](#) and ► [Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space](#)) have introduced a mean-field model (resulting from a uniform long-range Green function) of a single fault, whose dynamical organization is controlled by two control parameters,  $\epsilon$  which measures the dynamic stress weakening and  $c$  which is the deviation from stress conservation (due for instance to coupling with ductile layers). The point ( $\epsilon = 0$ ;  $c = 0$ ) is critical in the sense of a phase transition in statistical physics, with its associated scale invariant fluctuations described by power laws. Dynamic stress strengthening ( $\epsilon < 0$ ) leads to truncated Gutenberg–Richter power laws. Dynamic stress weakening ( $\epsilon > 0$ ) is either associated with a truncated Gutenberg–Richter power law for  $c > 0$  or with characteristic earthquakes decorating a truncated power law for  $c < 0$ . The coexistence of a characteristic earthquake regime with a power-law regime is particularly interesting as it suggests that they are not exclusive properties but may characterize the same underlying physics under slightly different conditions. This could provide a step towards explaining the variety of empirical observations in seismology [5,110,185,232].



Sornette et al. [214] have obtained similar conclusions using a quasi-static model in which faults grow and self-organize into optimal structures by repeated earthquakes. Depending on the value of dynamical stress drop (controlling the coupling strength between elements) relative to the amplitude of the frozen heterogeneity of the stress thresholds controlling the earthquake nucleation on each fault segment, a characteristic earthquake regime with a truncated power law is found for small heterogeneity or large stress drop while the power law SOC regime is recovered by large heterogeneity or small stress drop. The two approaches of [24,47,61] and ► **Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of** on the one hand and of [214] on the other hand can be reconciled conceptually by noting that the dynamic stress weakening of ► **Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of** controls the dynamical generated stress heterogeneity while the lack of stress conservation  $c$  controls the coupling strength. Fundamentally, the relevant control parameter is the degree of coupling between fault elements seen as threshold oscillators of relaxation versus the variance of the disorder in their spontaneous large earthquake recurrence times. Generically, power law statistics are expected to co-exist with synchronized behavior in a general phase diagram in the heterogeneity-coupling strength plane [152,214].

Let us finally mention a promising but challenging theoretical approach, which has the ambition to bridge the small-scale physics controlled by thermal nucleation of rupture to the large-scale organization of earthquakes and faults [155,205]. Partial success has been obtained with a remarkable prediction on the (“multifractal”) dependence of the Omori law exponent of aftershocks on the magnitude of the mainshock, verified by careful empirical analyzes on earthquakes in California, Japan and worldwide [158].

### Empirical Studies of Seismicity Inspired by Statistical Physics

*“False facts are highly injurious to the progress of science, for they often endure long; but false views, if supported by some evidence, do little harm, for everyone takes a salutary pleasure in proving their falseness.”*

Charles Darwin, in *The Origin of Man*, Chap. 6.

### Early Successes and Subsequent Challenges

A significant benefit of the statistical physics approach to seismology has been the introduction of novel techniques

to analyze the available empirical data sets, with the goal of obtaining new insights into the spatio-temporal organization of seismicity and of revealing novel regularities and laws that may guide the theoretical analysis.

A prominent forerunner is the application of the concept of fractals introduced by Mandelbrot [137] and of the measures of fractal dimensions to describe complex sets of earthquake epicenters, hypocenters and fault patterns. The use of fractals has constituted an epistemologic breakthrough in many fields, and not only in seismology. Indeed, before Mandelbrot, when dealing with most complex systems, one used to say: “this is too complicated for a quantitative analysis” and only qualitative descriptions were offered. After Mandelbrot, one could hear: “this is a fractal, with a fractal dimension equal to xxx!” By providing a new geometrical way of thinking about complex systems associated with novel metrics, Mandelbrot and his fractals have extended considerably the reach of quantitative science to many complex systems in all fields.

However, while there have been some attempts to use fractal dimensions as guidelines to infer the underlying organization processes, as for instance in [195,196], most of the initial reports have lost their early appeal [18,19,183,219] since the complexity of seismicity and faulting is much too great to be captured by scaling laws embodying solely a simple scale invariance symmetry. Among others, multifractal and adapted wavelet tools are needed to quantify this complexity, see for instance [68,146,156,157,159]. It should also be noted that few studies of the fractal dimensions of seismicity address the significant issues of errors, biases and incomplete records in earthquake catalogs – a notable exception being [115].

Since the beginning of the 21st century, a renewal of interest and efforts have burgeoned as groups of statistical physicists, interested in earthquakes as a potential instance of self-organized criticality (SOC), have claimed “novel”, “universal” and “robust” scaling laws from their analysis of the spatio-temporal organization of seismicity. The authors purport to have discovered universal and hitherto unknown features of earthquakes that give new insights into the dynamics of earthquakes and add to the evidence that earthquakes are self-organized critical. We now discuss a few of these recent studies to illustrate the existence of potential problems in the “statistical physics” approach. In a nutshell, we show that perhaps most of these “novel scaling laws” can be explained entirely by already known statistical seismicity laws. This claim has been defended by other experts of statistical seismology, the most vocal being perhaps Yan Kagan at UCLA [111].

The flurry of interest from physicists comes from their fascination with the self-similar properties exhibited by

seismicity (e. g. the Gutenberg–Richter power law of earthquake seismic moments, the Omori–Utsu law of the decay of aftershock rates after large earthquakes, the fractal and multifractal space-time organization of earthquakes and faults, etc.), together with the development of novel concepts and techniques that may provide new insights. But, and this is our main criticism based on several detailed examples discussed below, many of the new approaches and results do not stand close scrutiny. This failure is rooted in two short-comings: (i) the lack of testing of new methods on synthetic catalogs generated by benchmark models which are based on well-known statistical laws of seismicity, and (ii) the failure to consider earthquake catalog bias, incompleteness and errors. The latter may cause catalog artifacts to appear as genuine characteristics of earthquakes. Testing the results on a variety of catalogs and considering the influence of various catalog errors can help minimize their influence. The former short-coming often leads to the following scenario: authors fail to realize that a simpler null hypothesis could not be rejected, namely that their “discovery” could actually be explained by just a combination of basic statistical laws known to seismologists for decades.

The well-established laws of statistical seismicity that authors should consider before claiming for novelty include the following:

1. The Gutenberg–Richter law for the distribution of earthquake magnitudes with a  $b$ -value close to 1 (corresponding to an exponent  $\simeq 2/3$  for the probability density function of seismic moments),
2. The Omori–Utsu law for the decay of the rate of aftershocks following a mainshock,
3. The inverse Omori law for foreshocks,
4. The fact that aftershocks also trigger their own aftershocks and so on, and that aftershocks do not seem to exhibit any distinguishable physical properties,
5. The fact that the distribution of distances between mainshocks and aftershocks has a power law tail,
6. The fertility law (the fact that earthquakes of magnitude  $M$  trigger of the order of  $10^{aM}$  aftershocks with  $a \sim b \simeq 1$ ),
7. The fractal distribution of faults which are concentration centers for earthquakes.

This above non-exhaustive list selects “laws” which are arguably non-redundant, in the sense that it is likely not possible to derive one of these laws from the others (a possible exception is the inverse Omori law for foreshock, which can be derived from the direct Omori law for aftershocks in the context of the ETAS model [78,81]). Some

experts would argue that we should add to this list other claimed regularities, such as “Båth’s law” (see e. g. [190] for a recent discussion emphasizing the importance of this law), that states that the differences in magnitudes between mainshocks and their largest aftershocks are approximately constant, independent of the magnitudes of mainshocks [20]. However, Helmstetter and Sornette [79] and Saichev and Sornette [173] have shown that Bath’s law can be accurately recovered in ETAS-type models combining the first, second, fourth, and sixth laws stated above, with the assumption that any earthquake can trigger subsequent earthquakes.

### Entropy Method for the Distribution of Time Intervals Between Mainshocks

Mega et al. [140] used the “diffusion entropy” method to argue for a power-law distribution of time intervals between a large earthquake (the mainshock of a seismic sequence or cluster) and the next one. Helmstetter and Sornette [80] showed that all the “new” discoveries reported by [140] (including the supposedly new scaling) can be explained solely by Omori’s law for intra-cluster times, without correlation between clusters, thus debunking the claim for novelty.

### Scaling of the PDF of Waiting Times

Bak et al. [14] analyzed the scaling of the probability density function of waiting times between successive earthquakes in southern California as a function of “box size” or small regions in which subsequent earthquakes are considered. They found an approximate collapse of the pdfs for different seismic moment thresholds  $S$  and box sizes  $L$  which suggested the following scaling ansatz for the waiting times  $T$ :

$$T^\alpha P_{S,L}(T) = f(TS^{-b}L^{d_f}), \quad (3)$$

where  $b = 1$  is the Gutenberg–Richter exponent,  $d_f \simeq 1.2$  was claimed to be a spatial fractal dimension of seismicity (see [146] and [115] for more in-depth studies),  $\alpha = 1$  was identified as the exponent in the Omori law and  $f(\cdot)$  is a scaling function which was proposed to be roughly constant up to a constant (“kink”) beyond which it quickly decays. The scaling (3) was claimed to be a unified law for earthquakes that revealed a novel feature in the spatio-temporal organization of seismicity in that the Gutenberg–Richter, the Omori law and the spatial distribution of earthquakes were unified into a single picture that made no distinction between fore-, main- and aftershocks. The

scaling relations and critical exponents were claimed to be contained in the scaling ansatz. Corral [40,41,42,43] and others broadened the analysis to other regions of the world. Corral [41] proposed a slightly different scaling ansatz for a modified data analysis.

Early criticism came from Lindman et al. [132], who noted that synthetic data generated using a non-homogeneous Poisson process derived from Omori's law was able to reproduce some of the results of [14], indicating a rather trivial origin of the unified scaling law. Molchan [145] showed that, if at least two regions in the data set are independent, then, if a scaling relation were to hold exactly, this scaling function could only be exponential. All other functions could only result in approximate data collapses. Proponents of the unified scaling law, e. g. [44], argued that indeed all regions were correlated, as expected in systems near a critical point so that the assumption of independence between different regions should not hold. But Molchan [145] also showed that a simple Poisson cluster model (Poissonian mainshocks that trigger Omori-type aftershock sequences) could reproduce the short and long time limits of the observed statistics, indicating that the Omori law, the Gutenberg–Richter relationship and simple clustering were the sole ingredients necessary for the observed short and long time limit, and no spatial correlation was needed.

Saichev and Sornette [176,177] extended Molchan's arguments to show that the approximate data collapse of the waiting times could be explained completely by the Epidemic-Type Aftershock Sequence (ETAS) model of [149]. This provided further evidence that the apparent data collapse was only approximate. Remarkably, the theoretical predictions of the ETAS model seem to fit the observed data better than the phenomenological scaling function proposed by [41] to fit the data. Saichev and Sornette [176,177] thus showed that a benchmark model of seismicity was able to reproduce the apparent unified scaling law and that therefore the distribution of interevent times did not reveal new information beyond what was already known via statistical laws: The combination of the Gutenberg–Richter law, the Omori law, and the concept of clustering suffice to explain the apparent "universal" scaling of the waiting times.

Sornette et al. [215] developed an efficient numerical scheme to solve accurately the set of nonlinear integral equations derived previously in [177] and found a dramatic lack of power for the distribution of inter-event times to distinguish between quite different sets of parameters, casting doubt on the usefulness of this statistics for the specific purpose of identifying the clustering parameter (e. g. [72]).

### Scaling of the PDF of Distances Between Subsequent Earthquakes

Davidson and Paczuski [49] claimed evidence contradicting the theory of aftershock zone scaling in favor of scale-free statistics. Aftershock zone scaling refers to the scaling of the mainshock rupture length, along which most aftershocks occur, with the mainshock magnitude [112]. Davidson and Paczuski [49] suggested that the probability density function of spatial distances between successive earthquakes obeys finite size scaling with a novel dynamical scaling exponent, suggesting that the mainshock rupture length scale has no impact on the spatial distribution of aftershocks and that earthquakes are self-organized critical.

Werner and Sonette [230] debunked this claim by showing that (i) the purported power law scaling function is not universal as it breaks down in other regions of the world; (ii) the results obtained by [49] for southern California depend crucially on a single earthquake (the June 28, 1992, M7.3 Landers earthquake): without Landers and its aftershocks, the power law disappears; (iii) a model of clustered seismicity, with aftershock zone scaling explicitly built in, is able to reproduce the apparent power law, indicating that an apparent lack of scales in the data does not necessarily contradict aftershock zone scaling and the existence of scales associated with mainshock rupture length scales.

### The Network Approach

The recent boom in the statistical mechanics of network analysis has recently extended to applications well beyond physics (for reviews, see [6,16,57,148]). Earthquake seismology is no exception [1,2,3,4,8,9,10,160,161]. The resulting impact has been limited so far for several reasons.

A major concern is the assumption that earthquake catalogs as downloaded from the web are data sets fit for immediate analysis. References [82,113,229] and [231] present modern and complementary assessments of the many issues of incompleteness spoiling even the best catalogs. In particular, we should stress that magnitude determinations are surprisingly inaccurate, leading to large errors in seismic rate estimates [231]. Furthermore, there is no such thing as a complete catalog above a so-called magnitude of completeness, due to the fact that a non-negligible fraction of earthquakes are missed in the aftermath of previous earthquakes [82,113]. One should be concerned that analyses in terms of network metrics could be particularly sensitive to these defects. Nevertheless, Abe and Suzuki [1,2,3,4] applied metrics of network analysis to "raw" catalogs which included events well below the esti-

mated magnitude of completeness. As a result of neglecting to use a (reasonably) homogeneous and trustworthy data set, the results of their analysis may be severely biased, because the reliability of the inferred network structure is probably more sensitive than other metrics to the correct spatio-temporal ordering of the earthquake catalog. No serious study has yet been performed to quantify the usually serious impact of quality issues on the metrics used in network analysis. As a consequence, it is also not clear how to interpret the “success” of [160,161] in reproducing the “features” of Abe and Suzuki’s analysis on the synthetic seismicity generated by a spring-block model.

In addition, at best limited attempts have been made to interpret the results of the new network metrics using well-known, established facts in seismology. Many of the claimed novel features are probably very well understood – they are mostly related to scale-invariance and clustering of seismicity, facts documented for decades. The authors should always strive to show that the new metrics that they propose give results that cannot be explained by the standard laws in statistical seismology. Toward this end, there are well-defined benchmark models that incorporate these laws and that can generate synthetic catalogs on which the new metrics can be tested and compared.

A few exceptions are worth mentioning. Motivated by the long-standing and unresolved debate over “aftershock” identification, Baiesi and Paczusi [9,10] and Baiesi [8] provided a new metric for the correlations between earthquakes based on the space-time-magnitude nearest-neighbor distance between earthquakes. The authors compared their results with known statistical laws in seismology and with the predictions of the ETAS model, actually confirming both. While no new law has been unearthed here, such efforts are valuable to validate known laws and continue to test the possible limits. Zaliapin et al. [238] extended their study and investigated the theoretical properties of the metric and its ability to decluster catalogs (i. e., separate mainshocks from aftershocks). They concluded that aftershocks defined from this metric seem to be different from the rest of earthquakes. It will be interesting to see head-to-head comparisons with current state-of-the-art probabilistic declustering techniques that are based on empirical statistical laws and likelihood estimation [105, 240,241].

### Future Directions

The study of the statistical physics of earthquakes remains wide-open with many significant discoveries to be made. The promise of a holistic approach – one that emphasizes

the interactions between earthquakes and faults – is to be able to neglect some of the exceedingly complicated micro-physics when attempting to understand the large scale patterns of seismicity. The marriage between this conceptual approach, based on the successes of statistical physics, and seismology thus remains a highly important domain of research. In particular, statistical seismology needs to evolve into a genuine physically-based statistical physics of earthquakes.

The question of renormalizability of models of earthquake occurrence and the role of small earthquakes in the organization of seismicity is likely to remain an important topic. It connects with the problem of foreshocks and the predictability of large events from small ones and therefore has real and immediate practical applications as well as physical implications.

More detailed and rigorous empirical studies of the frequency-size statistics of earthquake seismic moments and how they relate to seismo-tectonic conditions are needed in order to help settle the controversy over the power-law versus the characteristic event regime, and the role of regime-switching and universality.

Spatially extended, dynamically evolving fault networks and their role in the generation of earthquakes are mostly ignored in the statistical physics approach to seismicity. Akin to the filaments in turbulence, these may provide key insights into the spatio-temporal organization of earthquakes. Novel methods combining information from seismology to faulting will be required (e. g., [71,159,195, 196,197]) to build a real understanding of the self-organization of the chicken-and-egg structures that earthquakes-faults constitute. Furthermore, a true physical approach requires understanding the spatio-temporal evolution of stresses, their role in earthquake nucleation via thermally activated processes, in the rupture propagation and in the physics of arrest, both involved in the generation of complex stress fields.

The important debate regarding statistical physics approaches to seismicity would benefit significantly from two points. Firstly, earthquake catalogs contain data uncertainties, biases and subtle incompleteness issues. Investigating their influence on the results of data analyses inspired by statistical physics increases the relevance of the results. Secondly, the authors should make links with the literature on statistical seismology which deals with similar questions. It is their task to show that the new metrics that they propose give results that cannot be explained by the standard laws in statistical seismology. For this, there are well-defined benchmark models that incorporate these laws and that can generate synthetic catalogs on which the new metrics can be tested.

## Bibliography

1. Abe S, Suzuki N (2004) Scale-free network of earthquakes. *Europhys Lett* 65:581–586. doi:10.1209/epl/i2003-10108-1
2. Abe S, Suzuki N (2004) Small-world structure of earthquake network. *Physica A: Stat Mech Appl* 337:357–362. doi:10.1016/j.physa.2004.01.059
3. Abe S, Suzuki N (2005) Scale-invariant statistics of period in directed earthquake network. *Eur Phys J B* 44:115–117. doi:10.1140/epjb/e2005-00106-7
4. Abe S, Suzuki N (2006) Complex earthquake networks: Hierarchical organization and assortative mixing. *Phys Rev E* 74(2):026, 113–+. doi:10.1103/PhysRevE.74.026113
5. Aki K (1995) Earthquake prediction, societal implications. *Rev Geophys* 33:243–248
6. Albert R, Barabási AL (2002) Statistical mechanics of complex networks. *Rev Mod Phys* 74(1):47–97. doi:10.1103/RevModPhys.74.47
7. Allègre CJ, Le Mouél JL, Provost A (1982) Scaling rules in rock fracture and possible implications for earthquake prediction. *Nature* 297:47–49. doi:10.1038/297047a0
8. Baiesi M (2006) Scaling and precursor motifs in earthquake networks. *Physica A: Stat Mech Appl* 359:775–783. doi:10.1016/j.physa.2005.05.094
9. Baiesi M, Paczuski M (2004) Scale-free networks of earthquakes and aftershocks. *Phys Rev E* 69(6):066, 106. doi:10.1103/PhysRevE.69.066106
10. Baiesi M, Paczuski M (2005) Complex networks of earthquakes and aftershocks. *Nonlin Proc Geophys* 12:1–11
11. Bak P (1996) *How Nature Works: The Science of Self-Organized Criticality*. Springer, New York, p 212
12. Bak P, Tang C (1989) Earthquakes as a self-organized critical phenomena. *J Geophys Res* 94(B11):15635–15637
13. Bak P, Tang C, Wiesenfeld K (1987) Self-organized criticality: An explanation of the  $1/f$  noise. *Phys Rev Lett* 59(4):381–384. doi:10.1103/PhysRevLett.59.381
14. Bak P, Christensen K, Danon L, Scanlon T (2002) Unified scaling law for earthquakes. *Phys Rev Lett* 88(17):178,501. doi:10.1103/PhysRevLett.88.178501
15. Bakun, WH, Aagaard B, Dost B, Ellsworth WL, Hardebeck JL, Harris RA, Ji C, Johnston MJS, Langbein J, Lienkaemper JJ, Michael AJ, Murray JR, Nadeau RM, Reasenber PA, Reichle MS, Roeloffs EA, Shakal A, Simpson RW, Waldhauser F (2005) Implications for prediction and hazard assessment from the 2004 Parkfield earthquake. *Nature* 437:969–974. doi:10.1038/nature04067
16. Barabási AL, Albert R (1999) Emergence of Scaling in Random Networks. *Science* 286(5439):509–512. doi:10.1126/science.286.5439.509
17. Barabási AL, Albert R, Jeong H (1999) Mean-field theory for scale-free random networks. *Physica A* 272:173–187. doi:10.1016/S0378-4371(99)00291-5
18. Barton CC, La Pointe PR (eds) (1995) *Fractals in the Earth Sciences*. Plenum Press, New York, London
19. Barton CC, La Pointe PR (eds) (1995) *Fractals in petroleum geology and earth processes*. Plenum Press, New York, London
20. Báth M (1965) Lateral inhomogeneities in the upper mantle. *Tectonophysics* 2:483–514
21. Ben-Zion Y, Rice JR (1993) Earthquake failure sequences along a cellular fault zone in a 3-dimensional elastic solid containing asperity and nonasperity regions. *J Geophys Res* 93:14109–14131
22. Ben-Zion Y, Rice JR (1995) Slip patterns and earthquake populations along different classes of faults in elastic solids. *J Geophys Res* 100:12959–12983
23. Ben-Zion Y, Rice JR (1997) Dynamic simulations of slip on a smooth fault in an elastic solid. *J Geophys Res* 102:17771–17784
24. Ben-Zion Y, Dahmen K, Lyakhovskiy V, Ertas D, Agnon A (1999) Self-driven mode switching of earthquake activity on a fault system. *Earth Planet Sci Lett* 172:11–21
25. Bird P, Kagan YY (2004) Plate-tectonic analysis of shallow seismicity: Apparent boundary width, beta, corner magnitude, coupled lithosphere thickness, and coupling in seven tectonic settings. *Bull Seismol Soc Am* 94(6):2380–2399
26. Bowman DD, King GCP (2001) Stress transfer and seismicity changes before large earthquakes. *C Royal Acad Sci Paris, Sci Terre Planetes* 333:591–599
27. Bowman DD, King GCP (2001) Accelerating seismicity and stress accumulation before large earthquakes. *Geophys Res Lett* 28:4039–4042
28. Bowman DD, Oullion G, Sammis CG, Sornette A, Sornette D (1998) An observational test of the critical earthquake concept. *J Geophys Res* 103:24359–24372
29. Brehm DJ, Braille LW (1998) Intermediate-term earthquake prediction using precursory events in the New Madrid Seismic Zone. *Bull Seismol Soc Am* 88(2):564–580
30. Bufe CG, Varnes DJ (1993) Predictive modeling of the seismic cycle of the greater San Francisco Bay region. *J Geophys Res* 98:9871–9883
31. Burridge R, Knopoff L (1964) Body force equivalents for seismic dislocation. *Seism Soc Am Bull* 54:1875–1888
32. Cardy JL (1996) *Scaling and Renormalization in Statistical Physics*. Cambridge University Press, Cambridge
33. Carlson JM, Langer JS (1989) Properties of earthquakes generated by fault dynamics. *Phys Rev Lett* 62:2632–2635
34. Carlson JM, Langer JS, Shaw BE (1994) Dynamics of earthquake faults. *Rev Mod Phys* 66:657–670
35. Chelidze TL (1982) Percolation and fracture. *Phys Earth Planet Interiors* 28:93–101
36. Christensen K, Farid N, Pruessner G, Stapleton M (2008) On the finite-size scaling of probability density functions. *Eur Phys B* 62:331–336
37. Clauset A, Shalizi CR, Newman MEJ (2007) Power-law distributions in empirical data. E-print arXiv:0706.1062
38. Cochard A, Madariaga R (1994) Dynamic faulting under rate-dependent friction. *Pure Appl Geophys* 142:419–445
39. Cochard A, Madariaga R (1996) Complexity of seismicity due to highly rate-dependent friction. *J Geophys Res* 101:25321–25336
40. Corral A (2003) Local distributions and rate fluctuations in a unified scaling law for earthquakes. *Phys Rev E* 68(3):035, 102. doi:10.1103/PhysRevE.68.035102
41. Corral A (2004) Universal local versus unified global scaling laws in the statistics of seismicity. *Physica A* 340:590–597
42. Corral A (2004) Long-term clustering, scaling, and universality in the temporal occurrence of earthquakes. *Phys Rev Lett* 92:108, 501
43. Corral A (2005) Mixing of rescaled data and bayesian inference for earthquake recurrence times. *Nonlin Proc Geophys* 12:89–100

44. Corral A (2005) Renormalization-group transformations and correlations of seismicity. *Phys Rev Lett* 95:028, 501
45. Corral A (2006) Universal earthquake-occurrence jumps, correlations with time, and anomalous diffusion. *Phys Rev Lett* 97:178, 501
46. Corral A, Christensen K (2006) Comment on "earthquakes descaled: On waiting time distributions and scaling laws". *Phys Rev Lett* 96:109, 801
47. Dahmen K, Ertas D, Ben-Zion Y (1998) Gutenberg–Richter and characteristic earthquake behavior in simple mean-field models of heterogeneous faults. *Phys Rev E* 58:1494–1501. doi:10.1103/PhysRevE.58.1494
48. Davidsen J, Goltz C (2004) Are seismic waiting time distributions universal? *Geophys Res Lett* 31:L21612. doi:10.1029/2004GL020892
49. Davidsen J, Paczuski M (2005) Analysis of the spatial distribution between successive earthquakes. *Phys Rev Lett* 94:048, 501. doi:10.1103/PhysRevLett.94.048501
50. Davidsen J, Grassberger P, Paczuski M (2006) Earthquake recurrence as a record breaking process. *Geophys Res Lett* 33:L11304. doi:10.1029/2006GL026122
51. Dhar D (1990) Self-organized critical state of sandpile automaton models. *Phys Rev Lett* 64:1613–1616
52. Dhar D (1999) The Abelian sandpile and related models. *Physica A* 263:4–25
53. Dieterich JH (1987) Nucleation and triggering of earthquake slip; effect of periodic stresses. *Tectonophysics* 144:127–139
54. Dieterich JH (1992) Earthquake nucleation on faults with rate-dependent and state-dependent strength. *Tectonophysics* 211:115–134
55. Dieterich J (1994) A constitutive law for rate of earthquake production and its application to earthquake clustering. *J Geophys Res* 99:2601–2618
56. Dieterich J, Kilgore BD (1994) Direct observation of frictional contacts- New insight for state-dependent properties. *Pure Appl Geophys* 143:283–302
57. Dorogovtsev SN, Mendes JFF (2003) *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press, New York
58. Eckman JP (1981) Roads to Turbulence in Dissipative Dynamical Systems. *Rev Mod Phys* 53:643–654
59. Ellsworth WL, Lindh AG, Prescott WH, Herd DJ (1981) The 1906 San Francisco Earthquake and the seismic cycle. *Am Geophys Union Maurice Ewing Monogr* 4:126–140
60. Felzer KR, Becker TW, Abercrombie RE, Ekstrom G, Rice JR (2002) Triggering of the 1999 Mw 7.1 Hector Mine earthquake by aftershocks of the 1992 Mw 7.3 Landers earthquake. *J Geophys Res* 107(B09):2190
61. Fisher DS, Dahmen K, Ramanathan S, Ben-Zion Y (1997) Statistics of Earthquakes in Simple Models of Heterogeneous Faults. *Phys Rev Lett* 78:4885–4888. doi:10.1103/PhysRevLett.78.4885
62. Freund F, Sornette D (2007) Electro-Magnetic Earthquake Bursts and Critical Rupture of Peroxy Bond Networks in Rocks. *Tectonophysics* 431:33–47
63. Frisch U (1995) *Turbulence. The legacy of A.N. Kolmogorov*. Cambridge University Press, Cambridge
64. Gabriellov A, Keilis-Borok V, Jackson DD (1996) Geometric Incompatibility in a Fault System. *Proc Nat Acad Sci* 93:3838–3842
65. Gabriellov A, Keilis-Borok V, Zaliapin I, Newman W (2000) Critical transitions in colliding cascades. *Phys Rev E* 62:237–249
66. Gabriellov A, Zaliapin I, Newman W, Keilis-Borok V, (2000) Colliding cascades model for earthquake prediction. *Geophys J Int* 143:427–437
67. Gallagher R, Appenzeller T (1999) Beyond Reductionism. *Science* 284(5411):79
68. Geilikman MB, Pisarenko VF, Golubeva TV (1990) Multifractal Patterns of Seismicity. *Earth Planet Sci Lett* 99:127–138
69. Gelfand IM, Guberman SA, Keilis-Borok VI, Knopoff L, Press F, Ranzman EY, Rotwain IM, Sadovsky AM (1976) Pattern recognition applied to earthquake epicenters in California. *Phys Earth Planet Interiors* 11:227–283
70. Geller RJ, Jackson DD, Kagan YY, Mulargia F (1997) Earthquakes cannot be predicted. *Science* 275:1616–1617
71. Gorshkov A, Kossobokov V, Soloviev A (2003) Recognition of earthquake-prone areas. In: Keilis-Borok V, Soloviev A (eds) *Nonlinear Dynamics of the Lithosphere and Earthquake Prediction*. Springer, Heidelberg, pp 239–310 [122]
72. Hainzl S, Scherbaum F, Beauval C (2006) Estimating Background Activity Based on Interevent-Time Distribution. *Bull Seismol Soc Am* 96(1):313–320. doi:10.1785/0120050053
73. Hanks TC (1992) Small earthquakes, tectonic forces. *Science* 256:1430–1432
74. Hardebeck JL, Felzer KR, Michael AJ (2008) Improved tests reveal that the accelerating moment release hypothesis is statistically insignificant. *J Geophys Res* 113:B08310. doi:10.1029/2007JB005410
75. Harris RA, Arrowsmith JR (2006) Introduction to the Special Issue on the 2004 Parkfield Earthquake and the Parkfield Earthquake Prediction Experiment. *Bull Seismol Soc Am* 96(4B):S1–10. doi:10.1785/0120050831
76. Helmstetter A (2003) Is earthquake triggering driven by small earthquakes? *Phys Rev Lett* 91(5):058, 501. doi:10.1103/PhysRevLett.91.058501
77. Helmstetter A, Sornette D (2002) Subcritical and supercritical regimes in epidemic models of earthquake aftershocks. *J Geophys Res* 107(B10):2237. doi:10.1029/2001JB001580
78. Helmstetter A, Sornette D (2003) Foreshocks explained by cascades of triggered seismicity. *J Geophys Res (Solid Earth)* 108(B10):2457 doi:10.1029/2003JB00240901
79. Helmstetter A, Sornette D (2003) Bath's law Derived from the Gutenberg–Richter law and from Aftershock Properties. *Geophys Res Lett* 30:2069. doi:10.1029/2003GL018186
80. Helmstetter A, Sornette D (2004) Comment on "Power-Law Time Distribution of Large Earthquakes". *Phys Rev Lett* 92:129801 (Reply is *Phys Rev Lett* 92:129802 (2004))
81. Helmstetter A, Sornette D, Grasso J-R (2003) Mainshocks are Aftershocks of Conditional Foreshocks: How do foreshock statistical properties emerge from aftershock laws. *J Geophys Res* 108(B10):2046. doi:10.1029/2002JB001991
82. Helmstetter A, Kagan YY, Jackson DD (2005) Importance of small earthquakes for stress transfers and earthquake triggering. *J Geophys Res* 110:B05508. doi:10.1029/2004JB003286
83. Helmstetter A, Kagan Y, Jackson D (2006) Comparison of short-term and long-term earthquake forecast models for Southern California. *Bull Seism Soc Am* 96:90–106
84. Hergarten S (2002) *Self-Organized Criticality in Earth Systems*. Springer, Berlin
85. Hinrichsen H (2000) *Non-equilibrium critical phenomena and*

- phase transitions into absorbing states. *Adv Phys* 49:815–958(144)
86. Holliday JR, Rundle JB, Tiampo KF, Klein W, Donnellan A (2006) Systematic procedural and sensitivity analysis of the Pattern Informatics method for forecasting large ( $M > 5$ ) earthquake events in Southern California. *Pure Appl Geophys* 163(11–12):2433–2454
  87. Huang J, Turcotte DL (1990) Evidence for chaotic fault interactions in the seismicity of the San Andreas fault and Nankai trough. *Nature* 348:234–236
  88. Huang J, Turcotte DL (1990) Are earthquakes an example of deterministic chaos? *Geophys Rev Lett* 17:223–226
  89. Huang Y, Saleur H, Sammis CG, Sornette D (1998) Precursors, aftershocks, criticality and self-organized criticality. *Europhys Lett* 41:43–48
  90. Ide K, Sornette D (2002) Oscillatory Finite-Time Singularities in Finance, Population and Rupture. *Physica A* 307(1–2):63–106
  91. Jackson DD, Kagan YY (2006) The 2004 Parkfield Earthquake, the 1985 Prediction, and Characteristic Earthquakes: Lessons for the Future. *Bull Seismol Soc Am* 96(4B):S397–409. doi:10.1785/0120050821
  92. Jaumé SC, Sykes LR (1999) Evolving Towards a Critical Point: A Review of Accelerating Seismic Moment/Energy Release Prior to Large and Great Earthquakes. *Pure Appl Geophys* 155:279–305
  93. Jensen HJ (1998) *Self-Organized Criticality: Emergent Complex Behavior in Physical and Biological Systems*. Cambridge University Press, Cambridge
  94. Johansen A, Sornette D, Wakita G, Tsunogai U, Newman WI, Saleur H (1996) Discrete scaling in earthquake precursory phenomena: evidence in the Kobe earthquake, Japan *J Phys I France* 6:1391–1402
  95. Johansen A, Saleur H, Sornette D (2000) New Evidence of Earthquake Precursory Phenomena in the 17 Jan. 1995 Kobe Earthquake, Japan. *Eur Phys J B* 15:551–555
  96. Jones LM (1994) Foreshocks, aftershocks, and earthquake probabilities: accounting for the Landers earthquake. *Bull Seismol Soc Am* 84:892–899
  97. Jordan TH (2006) Earthquake Predictability, Brick by Brick. *Seismol Res Lett* 77(1):3–6
  98. Kadanoff LP, Nagel SR, Wu L, Zhou S-M (1989) Scaling and universality in avalanches. *Phys Rev A* 39(12):6524–6537. doi:10.1103/PhysRevA.39.6524
  99. Kagan YY (1981), Spatial distribution of earthquakes: The three-point moment function. *Geophys J R Astron Soc* 67:697–717
  100. Kagan YY (1981) Spatial distribution of earthquakes: The four-point moment function. *Geophys J Roy Astron Soc* 67:719–733
  101. Kagan YY (1987) Point sources of elastic deformation: Elementary sources, static displacements. *Geophys J R Astron Soc* 90:1–34
  102. Kagan YY (1987) Point sources of elastic deformation: Elementary sources, dynamic displacements. *Geophys J R Astron Soc* 91:891–912
  103. Kagan YY (1988) Multipole expansions of extended sources of elastic deformation. *Geophys J R Astron Soc* 93:101–114
  104. Kagan YY (1989) Earthquakes and fractals. *Ann Rev Mater Sci: Fractal Phenom Disordered Syst* 19:520–522
  105. Kagan YY (1991) Likelihood analysis of earthquake catalogs. *Geophys J Int* 106:135–148
  106. Kagan YY (1992) Seismicity: Turbulence of solids. *Nonlinear Sci Today* 2:1–13
  107. Kagan YY (1992) On the geometry of an earthquake fault system. *Phys Earth Planet Interiors* 71:15–35
  108. Kagan YY (1993) Statistics of characteristic earthquakes. *Bull Seismol Soc Am* 83(1):7–24
  109. Kagan YY (1994) Observational evidence for earthquakes as a nonlinear dynamic process. *Physica D* 77:160–192
  110. Kagan YY (1994) Comment on “The Gutenberg–Richter or characteristic earthquake distribution, which is it?” by Wesnousky. *Bull Seismol Soc Am* 86:274–285
  111. Kagan YY (1999) Is earthquake seismology a hard, quantitative science? *Pure Appl Geophys* 155:33–258
  112. Kagan YY (2002) Aftershock Zone Scaling. *Bull Seismol Soc Am* 92(2):641–655. doi:10.1785/0120010172
  113. Kagan YY (2003) Accuracy of modern global earthquake catalogs. *Phys Earth Planet Interiors* 135:173–209
  114. Kagan YY (2006) Why does theoretical physics fail to explain and predict earthquake occurrence? In: Bhattacharyya P, Chakrabarti BK (eds) *Modelling Critical and Catastrophic Phenomena in Geoscience: A Statistical Physics Approach*. Lecture Notes in Physics, vol 705. Springer, Berlin, pp 303–359
  115. Kagan YY (2007) Earthquake spatial distribution: the correlation dimension. *Geophys J Int* 168:1175–1194. doi:10.1111/j.1365-246X.2006.03251.x
  116. Kagan YY, Knopoff L (1980) Spatial distribution of earthquakes: The two-point correlation function. *Geophys J R Astron Soc* 62:303–320
  117. Kagan YY, Knopoff L (1981) Stochastic synthesis of earthquake catalogs. *J Geophys Res* 86(B4):2853–2862
  118. Kagan YY, Knopoff L (1985) The first-order statistical moment of the seismic moment tensor. *Geophys J R Astron Soc* 81:429–444
  119. Kagan YY, Knopoff L (1985) The two-point correlation function of the seismic moment tensor. *Geophys J R Astron Soc* 83:637–656
  120. Keilis-Borok VI (ed) (1990) *Intermediate-term earthquake prediction: models, algorithms, worldwide tests*. *Phys Earth Planet Interiors* 61(1–2)
  121. Keilis-Borok VI, Malinovskaya LN (1964) One regularity in the occurrence of strong earthquakes. *J Geophys Res B* 69:3019–3024
  122. Keilis-Borok V, Soloviev A (2003) *Nonlinear Dynamics of the Lithosphere and Earthquake Prediction*. Springer, Heidelberg
  123. Keilis-Borok VI, Knopoff L, Rotwain IM, Allen CR (1988) Intermediate-term prediction of occurrence times of strong earthquakes. *Nature* 335:690–694
  124. King GCP, Bowman DD (2003) The evolution of regional seismicity between large earthquakes. *J Geophys Res* 108(B2):2096. doi:10.1029/2001JB000783
  125. Klein W, Rundle JB, Ferguson CD (1997) Scaling and nucleation in models of earthquake faults. *Phys Rev Lett* 78:3793–3796
  126. Knopoff L (1996) The organization of seismicity on fault networks. *Proc Nat Acad Sci USA* 93:3830–3837
  127. Landau LD, Lifshitz EM (1980) *Statistical Physics Course on Theoretical Physics*, vol 5, 3rd edn. Butterworth-Heinemann, Oxford

128. Langer JS, Carlson JM, Myers CR, Shaw BE (1996) Slip complexity in dynamical models of earthquake faults. *Proc Nat Acad Sci USA* 93:3825–3829
129. Lee MW, Sornette D, Knopoff L (1999) Persistence and Quiescence of Seismicity on Fault Systems. *Phys Rev Lett* 83(N20):4219–4222
130. Levin SZ, Sammis CG, Bowman DD (2006) An observational test of the stress accumulation model based on seismicity preceding the 1992 Landers, CA earthquake. *Tectonophysics* 413:39–52
131. Lindh AG (1990) The seismic cycle pursued. *Nature* 348:580–581
132. Lindman M, Jonsdottir K, Roberts R, Lund B, Bdvarsson R (2005) Earthquakes descaled: On waiting time distributions and scaling laws. *Phys Rev Lett* 94:108, 501
133. Lindman M, Jonsdottir K, Roberts R, Lund B, Bdvarsson R (2006) Reply to comment by A. Corral and K. Christensen. *Phys Rev Lett* 96:109, 802
134. Livina VN, Havlin S, Bunde A (2006) Memory in the occurrence of earthquakes. *Phys Rev Lett* 95:208, 501
135. Luebeck S (2004) Universal scaling behavior of non-equilibrium phase transitions. *Int J Mod Phys B* 18:3977
136. Manna S (1991) Critical exponents of the sandpile models in two dimensions. *Physica A* 179(2):249–268
137. Mandelbrot BB (1982) *The Fractal Geometry of Nature*. W.H. Freeman, San Francisco
138. Marsan D (2005) The role of small earthquakes in redistributing crustal elastic stress. *Geophys J Int* 163(1):141–151. doi:10.1111/j.1365-246X.2005.02700.x
139. May RM (1976) Simple mathematical models with very complicated dynamics. *Nature* 261:459–467
140. Mega MS, Allegrini P, Grigolini P, Latora V, Palatella L, Rapisarda A, Vinciguerra S (2003) Power law time distributions of large earthquakes. *Phys Rev Lett* 90:18850
141. Michael AJ, Jones LM (1998) Seismicity alert probabilities at Parkfield, California, revisited. *Bull Seismol Soc Am* 88(1):117–130
142. Miltenberger P, Sornette D, Vanneste C (1993) Fault self-organization as optimal random paths selected by critical spatiotemporal dynamics of earthquakes. *Phys Rev Lett* 71:3604–3607. doi:10.1103/PhysRevLett.71.3604
143. Mitzenmacher M (2004) A Brief History of Generative Models for Power Law and Lognormal Distributions. *Internet Math* 1(2):226–251
144. Mogi K (1969) Some features of recent seismic activity in and near Japan 2: activity before and after great earthquakes. *Bull Eq Res Inst Tokyo Univ* 47:395–417
145. Molchan G (2005) Intervent time distribution in seismicity: A theoretical approach. *Pure Appl Geophys* 162:1135–1150. doi:10.1007/s00024-004-2664-5
146. Molchan G, Kronrod T (2005) On the spatial scaling of seismicity rate. *Geophys J Int* 162(3):899–909. doi:10.1111/j.1365-246X.2005.02693.x
147. *Nature Debates* (1999) Nature debates: Is the reliable prediction of individual earthquakes a realistic scientific goal? available from [http://www.nature.com/nature/debates/earthquake/quake\\_frameset.html](http://www.nature.com/nature/debates/earthquake/quake_frameset.html)
148. Newman MEJ (2003) The structure and function of complex networks. *SIAM Rev* 45(2):167–256. doi:10.1137/S003614450342480
149. Ogata Y (1988) Statistical models for earthquake occurrence and residual analysis for point processes. *J Am Stat Assoc* 83:9–27
150. Ogata Y (1998) Space-time point-process models for earthquake occurrences. *Ann Inst Stat Math* 5(2):379–402
151. Olami Z, Feder HJS, Christensen K (1992) Self-organized criticality in a continuous, nonconservative cellular automaton modeling earthquakes. *Phys Rev Lett* 68(8):1244–1247
152. Osorio I, Frei MG, Sornette D, Milton J, Lai Y-C (2007) Seizures and earthquakes: Universality and scaling of critical far from equilibrium systems. submitted to *Phys Rev Lett*. <http://arxiv.org/abs/0712.3929>
153. Ouillon G, Sornette D (2000) The critical earthquake concept applied to mine rockbursts with time-to-failure analysis. *Geophys J Int* 143:454–468
154. Ouillon G, Sornette D (2004) Search for Direct Stress Correlation Signatures of the Critical Earthquake Model. *Geophys J Int* 157:1233–1246
155. Ouillon G, Sornette D (2005) Magnitude-Dependent Omori Law: Theory and Empirical Study. *J Geophys Res* 110:B04306. doi:10.1029/2004JB003311
156. Ouillon G, Sornette D, Castaing C (1995) Organization of joints and faults from 1 cm to 100 km scales revealed by Optimized Anisotropic Wavelet Coefficient Method and Multifractal analysis. *Nonlinear Process Geophys* 2:158–177
157. Ouillon G, Castaing C, Sornette D (1996) Hierarchical scaling of faulting. *J Geophys Res* 101(B3):5477–5487
158. Ouillon G, Ribeiro E, Sornette D (2007) Multifractal Omori Law for Earthquake Triggering: New Tests on the California, Japan and Worldwide Catalogs. submitted to *Geophys J Int*. <http://arxiv.org/abs/physics/0609179>
159. Ouillon G, Ducorbier C, Sornette D (2008) Automatic reconstruction of fault networks from seismicity catalogs: Three-dimensional optimal anisotropic dynamic clustering. *J Geophys Res* 113:B01306. doi:10.1029/2007JB005032
160. Peixoto TP, Prado CP (2004) Distribution of epicenters in the Olami–Feder–Christensen model. *Phys Rev E* 69(2):025101. doi:10.1103/PhysRevE.69.025101
161. Peixoto TP, Prado CPC (2006) Network of epicenters of the Olami–Feder–Christensen model of earthquakes. *Phys Rev E* 74(1):016, 126 doi:10.1103/PhysRevE.74.016126
162. Peters O, Neelin JD (2006) Critical phenomena in atmospheric precipitation. *Nature Phys* 2:393–396. doi:10.1038/nphys314
163. Pruessner G (2004) Studies in self-organized criticality, Ph D thesis, Imperial College London, available from [http://www.ma.imperial.ac.uk/%7Epruess/publications/thesis\\_final/](http://www.ma.imperial.ac.uk/%7Epruess/publications/thesis_final/)
164. Raleigh CB, Sieh K, Sykes LR, Anderson DL (1982) Forecasting Southern California Earthquakes. *Science* 217:1097–1104
165. Reynolds PJ, Klein W, Stanley HE (1977) Renormalization Group for Site and Bond Percolation. *J Phys C* 10:L167–L172
166. Rhoades DA, Evison FF (2004) Long-range earthquake forecasting with every earthquake a precursor according to scale. *Pure Appl Geophys* 161:47–72
167. Rhoades DA, Evison FF (2005) Test of the EEPAS forecasting model on the Japan earthquake catalogue. *Pure Appl Geophys* 162:1271–1290
168. Rice JR (1993) Spatio-temporal complexity of slip on a fault. *J Geophys Res* 98:9885–9907
169. Rundle JB, Klein W (1993) Scaling and critical phenomena in a cellular automaton slider block model for earthquakes. *J Stat Phys* 72:405–412



170. Rundle JB, Klein W (1995) New ideas about the physics of earthquakes. *Rev Geophys* 33:283–286
171. Rundle PB, Rundle JB, Tiampo KF, Sa Martins JS, McGinnis S, Klein W (2001) Nonlinear network dynamics on earthquake fault systems. *Phys Rev Lett* 87(14):148, 501. doi:10.1103/PhysRevLett.87.148501
172. Rundle JB, Turcotte DL, Shcherbakov R, Klein W, Sammis C (2003) Statistical physics approach to understanding the multiscale dynamics of earthquake fault systems. *Rev Geophys* 41(4):1019
173. Saichev A, Sornette D (2005) Distribution of the Largest Aftershocks in Branching Models of Triggered Seismicity: Theory of the Universal Bath's law. *Phys Rev E* 71:056127
174. Saichev A, Sornette D (2005) Vere-Jones' self-similar branching model. *Phys Rev E* 72:056, 122
175. Saichev A, Sornette D (2006) Renormalization of branching models of triggered seismicity from total to observable seismicity. *Eur Phys J B* 51:443–459
176. Saichev A, Sornette D (2006) "Universal" distribution of interearthquake times explained. *Phys Rev Lett* 97:078, 501
177. Saichev A, Sornette D (2007). Theory of earthquake recurrence times. *J Geophys Res* 112:B04313. doi:10.1029/2006JB004536
178. Saleur H, Sammis CG, Sornette D (1996) Renormalization group theory of earthquakes. *Nonlinear Process Geophys* 3:102–109
179. Saleur H, Sammis CG, Sornette D (1996) Discrete scale invariance, complex fractal dimensions and log-periodic corrections in earthquakes. *J Geophys Res* 101:17661–17677
180. Sammis SG, Sornette D (2002) Positive Feedback, Memory and the Predictability of Earthquakes. *Proc Nat Acad Sci USA* 99:SUPP1:2501–2508
181. Scholz CH (1991) Earthquakes and faulting: Self-organized critical phenomena with a characteristic dimension. In: Riste T, Sherrington D (eds) *Spontaneous Formation of Space Time Structure and Criticality*. Kluwer, Norwell, pp 41–56
182. Scholz CH (2002) *The Mechanics of Earthquakes and Faulting*, 2nd edn, Cambridge University Press, Cambridge
183. Scholz CH, Mandelbrot BB (eds) (1989) *Fractals in Geophysics*. Birkhäuser, Basel
184. Schorlemmer D, Wiemer S, Wyss M (2005) Variations in earthquake-size distribution across different stress regimes. *Nature* 437:539–542. doi:10.1038/nature04094
185. Schwartz DP, Coppersmith KJ (1984) Fault behavior and characteristic earthquakes: examples from the Wasatch and San Andreas Fault Zones. *J Geophys Res* 89:5681–5698
186. Shaw BE (1993) Generalized Omori law for aftershocks and foreshocks from a simple dynamics. *Geophys Res Lett* 20:907–910
187. Shaw BE (1994) Complexity in a spatially uniform continuum fault model. *Geophys Res Lett* 21:1983–1986
188. Shaw BE (1995) Frictional weakening and slip complexity in earthquake faults. *J Geophys Res* 102:18239–18251
189. Shaw BE (1997) Model quakes in the two-dimensional wave equation. *J Geophys Res* 100:27367–27377
190. Shcherbakov R, Turcotte DL (2004) A modified form of Bath's law. *Bull Seismol Soc Am* 94(5):1968–1975
191. Shnirman MG, Blanter EM (1998) Self-organized criticality in a mixed hierarchical system. *Phys Rev Lett* 81:5445–5448
192. Smalley RF Jr, Turcotte DL, Solla SA (1985) A renormalization group approach to the stick-slip behavior of faults. *J Geophys Res* 90:1894–1900
193. Sornette A, Sornette D (1989) Self-organized criticality and earthquakes. *Europhys Lett* 9:197–202
194. Sornette A, Sornette D (1999) Earthquake rupture as a critical point: Consequences for telluric precursors. *Tectonophysics* 179:327–334
195. Sornette A, Davy P, Sornette D (1990) Growth of fractal fault patterns. *Phys Rev Lett* 65:2266–2269
196. Sornette A, Davy P, Sornette D (1990) Fault growth in brittle-ductile experiments and the mechanics of continental collisions. *J Geophys Res* 98:12111–12139
197. Sornette D (1991) Self-organized criticality in plate tectonics. In: *Proceedings of the NATO ASI. vol 349, "Spontaneous formation of space-time structures and criticality"* Geilo, Norway 2–12 April 1991. Riste T, Sherrington D (eds) Kluwer, Dordrecht, Boston, pp 57–106
198. Sornette D (1992) Critical phase transitions made self-organized: a dynamical system feedback mechanism for self-organized criticality. *J Phys I France* 2:2065–2073. doi:10.1051/jp1:1992267
199. Sornette D (1998) Discrete scale invariance and complex dimensions. *Phys Rep* 297(5):239–270
200. Sornette D (1999) Earthquakes: from chemical alteration to mechanical rupture. *Phys Rep* 313(5):238–292
201. Sornette D (2000) Mechanochemistry: an hypothesis for shallow earthquakes. In: Teisseyre R, Majewski E (eds) *Earthquake Thermodynamics and Phase Transformations in the Earth's Interior*. *Int Geophys Series*, vol 76. Cambridge University Press, Cambridge, pp 329–366, e-print at <http://xxx.lanl.gov/abs/cond-mat/9807400>
202. Sornette D (2002) Predictability of catastrophic events: material rupture, earthquakes, turbulence, financial crashes and human birth. *Proc Nat Acad Sci USA* 99:2522–2529
203. Sornette D (2004) *Critical Phenomena in Natural Sciences: Chaos, Fractals, Selforganization and Disorder: Concepts and Tools*, 2nd edn. Springer, Berlin, p 529
204. Sornette D, Helmstetter A (2002) Occurrence of Finite-Time-Singularity in Epidemic Models of Rupture, Earthquakes and Starquakes. *Phys Rev Lett* 89(15):158501
205. Sornette D, Ouillon G (2005) Multifractal Scaling of Thermally-Activated Rupture Processes. *Phys Rev Lett* 94:038501
206. Sornette D, Pisarenko VF (2003) Fractal Plate Tectonics. *Geophys Res Lett* 30(3):1105. doi:10.1029/2002GL015043
207. Sornette D, Sammis CG (1995) Complex critical exponents from renormalization group theory of earthquakes: Implications for earthquake predictions. *J Phys I France* 5:607–619
208. Sornette D, Virieux J (1992) A theory linking large time tectonics and short time deformations of the lithosphere. *Nature* 357:401–403
209. Sornette D, Werner MJ (2005) Constraints on the size of the smallest triggering earthquake from the epidemic-type aftershock sequence model, Bath's law, and observed aftershock sequences. *J Geophys Res* 110:B08304. doi:10.1029/2004JB003535
210. Sornette D, Werner MJ (2005) Apparent clustering and apparent background earthquakes biased by undetected seismicity. *J Geophys Res* 110:B09303. doi:10.1029/2005JB003621
211. Sornette D, Davy P, Sornette A (1990) Structuration of the lithosphere in plate tectonics as a self-organized critical phenomenon. *J Geophys Res* 95:17353–17361

212. Sornette D, Vanneste C, Sornette A (1991) Dispersion of b-values in Gutenberg–Richter law as a consequence of a proposed fractal nature of continental faulting. *Geophys Res Lett* 18:897–900
213. Sornette D, Miltenberger P, Vanneste C (1994) Statistical physics of fault patterns self-organized by repeated earthquakes. *Pure Appl Geophys* 142:491–527. doi:10.1007/BF00876052
214. Sornette D, Miltenberger P, Vanneste C (1995) Statistical physics of fault patterns self-organized by repeated earthquakes: synchronization versus self-organized criticality. In: Bouwknecht P, Fendley P, Minahan J, Nemeschansky D, Pilch K, Saleur H, Warner N (eds) *Recent Progresses in Statistical Mechanics and Quantum Field Theory*. Proceedings of the conference ‘Statistical Mechanics and Quantum Field Theory’, USC, Los Angeles, May 16–21, 1994. World Scientific, Singapore, pp 313–332
215. Sornette D, Utkin S, Saichev A (2008) Solution of the Nonlinear Theory and Tests of Earthquake Recurrence Times. *Phys Rev E* 77:066109
216. Stanley HE (1999) Scaling, universality, and renormalization: Three pillars of modern critical phenomena. *Rev Mod Phys* 71(2):S358–S366. doi:10.1103/RevModPhys.71.S358
217. Sykes LR, Jaumé S (1990) Seismic activity on neighboring faults as a long-term precursor to large earthquakes in the San Francisco Bay Area. *Nature* 348:595–599
218. Tiampo KF, Rundle JB, Klein W (2006) Stress shadows determined from a phase dynamical measure of historic seismicity. *Pure Appl Geophys* 163(11–12):2407–2416
219. Turcotte DL (1986) Fractals and fragmentation. *J Geophys Res* 91:1921–1926
220. Turcotte DL (1997) *Fractals and Chaos in Geology and Geophysics*, 2nd edn. Cambridge University Press, Cambridge, p 398
221. Turcotte DL, Newman WI, Gabrielov A (2000) A statistical physics approach to earthquakes. In: Rundle JB, Turcotte DL, Klein W (eds) *GeoComplexity and the Physics of Earthquake*. American Geophysical Union, Washington, pp 83–96
222. Tumarkin AG, Shnirman MG (1992) Computational seismology 25:63–71
223. Vere-Jones D (1977) Statistical theories of crack propagation. *Math Geol* 9:455–481
224. Vere-Jones D (2005) A class of self-similar random measure. *Adv Appl Probab* 37(4):908–914
225. Vere-Jones D (2006) The development of statistical seismology: A personal experience. *Tectonophysics* 413(1–2):5–12
226. Vere-Jones D, Robinson R, Yang W (2001) Remarks on the accelerated moment release model: problems of model formulation, simulation and estimation. *Geophys J Int* 144:517–531. doi:10.1046/j.1365-246X.2001.01348.x
227. Voight B (1988) A method for prediction of volcanic eruptions. *Nature* 332:125–130
228. Voight B (1989) A relation to describe rate-dependent material failure. *Science* 243:200–203
229. Werner MJ (2007) On the fluctuations of seismicity and uncertainties in earthquake catalogs: Implications and methods for hypothesis testing. Ph D thesis, University of California, Los Angeles
230. Werner MJ, Sornette D (2007) Comment on “Analysis of the Spatial Distribution Between Successive Earthquakes” by Davidsen and Paczuski. [*Phys Rev Lett* 94:048501 (2005)]. *Phys Rev Lett* 99:179801
231. Werner MJ, Sornette D (2008) Magnitude Uncertainties Impact Seismic Rate Estimates, Forecasts and Predictability Experiments. *J Geophys Res* 113:B08302. doi:10.1029/2007JB005427
232. Wesnousky SG (1994) The Gutenberg–Richter or characteristic earthquake distribution, which is it? *Bull Seismol Soc Am* 84(6):1940–1959
233. Wiemer S, Katsumata K (1999) Spatial variability of seismicity parameters in aftershock zones. *J Geophys Res* 104:13135–13152. doi:10.1029/1999JB900032
234. Wilson K (1979) Problems in physics with many scales of length. *Sci Am* 241:140–157
235. Yeomans JM (1992) *Statistical Mechanics of Phase Transitions*. Oxford University Press Inc, New York
236. Zaliapin I, Keilis-Borok V, Ghil M (2003) A Boolean delay equation model of colliding cascades. Part I: Multiple seismic regimes. *J Stat Phys* 111:815–837
237. Zaliapin I, Keilis-Borok V, Ghil M (2003) A Boolean delay equation model of colliding cascades. Part II: Prediction of critical transitions. *J Stat Phys* 111:839–861
238. Zaliapin I, Gabrielov A, Keilis-Borok V, Wong H (2008) Clustering analysis of seismicity and aftershock identification. *Phys Rev Lett* 101:018501. doi:10.1103/PhysRevLett.101.018501
239. Zee A (2003) *Quantum Field Theory in a Nutshell*. Princeton University Press, Princeton
240. Zhuang J, Ogata Y, Vere-Jones D (2002) Stochastic declustering of space-time earthquake occurrences. *J Am Stat Assoc* 97:369–380
241. Zhuang J, Ogata Y, Vere-Jones D (2004) Analyzing earthquake clustering features by using stochastic reconstruction. *J Geophys Res* 109:B05301. doi:10.1029/2003JB002879
242. Zöller G, Hainzl S (2002) A systematic spatiotemporal test of the critical point hypothesis for large earthquakes. *Geophys Res Lett* 29:53–1
243. Zöller G, Hainzl S, Kurths J (2001) Observation of growing correlation length as an indicator for critical point behavior prior to large earthquakes. *J Geophys Res* 106:2167–2176. doi:10.1029/2000JB900379

# Single Column Modeling of Atmospheric Boundary Layers and the Complex Interactions with the Land Surface

ALBERT A. M. HOLTSLAG, GERT-JAN STEENEVELD  
Department of Meteorology and Air Quality,  
Wageningen University, Wageningen, The Netherlands

## Article Outline

Glossary  
Definition of the Subject  
Introduction  
Background  
Atmospheric Boundary-Layer Structure  
Modeling Basics  
Local Mixing Parameterization  
More Advanced Mixing Parameterizations  
Intercomparison of Single Column Models for Stable Conditions  
Modeling Boundary Layers over Land  
Impact of Land Surface Conditions on Model Results  
Summary  
Acknowledgments  
Bibliography

## Glossary

**Atmospheric boundary layer** The Atmospheric Boundary Layer (ABL) is the lower part of the atmosphere which is directly influenced by the presence of the earth's surface. As such its major characteristics are turbulence and the diurnal cycle.

**Diurnal cycle** The depth of the dry atmospheric boundary layer (ABL) can vary over land between tens of meters during night up to kilometers during daytime (see Fig. 2). Over sea the depth is often typical a few hundred meters and rather constant on the time scale of a day.

**Turbulence** Turbulence in the atmospheric boundary layer is the three-dimensional, chaotic flow of air with time scales typically between a second and an hour. The corresponding length scales are from a millimeter up to the depth of the boundary layer (or more in the case of clouds). Turbulence in the ABL originates due to friction of the flow and heating (convection) at the surface.

## Definition of the Subject

In this article we deal with the single column modeling of the Atmospheric Boundary layer (ABL) and the complex interactions which may occur with the land surface. As such we review the major characteristics of the ABL over land, and summarize the basic parameterizations for the representation of atmospheric turbulence and the surface fluxes. The modeling principles are illustrated with the outcome of single-column models for a variety of conditions using field data and fine-scale model results. Our emphasis is on stable conditions which occur over land at night-time under clear skies. For readers not familiar with atmospheric turbulence and meteorological definitions, some background and basic definitions are also given.

## Introduction

The Atmospheric Boundary Layer (ABL) is generally characterized by turbulence. Because of its capability to mix air with different properties efficiently, the representation of turbulence is directly relevant for atmospheric and environmental modeling. For instance, turbulence directly impacts on the transfer of momentum, sensible heat, water vapor, ozone, and methane, among many other quantities, between the earth's surface and the atmosphere. Turbulence also defines the mixing of properties inside the atmospheric boundary layer, the transfer of quantities between the boundary layer and the clear or cloudy atmosphere aloft, and the mixing inside clouds.

Turbulence in the ABL is mainly due to the mechanical turbulence by vertical wind shear and turbulence by convection. Most of the atmosphere above the ABL is not turbulent, although turbulence can occur throughout the whole atmosphere. For instance, cumulus-type clouds, which may grow into thunderstorms, are always turbulent through convection produced by the heat released due to the condensation of water vapor. Turbulence can also occur in clear air above the ABL; most of this is produced in layers of strong vertical wind shear at the boundary between air masses (so-called 'Clear-Air Turbulence').

Because of the mixing capacity of turbulence, modeling atmospheric boundary layers is also relevant for many practical applications. For instance, chimney plumes are diluted and spread over larger volumes than they would be without turbulence. As such, strong local peaks of pollution are prevented and otherwise clean air is polluted. In practice turbulence may also cause engineering problems, because it shakes structures such as bridges, towers, and airplanes, causing failure of such systems in extreme cases. Turbulent fluctuations in the horizontal motions during

severe storms can be fatal to tall buildings or bridges, particularly if resonance (e. g., forcing of a system at its natural frequency) occurs.

The correct formulation of the overall effects by turbulence, either inside or outside the atmospheric boundary layer, is an essential part of atmospheric models dealing with the prediction and study of weather, climate and air quality. These models are based on solving the equations dealing with atmosphere behavior. With state-of-the-art computers, the number of grid points in atmospheric models is limited to a number of typically  $10^8$  or so. This implies that on the regional and global scale the atmospheric model equations are usually applied too fairly large ‘air boxes’. Such boxes are often in the order of ten to hundred kilometers wide and ten to a few hundred meters thick. In these large boxes, smaller scale motions make air parcels interact and mix. For example, if a hot parcel is located next to a cold parcel, turbulent motion at their boundaries will heat the cool and cool the hot parcel. Thus, a closure formulation is needed to reproduce mixing by the turbulent motions into the model-resolved scales using the equations for the larger-scale ‘mean’ motions. It is important to realize that the closure formulation needs to be expressed in terms of variables available in the modeling context. This is called a ‘parameterization’.

In this contribution we provide an overview of the modeling principles, the turbulent closures and parameterizations in use for of the atmospheric boundary layer, where we emphasize the modeling and parameterization of turbulence in the atmospheric boundary layer without clouds. Additionally we discuss the performance of models in current use [7,28], and we study the impact of the surface boundary condition over land [18].

## Background

Atmospheric models for the forecasting and study of weather, climate, and air quality are typically based on integration of the basic equations governing atmospheric behavior. These equations are the gas law, the equation of continuity (mass), the first law of thermodynamics (heat), the conservation equations for momentum (the so-called “Navier–Stokes equations”), and usually equations expressing the conservation of moisture, trace gases and air pollutants. At one extreme, atmospheric models may deal with the world’s climate and climate change; at the other, they may account for the behavior of local flows at coasts, in mountain-valley areas, or even deal with individual clouds. This all depends on the selected horizontal modeling domain and the available computing resources.

Since there is an enormous range of scales in atmospheric motion and turbulence, there is a need to separate the scales of atmospheric turbulence from larger-scale motions. Let  $C$  denote an atmospheric variable, such as specific humidity. Then  $\bar{C}$  represents a mean or “smoothed” value of  $C$ , typically taken on a horizontal scale of order 10 (or more) km and a corresponding time scale in the order of 10 min to one hour. A local or instantaneous value of  $C$  would differ from  $\bar{C}$ . Thus, we have

$$C = \bar{C} + c. \quad (1)$$

Here  $c$  represents the smaller-scale fluctuations. Note that we use lower case for the latter (often primes are used as well to indicate fluctuations). In principle, the fluctuations around the mean motion also reflect gravity waves and other smaller scale motions, in addition to turbulence. Gravity waves often co-exist with turbulence or are generated by turbulence. If the wind at the same time is weak, there may be no turbulence at all. Anyhow, if turbulence exists, it is usually more important for most atmospheric applications, because it mixes more efficiently than the other small-scale motions.

To make the mathematical handling of  $c$  tractable, it must satisfy the so-called “Reynolds postulates”. These require, for example, that  $\bar{c} = 0$  and that small- and larger-scale values must not be correlated. After a quantity has been averaged to create a larger-scale quantity, further averaging should produce no further changes, in order for this postulate to apply. The mean of the summation of two variables  $A$  and  $C$  will produce  $\overline{A \pm C} = \bar{A} \pm \bar{C}$ . A further condition is that a mean variable  $\bar{C}$  must be differentiable, since differentials show up in the atmospheric equations (see below). In practice, not all these conditions are rigorously satisfied. If the Reynolds postulates are fulfilled, then the averaging for the product of two variables provides

$$\overline{AC} = \bar{A}\bar{C} + \overline{ac}. \quad (2)$$

The second term at the right hand side of Eq. (2) is known as the turbulent covariance. Similarly, the turbulence variance of a quantity is given by  $\overline{C^2} - (\bar{C})^2$  (which is the square of the standard deviation).

If in Eq. (2), the variable  $A$  represents one of the velocity components ( $U, V, W$  in the  $x, y, z$  direction, respectively), then  $\bar{AC}$  is the total flux of  $C$  and the second term at the right hand side of Eq. (2) represents a turbulent flux of  $C$ . For instance,  $\overline{uc}$  and  $\overline{wc}$  are the horizontal and vertical turbulent fluxes of the variable  $C$ , respectively. Here  $u$  and  $w$  are the turbulent fluctuations of the horizontal and vertical velocities. Near the surface, the mean vertical wind  $\bar{W}$  is usually small, and thus the total vertical fluxes are normally dominated by the turbulent contributions.

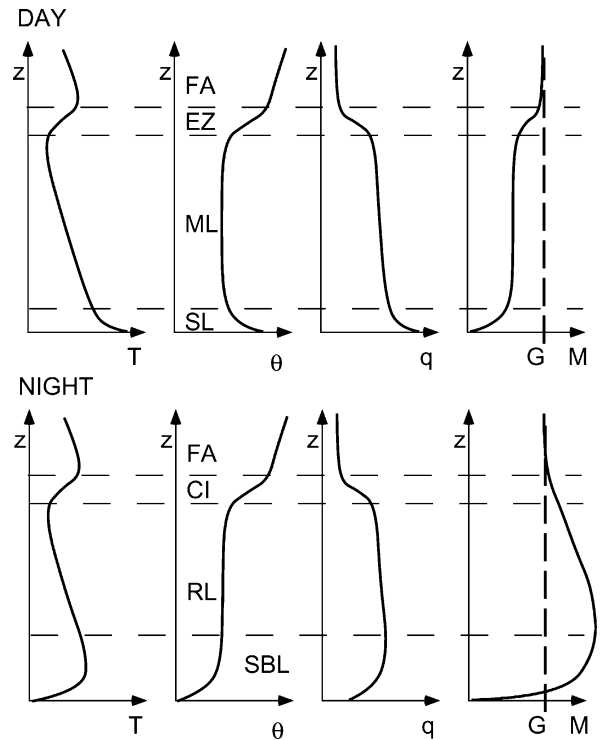
## Atmospheric Boundary-Layer Structure

Turbulent fluctuations, variances and fluxes of variables are influenced by the vertical boundary-layer structure. Here the variation of temperature in the atmospheric boundary-layer plays an important role. Since pressure decreases with altitude, air parcels, which are forced to rise (sink), do expand (compress). According to the first law of thermodynamics, a rising (sinking) parcel will cool (warm) if there is no additional energy source such as condensation of water vapor. Then this is called a dry adiabatic process.

It can be shown that in the atmospheric boundary layer, the temperature ( $T$ ) variation with height for a dry adiabatic process is  $dT/dz = -g/C_p$  (here  $g$  is gravity constant and  $C_p$  is specific heat at constant pressure). The value for  $g/C_p$  is approximately 1 K per 100 m. An atmospheric layer which has such a temperature variation with height, is called neutral for dry air (at least when there is no convection arising from other levels). In that case  $\Theta = T + (g/C_p)z$  is constant, where  $\Theta$  is called the potential temperature (Note that the previous definition for potential temperature is not accurate above the boundary layer). Since air normally contains water vapor and because moist air is lighter than dry air, we have to correct for the influence of this on vertical motions. Consequently, a virtual potential temperature is defined as  $\Theta_v = \Theta(1 + 0.61q)$ , where  $q$  is the specific humidity (defined as the mass of water vapor per unit mass of moist air).

In a neutral layer with constant  $\Theta_v$ , vertical motions of moist (not saturated) air can maintain themselves. If the virtual potential temperature of the atmospheric layer increases with height, vertical displacements are suppressed. This is called a stable condition (or ‘inversion’). At the other hand, when the virtual potential temperature decreases with height, vertical fluctuations may be accelerated. Consequently this is called an unstable condition. Thus in considerations with turbulent fluctuations and atmospheric stability, we have to deal with the virtual potential temperature and not with the actual temperature. Similarly, the vertical flux of sensible heat is connected to turbulent fluctuations of (virtual) potential temperature; e. g. it reads as  $w\theta_v$  (in m K/s). The latter relates directly to the energy per time and unit area  $H$  by  $H = \rho C_p w\theta_v$  (in W/m<sup>2</sup>), where  $\rho$  is density of the air (in kg/m<sup>3</sup>).

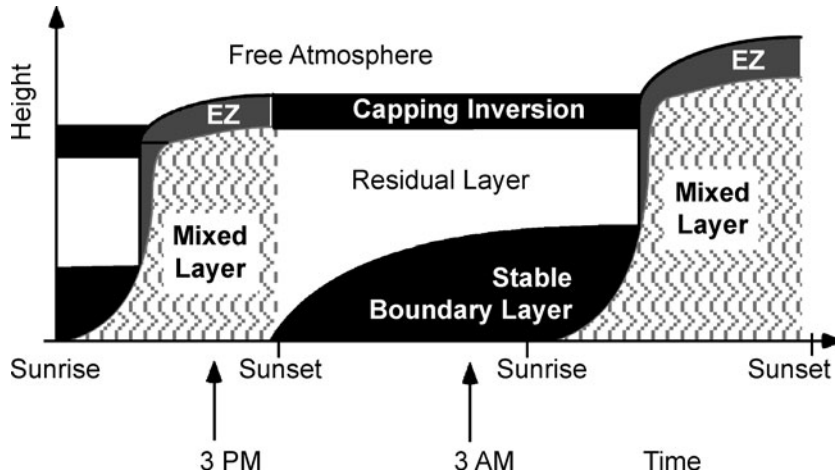
Figure 1 (after [30]), provides the typical, idealized, mean vertical profiles for temperature  $T$ , potential temperature  $\Theta$ , specific humidity  $q$ , in addition to the horizontal wind  $M$  (defined by  $M^2 = U^2 + V^2$ ). These profiles apply for an atmospheric boundary layer over land in clear sky conditions in the afternoon and around midnight. Note



Single Column Modeling of Atmospheric Boundary Layers and the Complex Interactions with the Land Surface, Figure 1  
Idealized vertical profiles of mean variables in the Atmospheric Boundary Layer over land in fair weather (after [30]). See text for additional information

that in the free atmosphere the horizontal wind is mostly a result of the acting of the larger scale pressure differences and the Coriolis force due to the rotation of the earth (but other effects may play a role as well). The resulting wind is known as the ‘geostrophic’ wind and indicated with  $G$  in Fig. 1 (see dashed line). In the daytime boundary layer the actual wind is smaller due to surface friction, while at clear nights the actual wind away from the surface may be substantially stronger than  $G$  due to inertial effects (resulting in the so-called ‘low level jet’).

The temporal variation of the mean boundary-layer profiles over land can be quite substantial due to the strong diurnal variation of solar incoming radiation and the nighttime cooling at the land surface. During daytime the turbulent boundary layer may grow to several kilometers into the non-turbulent ‘free atmosphere’ (indicated as FA in Fig. 1). At night the turbulent part of the stable boundary layer (SBL) may only extend up to a few hundred meters or less (the lowest dashed line in the lower figure). An idealized picture for the temporal variation of the boundary layer over land is given in Fig. 2 (after [30]).



Single Column Modeling of Atmospheric Boundary Layers and the Complex Interactions with the Land Surface, Figure 2  
Idealized diurnal evolution of the Atmospheric Boundary Layer over land in fair weather (after [30])

Here the arrows with local time indications refer to the day and nighttime figures of Fig. 1.

Figure 1 also indicates that the boundary layer during daytime shows a three-layer structure: an unstable ‘surface layer (SL)’, a ‘well-mixed layer (ML)’ with rather uniform (virtual) potential temperature, and a stably-stratified ‘entrainment zone (EZ)’. In the latter zone, turbulence acts to exchange heat, momentum, water-vapor and trace gasses between the boundary layer and the free atmosphere. During nighttime, often the vertical structure of the previous day persists above the SBL. As such a ‘residual layer (RL)’ with sporadic turbulence (remaining from the previous day) can be identified as well as a ‘capping inversion (CI)’.

### Modeling Basics

The challenge of modeling the atmospheric boundary layer is the prediction of the temporal variation of the vertical and horizontal structures in response to the influence of the major processes acting in the atmosphere and at the earth’s surface. As such the governing equations have to be integrated. In practice, the variables are split into ‘mean’ larger-scale motions and smaller-scale fluctuations as in Eq. (1). Inserting this into the basic equations and after averaging this provides a set of equations for the behavior of the larger-scale (mean) variables. The larger-scale variables are then used explicitly in atmospheric models. This can be demonstrated as follows below.

The general character of any of the budget equations dealing with atmospheric motions is

$$\frac{DC}{Dt} = S_i . \quad (3a)$$

Here  $S_i$  represents the subsequent sources and sinks for the variable  $C$  (such as radiation or chemistry effects). The notation  $DC/Dt$  represents the total rate of change for the variable  $C$  by local changes ( $\partial/\partial t$ ), and changes transported with the fluid motion in the three directions. As such, we have

$$\frac{\partial C}{\partial t} + U \frac{\partial C}{\partial x} + V \frac{\partial C}{\partial y} + W \frac{\partial C}{\partial z} = S_i . \quad (3b)$$

Here  $U, V, W$  are the wind speeds in the three directions  $x, y, z$ , respectively.

If in the atmospheric motion each variable is split into a mean component and a fluctuation then (3b) provides after Reynolds-averaging, some algebraic manipulations and simplifying assumptions, a budget equation for the mean variable  $\bar{C}$ . This reads as

$$\begin{aligned} \frac{D\bar{C}}{Dt} &= \frac{\partial \bar{C}}{\partial t} + \bar{U} \frac{\partial \bar{C}}{\partial x} + \bar{V} \frac{\partial \bar{C}}{\partial y} + \bar{W} \frac{\partial \bar{C}}{\partial z} \\ &= \bar{S}_i - \frac{\partial \overline{u\bar{c}}}{\partial x} - \frac{\partial \overline{v\bar{c}}}{\partial y} - \frac{\partial \overline{w\bar{c}}}{\partial z} . \end{aligned} \quad (4)$$

We may note that in the derivation of (4), single terms representing fluctuations have disappeared (as above in Eq. (2)). However, terms involving the product of two fluctuations did remain.

Thus because the basic equations are nonlinear, the budget equations for the mean variables contain terms involving smaller-scale motions. The latter terms are of the form of a divergence of fluxes produced by such motions in the three directions and appear as the last three terms in Eq. (4). These motions are said to be sub-grid and consequently, closure formulations or parameterizations are

needed to introduce mixing by the smaller-scale, sub-grid, motions into the equations for the larger-scale motions (as resolved by the model). Note that additional terms may also appear in (4) when the source or sink term  $S_i$  incorporates nonlinear effects (such as in the case of chemistry).

The atmospheric model equations can also be applied on much smaller spatial and temporal scales than discussed here, for instance by using vertical and horizontal grid elements of 10 to 100 m, and time steps of seconds only. It is important to realize that in such cases a significant part of the turbulent fluctuations are resolved by the model equations. This type of modeling is nowadays known as ‘Large-eddy simulation (LES)’. This has become a powerful and popular tool in the last decade to study turbulence in clear and cloudy boundary layers under well-defined conditions. It is important to realize that in the case of LES the simplifying assumptions leading to Eq. (2) are normally not valid.

A special and simple form of Eq. (4) arises for horizontally homogeneous conditions. In such cases the terms including horizontal derivatives are negligible. If in addition the mean vertical wind is small and if there are no other sources and sinks, then (4) provides

$$\frac{\partial \bar{C}}{\partial t} = -\frac{\partial \bar{w}c}{\partial z}. \quad (5)$$

This equation is known as the one-dimensional, vertical diffusion equation. It shows that the local time rate of change for the mean of a variable (such as temperature or wind) at a certain height, is given by the divergence of the turbulent (corresponding heat or momentum) flux in the vertical direction. As such, information on the turbulent flux may produce a local forecast of the variation of a mean variable (but only under the simplifications mentioned).

Equation (5) can be seen as the basis of a single-column model where only local information of the atmosphere is relevant. However, normally the other terms in (4) are also relevant, in particular the terms with mean wind speed (the so-called ‘‘advection terms’’). This means that in general the budget equations for momentum, heat, and the various scalars are closely coupled in any atmospheric model. Still one can solve for the local time rate of change in a single column model once the advection terms are known from observations or other means. This approach is widely adopted to study atmospheric boundary layers in comparison with observations on a local scale.

Before we proceed with more detailed parameterizations for the fluxes in the boundary layer, let us deal with the derivation of the surface fluxes. These fluxes enter as boundary conditions when solving the budget equations for all the relevant mean variables (in any approach). It

is important to realize that near the surface, the average wind must vanish because the mean wind is zero at the earth’s surface. At the other hand, we know from observations that the fluxes of heat, momentum and trace gasses are nonzero. Consequently, it is convenient to model an ‘effective’ surface flux  $\bar{w}c_0$  of a conserved variable due to the combined effect of molecular diffusion and turbulence at the surface. This can be achieved by writing

$$\bar{w}c_0 = \beta_t w_t (C_0 - C_a). \quad (6)$$

Here  $C_0$ , and  $C_a$  are the values of the transported variable at the surface and in the air, respectively;  $\beta_t$  is a transfer coefficient, and  $w_t$  is an effective transport velocity representing the turbulence. For example, in near-neutral conditions the effective transport velocity is well represented by the well-known surface friction velocity  $u_{*0}$ . Then it can be shown that  $\beta_t = \kappa / \ln(z/z_0)$ , where  $\kappa$  is the ‘Von Karman’ constant (often specified as  $\kappa \cong 0.4$ ),  $z$  is the corresponding height of  $C_a$  in the lowest part of the boundary layer and  $z_0$  is the so-called surface roughness length for the variable  $C$ . We refer to the literature for a more detailed treatment (e. g., Beljaars and Holtslag [4]).

### Local Mixing Parameterization

To solve the budget Eq. (4) for all the mean atmospheric variables involved, the terms involving turbulent fluxes need to be parametrized. As mentioned before, this means that the fluxes need to be expressed in terms of available mean model quantities, both in the atmosphere and at the surface. Once this has been achieved, the atmospheric model equations can be integrated. Thus, starting with proper initial values, new values can be calculated for the following time step and so on.

The most frequently used parameterization for environmental and atmospheric models, is known as first-order closure or often also called  $K$ -theory. In this theory it is assumed that the flux  $\bar{w}c$  of a variable  $C$  in the vertical direction  $z$ , is down the vertical gradient of the mean concentration of  $C$  per unit mass. Thus

$$\bar{w}c = -K_c \frac{\partial \bar{C}}{\partial z}. \quad (7)$$

Here,  $K_c$  is known as the ‘eddy-diffusivity’ or mixing coefficient for the variable  $C$ . Similarly, the horizontal fluxes can be represented in terms of horizontal gradients. Note that the corresponding eddy-diffusivities typically are not constant, but that they generally depend on properties of the flow and the variable of interest. This also means that normally no analytic solutions are possible, not even for the simple case in which Eqs. (5) and (8) are combined.

We may note that the dimension of an eddy-diffusivity is a length scale  $\ell$  times a velocity scale. These are proportional to the products of effective eddy sizes and eddy velocities in the corresponding directions. Often a diagnostic expression is used for the eddy-diffusivity, on basis of what is called ‘mixing length theory’ (in analogy with molecular diffusion). The result reads as

$$K_c = \ell^2 S f(\text{Ri}) . \quad (8)$$

Here  $S$  is vertical wind shear (that is the variation of mean horizontal wind with height). Note that the combination  $\ell S$  in Eq. (9) has units of velocity. In Eq. (8),  $f(\text{Ri})$  denotes a functional dependence on local stability as represented by the gradient Richardson-number  $\text{Ri}$  defined by

$$\text{Ri} = \frac{g}{\Theta_v} \frac{\partial \overline{\Theta}_v / \partial z}{(\partial \overline{U} / \partial z)^2 + (\partial \overline{V} / \partial z)^2} . \quad (9)$$

Here  $g$  is the acceleration due to gravity, and  $\overline{\Theta}_v$  is the mean ‘virtual potential temperature’.

The specification of the length scale  $\ell$  is not at all straightforward, except near the surface where so-called ‘surface-layer similarity theory’ (see cited literature) provides that  $\ell \propto z$ . A frequently used form for  $\ell$  is:

$$\frac{1}{\ell} = \frac{1}{\kappa z} + \frac{1}{\lambda} . \quad (10)$$

Here  $\lambda$  is a turbulent length scale, which should be valid for the turbulence far above the surface. We note that the latter has a rather empirical nature and consequently there is no agreement on the specification of  $\lambda$  in the literature.

Equation (8) is a diagnostic equation, which indicates that the eddy-diffusivity may vary with height, wind speed, stability, et cetera. In combination with the flux parameterization of Eq. (7), it follows that the flux at a certain height depends on the local gradient of the mean variable involved. Consequently the approach is referred to as a ‘diagnostic local mixing approach’. Such an approach is mostly suitable for relatively homogeneous conditions with neutral and stable stratification, and is not so suitable for cases with convection (see non-local mixing parameterizations below).

### More Advanced Mixing Parameterizations

A physically realistic alternative to the diagnostic approach is to relate the eddy-diffusivity of Eq. (7) to the actual turbulent kinetic energy of the flow, by using the prognostic turbulent kinetic energy equation and an appropriate choice for the turbulent length scale. It is important to realize that the kinetic energy of atmospheric motion per unit of mass  $E$  is given by the half of the sum of

the velocities squared in the three directions (as in classic mechanics), e.g.  $E = (U^2 + V^2 + W^2)/2$ . Similar as with respect to Eq. (2), we can separate between the Mean Kinetic Energy  $\overline{E}$  of the mean atmospheric motions and the Turbulent Kinetic Energy (TKE or  $e$ ) of the smaller-scale fluctuating motions by turbulence. Thus  $e$  is given by  $e = (\overline{u^2} + \overline{v^2} + \overline{w^2})/2$ .

The prognostic equation for  $e$  reads in its basic form as:

$$\frac{De}{Dt} = -\overline{uw} \frac{\partial \overline{U}}{\partial z} - \overline{vw} \frac{\partial \overline{V}}{\partial z} + \frac{g}{\Theta_v} \overline{w\theta_v} + D - \varepsilon . \quad (11)$$

Here  $De/Dt$  is the total variation of  $e$  with time (the sum of local variations and those transported with the mean air motion). The two terms at the immediate right hand side of (11) represent the shear production of turbulence. These depend primarily on vertical variations of wind or, near the ground, on wind speed and surface roughness. The terms are almost always positive. The third term in Eq. (11) represents the rate of production or breakdown of turbulence by buoyancy effects (such as heat convection). It depends directly on density effects, which can be written in terms of the virtual potential temperature  $\overline{\Theta}_v$ , and its turbulent flux  $\overline{w\theta_v}$ . The term  $D$  in Eq. (11) represents divergence and pressure redistribution terms. These have a tendency to cancel near the surface. Finally, the term  $\varepsilon$  reflects the molecular dissipation of turbulence into heat and this term is always positive. In fact  $\varepsilon$  is typically proportional to  $e/\tau$ , where  $\tau$  is the characteristic time scale for the turbulent mixing process.

Using Eq. (11), turbulent kinetic energy can be calculated for given mean profiles when the corresponding fluxes are calculated using Eq. (7) for all fluxes involved. In this approach the diffusivities are typically calculated with equations of the form

$$K_c = \alpha_c \ell \sqrt{e} . \quad (12)$$

Here  $\alpha_c$  is a constant depending on the variable of interest. The length scale is typically calculated with a similar type of diagnostic equation as (10) provides. This approach is known as the ‘TKE-length scale approach’ and it is an example of so-called 1.5 order closure. Sometimes a prognostic equation is used for the length scale as well, but such an approach is more popular in engineering applications than in the atmospheric sciences.

It can be shown that Eq. (8) is a solution of (11) and (12) in stationary conditions and when other simplifications are made such as the neglect of the influences by advection and turbulence divergence in the TKE equation. A more advanced turbulence scheme is known as ‘second-order closure’. In such an approach, prognostic



equations are developed for the fluxes and variances themselves. Such equations have a very similar structure as Eq. (12) for kinetic energy. Unfortunately, new unknowns are present in these equations. These must be related to the other variables in the model equations, always involving assumptions. Thus, second-order closure involves many more than the original equations and is therefore computationally more time consuming ('expensive') than first-order and 1.5 order closure.

One may expect that a model with 1.5 or second-order closure would produce more realistic results than a model with a first order closure. However, in practice this is often not the case, because of complex model interactions and the difficulty of representing all the relevant details with sufficient accuracy (see also below). That is the reason why diagnostic approaches remain popular. Nevertheless, second order equations are useful to gain insight in the governing physics, and after simplification useful extensions of the basic parameterizations may be achieved.

We continue our discussion with mixing parameterizations which have been proposed for boundary-layers with strong atmospheric convection. In such cases, the turbulent flux of a conserved quantity is typically not proportional to the local gradient alone as predicted by Eq. (7). In fact, in a large part of the ABL the mean gradients are small in conditions with dry convection, in particular for potential temperature (see Fig. 1). Then the fluxes depend mostly on the mixing characteristics of the large eddies across the ABL. Theories are available, which have modified  $K$ -theory to allow for the influence of convection, for example by including additional terms at the right hand side of Eq. (8) For details we refer to the literature (e. g., Holtslag and Moeng [15]).

In the next sections we apply the modeling concepts above and compare their results with field observations. In addition we present results from model intercomparison studies, and illustrate the role of boundary conditions.

### Intercomparison of Single Column Models for Stable Conditions

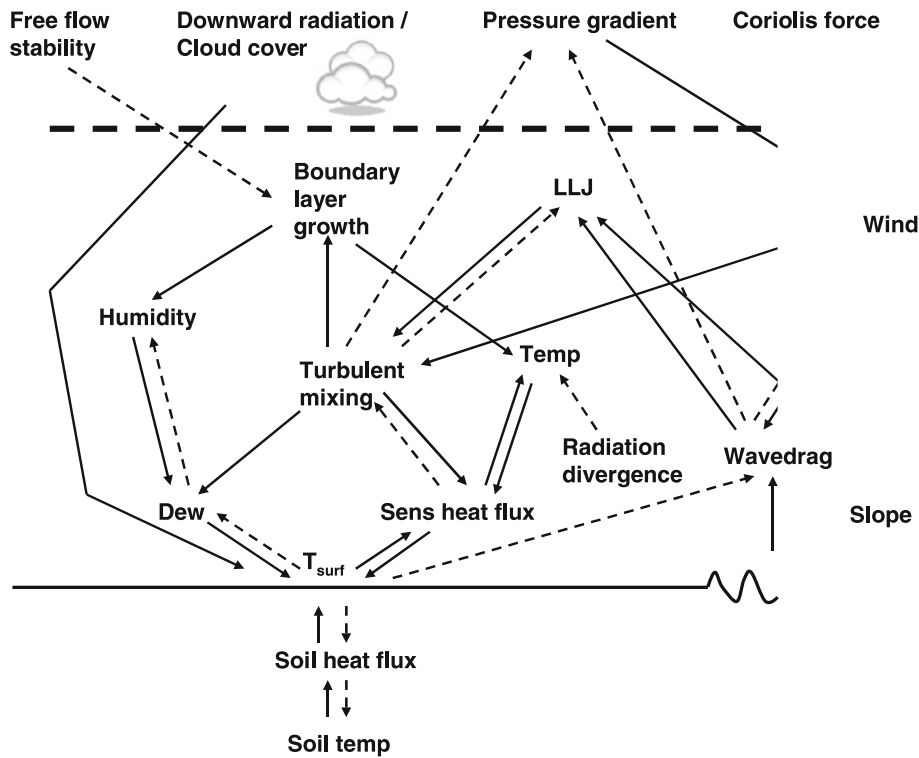
Atmospheric models for weather and climate need to make an overall representation of the smaller-scale boundary-layer and near surface processes. This appears to be more successful during daytime (e. g. Ek and Holtslag [10]; Holtslag and Ek [17]) than during nighttime stable conditions over land. The modeling of the stable boundary layer over land is rather complex because of the many different physical processes which are "at work" in stable conditions [21]. These small-scale processes are: clear air radiation divergence, drainage flow, generation of grav-

ity waves and shear instabilities, fog and dew formation, the occurrence of a low-level jet and generation of discontinuous or intermittent turbulence [33]. In addition, the phenomenology of stable atmospheric boundary layers is quite diverse, e. g. shallow and deep boundary layers with continuous turbulence through most of their depth, and on the other hand boundary layers with intermittent turbulence or even laminar flow.

The small-scale processes influence the vertical and horizontal exchange of quantities between the surface and the atmosphere as well as the mixing in the atmosphere on a variety of scales. In addition, it is known that turbulent mixing in stratified flow has an inherent non-linear character and may, as such, trigger positive feedbacks. These positive feedbacks, in turn, may cause unexpected transitions between totally different SBL regimes (e. g. van de Wiel et al. [34]).

Figure 3 depicts the interactions between relevant processes in the stable boundary layer. The non-linear behavior of the system is seen in e. g. the surface sensible heat flux ( $H$ ). A sudden change of the surface temperature can result in 2 different impacts on  $H$ . First, in weakly stable conditions (with strong wind and sufficient turbulence), a surface temperature decrease will provide a larger heat flux since  $H$  is proportional to the temperature difference between the surface and the atmosphere. The larger heat flux from the atmosphere to the surface compensates for the stronger cooling. In contrast for stronger stably stratified conditions, a surface temperature decrease will provide a stronger stratification and inhibits turbulent mixing, and consequently  $H$  will decrease. This allows for even stronger surface cooling (positive feedback). Note that a similar diagram for the daytime boundary layer can be found in Ek and Holtslag [10].

Having in mind the above mentioned complexity, one should not be surprised that atmospheric models encounter large forecast errors for stable conditions [20,24]. One strategy to improve model performance is to provide different models the same forecasting task, and analyze which model descriptions are in favor for which atmospheric stability. Recently such an intercomparison of boundary-layer schemes for stable conditions was made within the GEWEX Atmospheric Boundary Layer Study ('GABLS'). This GEWEX project aims to improve the understanding and the representation of the atmospheric boundary layer in regional and large-scale climate models [14]. A rather simple case was selected as a benchmark to review the state of the art and to compare the skills of single column (1D) models [7] and Large-Eddy Simulation models [3]. In this case a stable boundary layer is driven by an imposed, uniform geostrophic wind, with



Single Column Modeling of Atmospheric Boundary Layers and the Complex Interactions with the Land Surface, Figure 3 Interaction diagram for the processes and variables in the stable atmospheric boundary layer over land (after [26])

a specified constant surface-cooling rate over (homogeneous) ice. The case is initialized with  $\theta = 265$  K for  $0 < z < 100$  and a lapse rate of 1 K/100 m aloft.

It turns out that with the same initial conditions and model forcings, the models indicate a large range of results for the mean temperature and wind profiles. Figure 4 shows the mean profiles for several models after nine hours of constant surface cooling (sufficient to achieve a quasi-steady state). The variable results achieved are strongly related to the details of the boundary-layer mixing schemes [7]. An important finding is that the models in use at operational weather forecast and climate centers (as depicted at the left hand side of Fig. 4) typically allow for enhanced mixing resulting in too deep boundary layers, while the typical research models (at the right hand sides) show less mixing in more in agreement with the ‘Large Eddy Simulation’ results for this case [3].

Because of the enhanced mixing in weather and climate models, these models tend to show a too strong surface drag, too deep boundary layers, and an underestimation of the wind turning in the lower atmosphere [19]. At the other hand, by decreasing the mixing and surface drag, a direct impact on the atmospheric dynamics (‘Ekman

pumping’) is noted (e.g. Beljaars and Viterbo [5]). Consequently, cyclones may become too active, corresponding in too high extremes for wind and precipitation, etc.

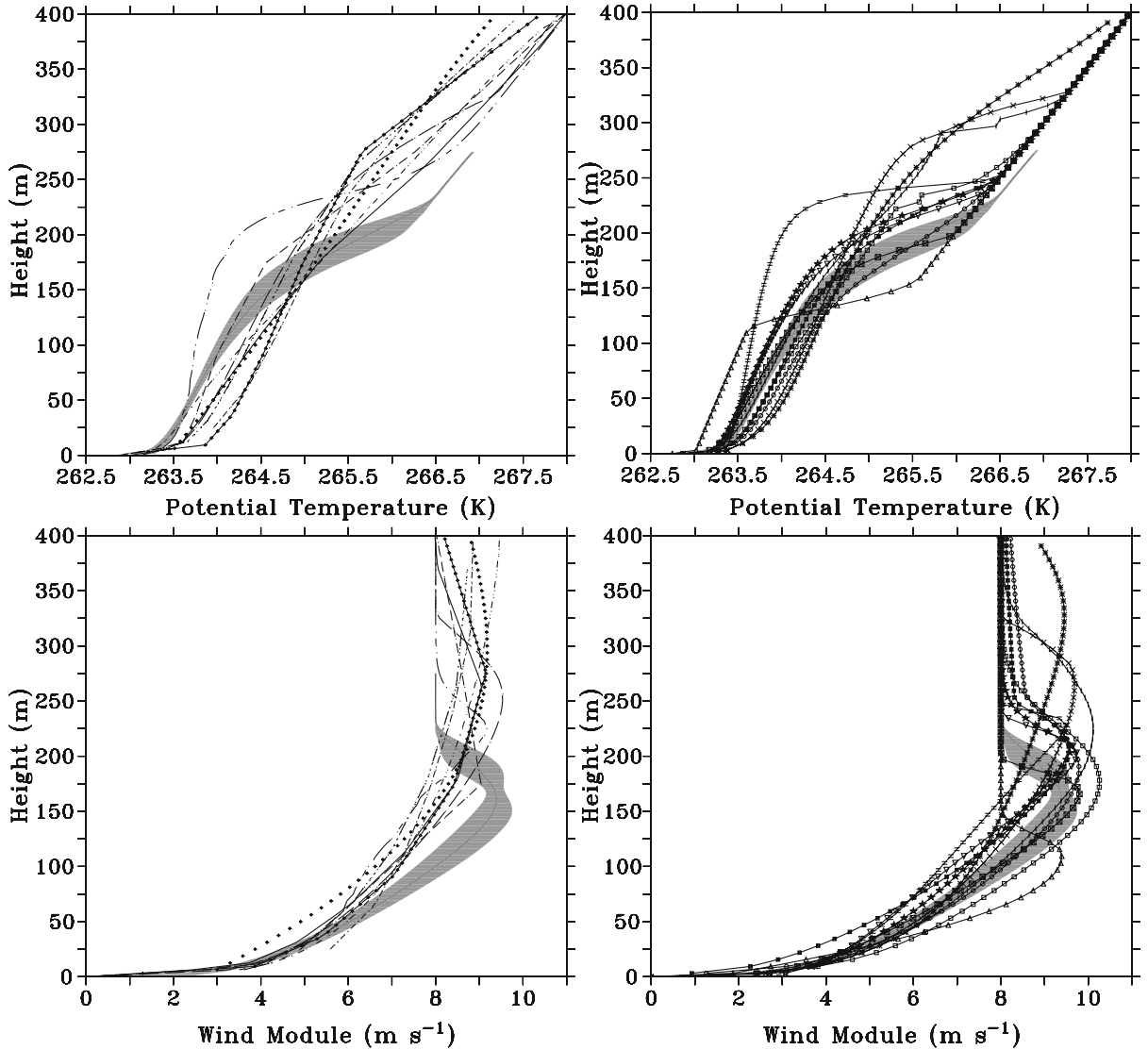
### Modeling Boundary Layers over Land

To study the interactions of the ABL with the land surface we utilize the model by Duynkerke [9] with the extensions and modifications by Steeneveld et al. [28]. This model has been validated against tower observations at Cabauw, The Netherlands and later with CASES-99 field observations [1,28], and also participated in the GABLS model comparison in the previous section.

Instead of prescribing the surface temperature, and to enable interaction between the ABL and the land surface, the model is extended with a soil and a vegetation layer. The soil temperature evolution is calculated by solving the diffusion equation (using a grid spacing of 1 cm) and the heat flux  $G_h$  from the soil to vegetation is calculated by:

$$G_h - (1 - f_{veg})K^\downarrow = r_g(T_{veg} - T_{s0}). \quad (13)$$

In Eq. (13)  $K^\downarrow$  is the incoming shortwave (solar) radiation,  $T_{veg}$  represents the vegetation surface temperature,



Single Column Modeling of Atmospheric Boundary Layers and the Complex Interactions with the Land Surface, Figure 4  
 Modeled potential temperature and wind profiles by an ensemble of column models (after 9 h). Grey areas indicate the ensemble of Large Eddy Simulation results [3]. Left panel shows the results for first order closure models and the right panel for higher order closure models (after [7])

and  $T_{s0}$  the soil temperature just below the vegetation. We use a vegetation fraction  $f_{veg} = 0.9$  and conductance  $r_g = 5.9 \text{ W m}^{-2} \text{ K}^{-1}$ , which are consistent with the observations of CASES99 [28]. Initial soil and surface temperatures are also taken from the CASES99 observations.

Subsequently, the evolution of  $T_{veg}$  is computed by solving the surface energy budget for the vegetation layer:

$$C_v \frac{\partial T_{veg}}{\partial t} = Q^* - G_h - H - L_v E. \quad (14)$$

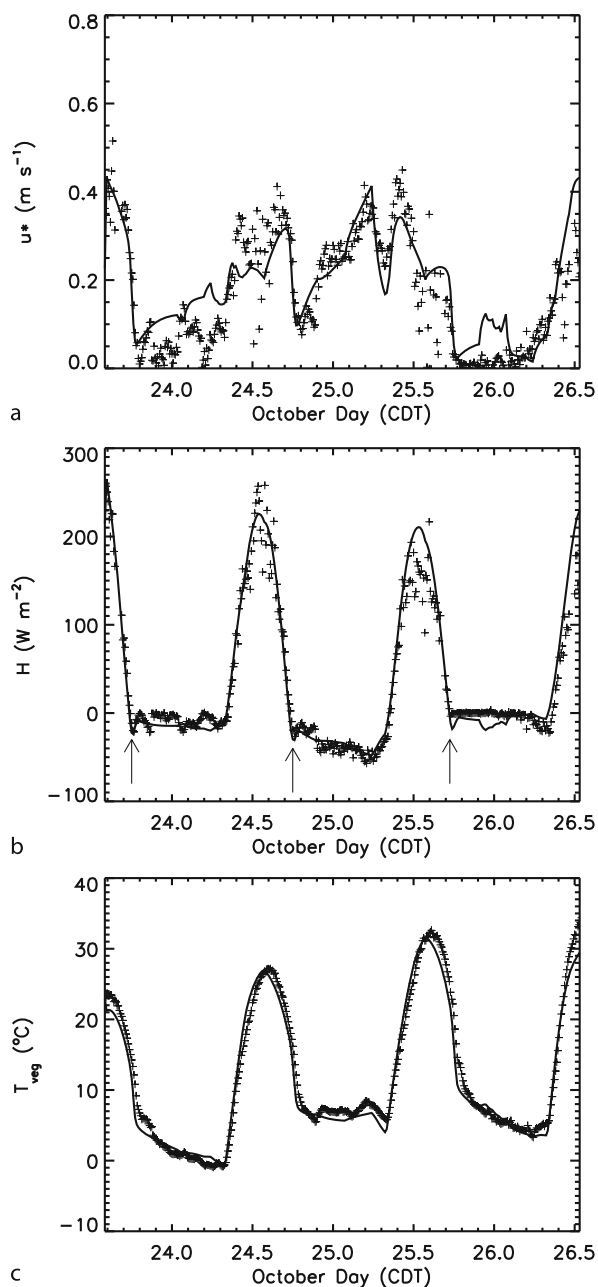
Here  $C_v$  is the heat capacity of the vegetation layer per unit of area ( $C_v = 2000 \text{ J m}^{-2} \text{ K}^{-1}$ , van de Wiel [34]),  $Q^*$  is the net radiation,  $H$  is the sensible heat flux and  $L_v E$  the latent heat flux. The turbulent fluxes are calculated basically with the format of Eq. (6) above. Finally,  $Q^*$  is calculated by adopting the Garratt and Brost [12] radiation scheme. Note that Eqs. (13) and (14) provide a rather strong coupling of the atmosphere to the vegetated land surface for the current parameter setting which is found to be important [28].

Model forecasts have been compared with CASES-99 observations for contrasting diurnal cycles (i. e. for different wind speeds) for 23–26 Oct. 1999. Note that the model is only initialized once and that the total run comprises three full days. We distinguish between “radiative nights” with weak winds and small turbulent mixing, when radiative cooling dominates the SBL development. On the contrary, so-called “continuous turbulent” nights are characterized by strong winds. The final archetype is the so called “intermittently turbulent” night where turbulent episodes alternate with calm periods, when the radiative and turbulent forcings are of similar order of magnitude. Here we restrict ourselves to the results for surface fluxes, surface vegetation temperature, and vertical profiles of temperature and wind speed during nighttime. The diurnal cycle of the modeled net radiation  $Q^*$  (the balance of all incoming and outgoing short- and longwave radiative fluxes) agrees with the observations (not shown). Net radiation amounts typically  $Q^* = 400 \text{ Wm}^{-2}$  during daytime and  $Q^* = -70 \text{ Wm}^{-2}$  during nighttime.

The friction velocity ( $u_*$ ) shows a clear diurnal cycle:  $u_*$  is large during the day and small at night, which is in general well captured by the model (Fig. 5a). Looking in more detail we find that during weak winds (1st and 3rd night) the model tends to overestimate  $u_*$ . The model lacks a clear turbulence collapse as observed during the first (intermittent) night. In the period 24 Oct., 700 CDT – 25 Oct., 1700 CDT the model performs well, while during the last (radiative) night  $u_*$  is slightly too high until midnight but follows the collapse at the end of the night. The overall bias amounts to  $0.03 \text{ ms}^{-1}$  for the last night. Sodar observations show much smaller wind speeds at 200 m AGL (which is above the SBL during this night) than the imposed G. This may suggest that G was overestimated, and this consequently may explain the bias in  $u_*$ . In general it is known that models correctly predict  $u_*$  for strong winds, but overestimate  $u_*$  for weak winds [25,32].

The sensible heat flux differs substantially between day ( $\sim 250 \text{ Wm}^{-2}$  here) and night (between 0 and  $-60 \text{ Wm}^{-2}$  depending on the wind speed). In the first (intermittent) night, the modeled  $H = -14.1 \text{ Wm}^{-2}$  on average, while  $-9.1 \text{ Wm}^{-2}$  was observed. However, the model does not simulate the observed intermittent character of the surface fluxes (Fig. 5b). Some models with more resolution [23,36] were also able to reproduce intermittent turbulence. On the other hand, the models by Sharan and Gopalakrishnan [25] and Derbyshire [8] did not show any intermittency. This subject needs further investigation.

Just after the day-night transition to the intermittent night, the observed magnitude of  $H$  shows a clear maximum (see arrows in Fig. 5b), which is well reproduced



Single Column Modeling of Atmospheric Boundary Layers and the Complex Interactions with the Land Surface, Figure 5 Modeled and observed (+) friction velocity (a), surface sensible heat flux (b), and surface vegetation temperature (c) for three diurnal cycles in CASES-99 (after [28])

by the model. This maximum is caused by a sudden reversal of the stratification near the surface due to longwave radiation emission during the day-night transition, and is maintained by residual turbulence of the convec-

tive boundary layer. This is an often observed *realistic feature* (e. g. during 11 of the 30 nights for CASES-99 and in FIFE observations shown by [6]). However, most modeling studies rarely show these minima. The reproduction of this detailed feature emphasizes the realism of the model outcome.

During the turbulent night (24–25 Oct.), the predicted  $H$  follows the observations. The specific minimum during the day-night transition (24 Oct., 1900 CDT) is present here as well. Radiative flux divergence dominates the last (radiative) night and the observed  $H$  is approximately zero. The model slightly overestimates the magnitude of  $H$  ( $-2.9 \text{ Wm}^{-2}$ ), mainly caused by an overestimation of  $u_*$ . This causes a weaker stratification and thus a larger magnitude of  $H$ . The second half of this night the model gives good results.

Reliable prediction of  $T_{\text{veg}}$  is a common problem for large-scale models. Some models show unphysical decoupling of the atmosphere from the surface resulting in so-called “runaway cooling” of  $T_{\text{veg}}$ . On the other hand, the pragmatic enhanced mixing approach which is commonly used for very stable conditions, leads to overestimation of  $T_{\text{veg}}$ . For both day- and nighttime  $T_{\text{veg}}$  is simulated in very good agreement with the data, despite the fact that we cover a broad range of stability (Fig. 5c).

We conclude that the present model generates surface fluxes which are in good agreement with observations, because of the detail in the description of the surface scheme, the soil heat flux and radiation physics (with high resolution). In general, the model is also able to estimate temperature and wind profiles (see results and discussion in Steeneveld et al. [28]). To examine the robustness of the results, we performed some sensitivity analysis on the initial conditions and model parameters. Disturbing the initial temperature (by 1 K), wind profiles (by 5%), soil temperature and vegetation temperature (both by 1 K) do not affect the results seriously. Also model re-initialization every 24 h (1400 CDT) with observed radiosonde information showed hardly any impact on the results (not shown).

### Impact of Land Surface Conditions on Model Results

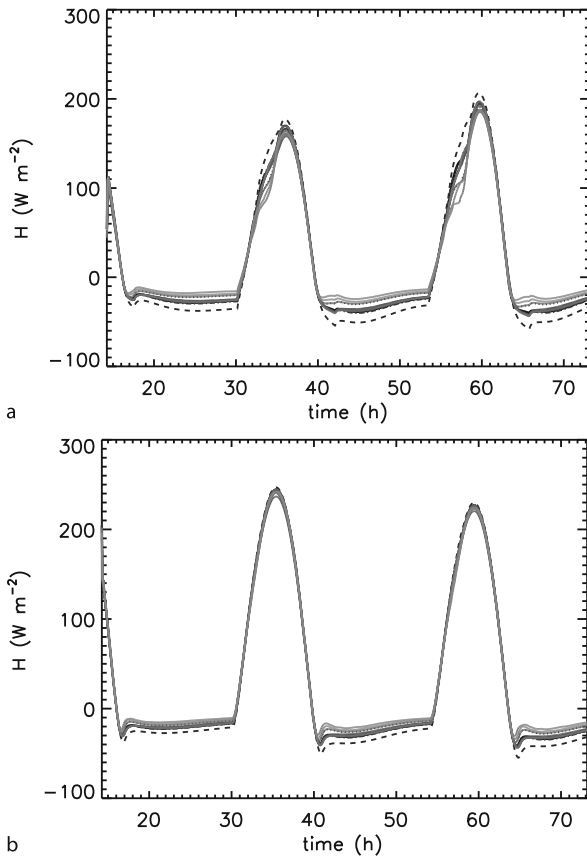
Inspired by the result in the previous section that a coupling with the land surface is necessary to obtain satisfactory model results, we now analyze the difference in variability of model results during model intercomparisons, as function of the chosen boundary condition. At first we use the first-order closure model and vary the parameters in the turbulence scheme for stable conditions in a reasonable range to mimic the apparent variability among

boundary-layer models. As such first model runs are performed with a prescribed surface temperature as inspired by (but not identical to) the observations in CASES99 [22] and as described in the GABLS2 case description [31]. Subsequently, the model runs are repeated, but then using an interactive prognostic heat budget equation for the surface temperature (Eq. (14)).

To study the impacts of parameter values on the model results, reference runs are made for coupled and uncoupled cases with alternative permutations in some of the parameter settings for stable conditions. The parameter modifications are chosen such that they cover a realistic range in comparison with existing models of the stable boundary layer [7]. The local starting time in the model runs is 14.00 LT on October 22, 1999 (rather than 16.00 LT in the GABLS2 runs). The duration of all runs is 59 hours (so that the axis of all the figures indicates 14.00 until 73.00 h, covering a period of 2.5 diurnal cycles). In all model runs the roughness length for heat  $z_{\text{oh}}$  and momentum  $z_{\text{om}}$  (3 mm and 3 cm respectively), and the canopy resistance are constant, and the geostrophic wind is taken at a reference value of  $9.5 \text{ ms}^{-1}$  (as in Svensson and Holtslag [31]). The reference model set up has 50 logarithmically distributed layers and the first atmospheric model level is at 2 m.

The model results for all parameter permutations are presented for the sensible heat flux (Fig. 6). In the upper sub-frame of the figure (labeled a) the results achieved with the uncoupled model are given (using prescribed surface temperature). Overall the variety of results in the upper frame is comparable to the variety within the GABLS2 intercomparison study in stable conditions for the uncoupled models (see Svensson and Holtslag [31]). Thus we have a range of  $-15$  to  $-50 \text{ Wm}^{-2}$  for the sensible heat flux (at the end of the first night e. g. at the time of 30 h). The variability is a result of the range of parameters chosen above and the impact is apparently sufficient to mimic the different parameterizations for stable conditions in the models used within GABLS2.

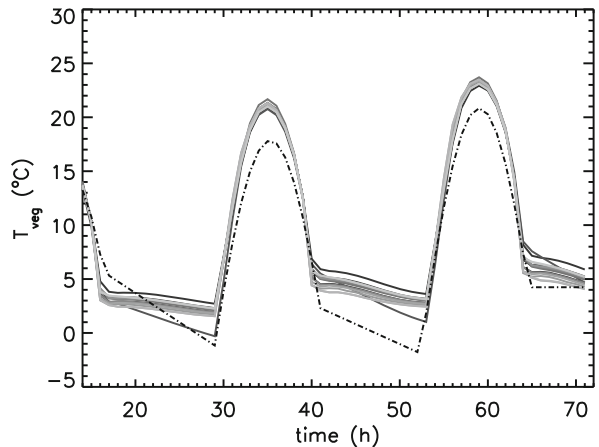
Next we repeat all model runs and allow for surface feedback using Eqs. (13) and (14). The results for the sensible heat flux with the coupled model are given in the lower frame (Fig. 6b). Now we have a range of  $-10$  to  $-25 \text{ Wm}^{-2}$  (again the values apply for the end of the first night e. g. at the time of 30 h). Thus it appears that the variety of model results is *smaller* for the sensible heat flux in the coupled case. At the same time it appears that the variability appears to be somewhat larger for friction velocity and boundary-layer depth, which seems to be related to the larger variability in the near surface air temperature and wind speed.



Single Column Modeling of Atmospheric Boundary Layers and the Complex Interactions with the Land Surface, Figure 6  
**Time series of an ensemble of model results for the sensible heat flux in a model intercomparison study with a prescribed surface temperatures, and b by solving the surface energy budget (after [18])**

During daytime the sensible heat fluxes are rather similar for all model runs within one category (either coupled or uncoupled), but the maximum values differ. In addition, due to the coupling the sensible heat fluxes show a more smooth behavior in the morning hours as compared with the uncoupled results. Thus, surface feedback is influencing the model results and is also able to compensate for some variation in the model parameter values. Note also that the variability in the friction velocities of the first night remains during the morning hours in the uncoupled runs, but not so much in the coupled case.

In Fig. 7 the surface temperatures are given as specified for the uncoupled case (the dashed line), and the temperatures as calculated in the various interactive model runs (various gray lines). It is seen that the latter ones are quite different from each other (in particular at night). It is also important to note that the surface temperature by the en-



Single Column Modeling of Atmospheric Boundary Layers and the Complex Interactions with the Land Surface, Figure 7  
**Time series of an ensemble of modeled surface temperature for coupled runs. Dash-dotted line: prescribed surface temperature in the uncoupled case (after [18])**

semble of coupled model runs is clearly different from the specified temperature in the uncoupled case. This impacts also on the absolute values and the range of air temperatures and the wind speeds.

Thus it is apparent that the treatment of the surface temperature impacts strongly on the outcome of the boundary-layer model results and their variety (see also Basu et al. [2]). By repeating the uncoupled model runs with a specified surface temperature as given by the ensemble mean value of the interactive runs, we achieve basically the same variety of model outputs for the potential temperature and wind as for the coupled cases. This confirms that in model evaluation studies the surface temperature should be taken consistent with the value of the geostrophic wind (although this may be model dependent).

## Summary

In this paper a summary is given of the basic approaches for the modeling and parameterization of turbulence in the atmospheric boundary layer. The treated approaches are in current use in regional and global-scale models for the forecasting and study of weather, climate and air quality. Here we have shown the results of such approaches by using single column models in comparison with field data and fine-scale model results. We have also studied the impact of the surface temperature condition on the variability of results by an atmospheric boundary-layer model. From the coupled model results we achieve that surface feedback can compensate for some of the variety introduced by

changing model parameters. Generally much work needs still to be done before we have a full understanding of the complexity of atmospheric turbulence and the interactions with the land surface. A better understanding of atmospheric turbulence hopefully also contributes to our capability in refining and unifying the turbulence parameterizations for modeling of the atmospheric boundary layer in response to the different type of surfaces which are found in reality.

### Acknowledgments

This contribution is a compilation of earlier works by the authors, in particular the works by Holtslag [13], Steeneveld et al. [28] and Holtslag et al. [18].

### Bibliography

1. Baas P, Steeneveld GJ, van de Wiel BJH, Holtslag AAM (2006) Exploring Self-correlation in flux-gradient relationships for stably stratified conditions. *J Atmos Sci* 63:3045–3054
2. Basu S, Holtslag AAM, van de Wiel BJH, Moene AF, Steeneveld GJ (2007) An inconvenient 'truth' about using the sensible heat flux as a surface boundary condition in models under stably stratified regimes. *Acta Geophys* 56:88–99. doi:10.2478/s11600-007-0038-y
3. Beare R, MacVean M, Holtslag AAM, Cuxart J, Esau I, Golaz J-C, Jimenez M, Khairoutdinov M, Kosovic B, Lewellen D, Lund T, Lundquist J, McCabe A, Moene A, Noh Y, Raasch S, Sullivan P (2006) An intercomparison of Large-Eddy Simulations of the stable boundary layer. *Bound-Layer Meteorol* 118:247–272
4. Beljaars ACM, Holtslag AAM (1991) Flux parameterization over land surfaces for atmospheric models. *J Appl Meteor* 30:327–341
5. Beljaars ACM, Viterbo P (1998) Role of the boundary layer in a numerical weather prediction model. In: Holtslag AAM, Duynkerke PG (eds) *Clear and Cloudy boundary layers*. Royal Netherlands Academy of Arts and Sciences, Amsterdam, 372 pp
6. Chen F, Dudhia J (2001) Coupling an Advanced Land Surface-Hydrology Model with the Penn State-NCAR MM5 Modeling System. Part II: Preliminary Model Validation. *Mon Wea Rev* 129:587–604
7. Cuxart J, Holtslag AAM, Beare RJ, Bazile E, Beljaars A, Cheng A, Conangla L, Ek MB, Freedman F, Hamdi R, Kerstein A, Kitagawa H, Lenderink G, Lewellen D, Mailhot J, Mauritsen T, Perov V, Schayes G, Steeneveld GJ, Svensson G, Taylor P, Weng W, Wunsch S, Xu K-M (2006) Single-column model intercomparison for a stably stratified atmospheric boundary layer. *Bound-Layer Meteorol* 118: 273–303
8. Derbyshire SH (1999) Boundary layer decoupling over cold surfaces as a physical boundary instability. *Bound-Layer Meteorol* 90:297–325
9. Duynkerke PG (1991) Radiation fog: A comparison of model simulation with detailed observations. *Mon Wea Rev* 119:324–341
10. Ek MB, Holtslag AAM (2004) Influence of Soil Moisture on Boundary Layer Cloud Development. *J Hydrometeor* 5:86–99
11. Garratt JR (1992) *The Atmospheric Boundary Layer*. Cambridge University Press, New York, 316 pp
12. Garratt JR, Brost RA (1981) Radiative Cooling effects within and above the nocturnal boundary layer. *J Atmos Sci* 38:2730–2746
13. Holtslag AAM (2002) Atmospheric Boundary Layers: Modeling and Parameterization. In: Holton JR, Pyle J, Curry JA (eds) *Encyclopedia of Atmospheric Sciences*, vol 1. Academic Press, pp 253–261
14. Holtslag AAM (2006) GEWEX Atmospheric Boundary Layer Study (GABLS) on Stable Boundary Layers. *Bound-Layer Meteorol* 118:243–246
15. Holtslag AAM, Moeng C-H (1991) Eddy diffusivity and counter-gradient transport in the convective boundary layer. *J Atmos Sci* 48:1690–1698
16. Holtslag AAM, de Bruin HAR (1988) Applied Modeling of the Nighttime Surface Energy Balance over Land. *J Clim Appl Meteor* 27:689–704
17. Holtslag AAM, Ek MB (2005) Atmospheric Boundary Layer Climates and Interactions with the Land Surface. In: *Encyclopedia of Hydrological Sciences*. Wiley
18. Holtslag AAM, Steeneveld GJ, van de Wiel BJH (2007) Role of land-surface feedback on model performance for the stable boundary layer. *Bound-Layer Meteorol* 125:361–376
19. Lenderink G, Van den Hurk B, van Meijgaard E, van Ulden A, Cuijpers H (2003) Simulation of present day climate in RACMO2: First results and model developments. KNMI Technical report TR-252, 24 p
20. Mahrt L (1998) Stratified atmospheric boundary layers and breakdown of models. *Theor Comp Fluid Phys* 11:263–279
21. Mahrt L (1999) Stratified atmospheric boundary layers, *Bound-Layer Meteorol* 90:375–396
22. Poulos GS et al (2002) CASES-99: A comprehensive investigation of the stable nocturnal boundary layer. *Bull Am Meteor Soc* 83:555–581
23. ReVelle DO (1993) Chaos and "bursting" in the planetary boundary layer. *J Appl Meteor* 32:1169–1180
24. Salmond JA, McKendry IG (2005) A review of turbulence in the very stable boundary layer and its implications for air quality. *Prog Phys Geogr* 29:171–188
25. Sharan M, Gopalakrishnan SG (1997) Comparative Evaluation of Eddy Exchange coefficients for strong and weak wind stable boundary layer modelling. *J Appl Meteor* 36:545–559
26. Steeneveld GJ (2007) Understanding and prediction of stable boundary layers over land. PhD thesis, Wageningen University, 199 pp
27. Steeneveld GJ, van de Wiel BJH, Holtslag AAM (2006) Modeling the arctic stable boundary layer and its coupling to the surface. *Bound-Layer Meteorol* 118:357–378
28. Steeneveld GJ, van de Wiel BJH, Holtslag AAM (2006) Modeling the Evolution of the Atmospheric Boundary Layer Coupled to the Land Surface for Three Contrasting Nights in CASES-99. *J Atmos Sci* 63:920–935
29. Steeneveld GJ, Mauritsen T, de Bruijn EIF, Vila-Guerau de Arellano J, Svensson G, Holtslag AAM (2008) Evaluation of limited area models for the representation of the diurnal cycle and contrasting nights in CASES99. *J Appl Meteor Clim* 47:869–887
30. Stull RB (1988) *An introduction to Boundary-Layer Meteorology*. Kluwer, Dordrecht, 666 pp. (reprinted 1999)
31. Svensson G, Holtslag AAM (2006) Single column modeling of the diurnal cycle based on CASES99 data -GABLS second intercomparison project. 17th Symposium on Boundary lay-

- ers and Turbulence, 22–25 May, San Diego. American Meteorol Soc, Boston, Paper 8.1 (available at <http://ams.confex.com/ams/pdfpapers>)
32. Tjemkes SA, Duynkerke PG (1989) The nocturnal boundary layer: model calculations compared with observations. *J Appl Meteor* 28:161–175
  33. van de Wiel BJH (2002) Intermittency and Oscillations in the Stable Boundary Layer over Land. Ph D thesis, Wageningen University, 129 pp
  34. van de Wiel BJH, Moene AF, Hartogensis OK, de Bruin HAR, Holtslag AAM (2003) Intermittent turbulence and oscillations in the stable boundary layer over land, Part III: a classification for observations during CASES99. *J Atmos Sci* 60:2509–2522
  35. van de Wiel BJH, Moene AF, Steeneveld GJ, Hartogensis OK, Holtslag AAM (2007) Predicting the Collapse of Turbulence in Stably Stratified Boundary Layers. *Turbul Flow Combust* 79:251–274
  36. Welch RM, Ravichandran MG, Cox SK (1986) Prediction of Quasi-Periodic Oscillations in Radiation Fogs. Part I: Comparison of Simple Similarity Approaches. *J Atmos Sci* 43:633–651



## Slug Flow: Modeling in a Conduit and Associated Elastic Radiation

LUCA D'AURIA, MARCELLO MARTINI

Osservatorio Vesuviano, Istituto Nazionale di Geofisica e Vulcanologia, Sezione di Napoli, Naples, Italy

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Slug Flow and Strombolian Activity](#)

[Numerical Modeling](#)

[Bubble Ascent](#)

[Slug Ascent in a Vertical Pipe](#)

[Slug Ascent in a Pipe with a Flare](#)

[Seismological Constraints on Numerical Models](#)

[Conclusions](#)

[Future Directions](#)

[Appendix A – Fluid Dynamics of a Two-Phase System](#)

[Appendix B – Numerical Implementation](#)

[Bibliography](#)

### Glossary

**Strombolian activity** Is a kind of volcanic activity consisting of discrete, intermittent ejections of gas and magma fragments. The height reached by the ejecta seldom exceeds a few hundred meters above the volcanic crater. They consist mostly of molten magma fragments that partially quench as they fly. Strombolian explosions are intermediate between small Hawaiian eruptions and stronger Plinian eruptions. This kind of activity takes its name from the Stromboli volcano (Southern Italy) famous for its perpetual activity consisting of repeated moderate explosions (usually about 10 every hours).

**Gas slug** A bubble whose diameter is close to the diameter of the conduit where it is flowing. The shape and the motion of the slug is largely controlled by the conduit walls.

**Modeling** A simplified representation of a system aimed at reproducing some of its features. Mathematical models describe a system through a set of variables related by analytical relations. When these relations are too complex to be solved exactly, they can be solved using approximate numerical methods. Actually these numerical techniques often involve the massive use of computers. An alternative kind of modeling (analogue

modeling) involves the use of versions of the system under study rescaled to fit spatial and temporal laboratory scales. Different materials are used to simulate original ones. For instance silicon oil is often used to simulate magma. The scaling relationships are rigorously stated in order to get physically meaningful results.

**Computational fluid dynamics** A set of mathematical, numerical and computational tools aimed at simulating complex fluid flows on computers. It developed simultaneously with computer science starting from the 1950s mostly with the aim of solving engineering problems. Today CFD spans a wider range of fields, from aeronautics to chemical engineering, to astrophysics to geophysics and much more. The most recent developments of CFD are related with the increase in computer performances and with the wider use of parallel computers.

**Diffuse interface theory** A molecular theory for describing the variation of the chemical composition across the interface on the basis of a rigorous thermodynamical approach. Starting from the definition of a free-energy function dependent on the chemical composition and its gradient, it is possible to compute all the chemical-physical properties of the interface. Beyond its physical meaning, this theory can be used also as a numerical tool for modeling of multiphase flows.

**Very-long-period events (VLP)** Seismic events recorded on active volcanoes and geothermal systems having a typical period of  $10^2$ – $10^0$  s. Their observation and study began during the 1990s, with the spreading of seismic broadband sensors and have shown to be one of the most powerful tools for investigating the geometries of volcanic conduits and the dynamics of volcanic eruptions. Until now they have been observed in tens of volcanoes with different eruptive styles such as: Aso (Japan), Erebus (Antarctica), Kilauea (Hawaii), Popocatepetl (Mexico), Sakurajima (Japan), Stromboli (Italy).

**Moment-tensor** Tensor representation of the force systems acting on seismic sources. It can be applied to common earthquake sources (rupturing faults) as well as to volcanic sources (fluid filled conduits). In the former case its trace is null, which from a physical point of view, means that there is no net volume change during common earthquakes. On the other hand in volcanic sources volumetric variations are very common and furthermore they are accompanied also by a single force component, related to the net acceleration of center of mass of the fluid filling the conduit.

## Definition of the Subject

Among the eruptive styles, the Strombolian activity is one of the more easy to study because of its repetitive behavior. For this reason large amount of data can be comfortably collected. Strombolian volcanoes are like natural laboratories repeating the same experiment (individual explosions) many times each day.

The development of quantitative models of eruptive dynamics is driven by the comparison of experimental observations and synthetic data obtained through mathematical, numerical or analogue modeling.

Since Strombolian activity offers a profuse amount of interesting seismic signals, during the last decades there has been growing attention on seismological techniques aimed at retrieving the conduit geometry and the eruption dynamics from the seismological recordings. One of these techniques, the source function inversion, is able to retrieve a summary of the forces acting on the volcanic conduit during the VLP event generation [5]. The comparison of observed source functions with synthetic ones, obtained through numerical modeling, allow us to put constraints on the proposed models.

Quantitative models, able to fit seismological observations, are a powerful tool for interpreting seismic recordings and therefor the seismological monitoring of active volcanoes.

## Introduction

In this paper we discuss the mechanism of generation of Very-Long-Period events related to Strombolian explosions. This eruptive style, occurring in many basaltic volcanoes worldwide, is characterized by the ascent and the bursting of large gas slugs. The mechanism of formation, ascent and explosion of bubbles and slugs and their relation to eruptive activity has been studied from a theoretical point of view and by means of analogue simulations. Here we introduce results from numerical simulations, focusing on the pressure variations induced on the conduit walls and responsible for the generation of seismic signals.

We will first illustrate the main features of the fluid dynamics related to Strombolian eruptive activity (Sect. “**Slug Flow and Strombolian Activity**”) and an overview of the numerical modeling (Sect. “**Numerical Modeling**”) Then we will show results obtained using simple conduit model (Sect. “**Bubble Ascent**”, Sect. “**Slug Ascent in a Vertical Pipe**” and Sect. “**Slug Ascent in a Pipe with a Flare**”) and we will compare the synthetic source functions with actual observations (Sect. “**Seismological Constraints on Numerical Models**”).

## Slug Flow and Strombolian Activity

A fundamental distinction can be made between eruptive regimes on the basis of the magma viscosity. In silicic systems, the magma viscosity ( $>10^5$  Pa s) is too high to allow an independent motion of gas bubbles [18]. They can only grow by diffusion processes and expand under the effects of pressure variations until fragmentation occurs. In basaltic magmas the viscosity ( $<10^3$  Pa s) allows independent motion of the gas bubbles leading to a different behavior with the possibility of bubble coalescence, splitting and turbulent fluid flow [11]. Theoretical laboratory and numerical studies have been published in order to understand the dynamics of Strombolian eruption in terms of gas/magma interaction [12,24].

The distinction between free bubble ascent and slug flow can be made using the ratio between the average bubble diameter  $d$  and the conduit diameter  $D$  [7]: the parameter  $\lambda$  (Table 1). Values of  $\lambda$  higher than 0.6 are related to slug flow. The range of behavior exhibited by slug flow depends on the physical properties of the fluids (density, viscosity, surface tension) and on the geometry of the conduit (diameter, shape, inclination). It is possible to define adimensional parameters for describing the particular flow regime (see Table 1). The Reynolds number is the ratio between inertial and viscous forces. Low values of  $Re$  are typical of laminar flows, while higher values are related to turbulent regimes. The Froude number  $Fr$  is the ratio between inertial and gravitational forces. The Eotvos number  $Eo$  is the ratio between buoyancy and surface tension effects. The values of  $Eo$  determines the bubble shape. High values of  $Eo$  are related to distorted bubble shapes, while lower values to sub-spherical shapes. The Morton number  $Mo$  has a similar meaning. The dimensionless inverse viscosity  $N_f$  assess the relative importance of viscous effects. Viscous flows are characterized by  $N_f < 2$  while inertial flows by  $N_f > 200$  [9].

The ascent of gas in a basaltic system has been studied both from a theoretical and an experimental point of view. Observations of basaltic systems have shown that usually the flow conditions are transitional between viscous dominated and inertia dominated systems [17]. This puts constraints on the parameter range to explore both in numerical and analogue modeling. The range of adimensional numbers for basaltic systems, summarized from [17] and [9] is reported in Table 2.

For a basaltic system with given physical properties another variable plays a fundamental role in determining the physics of the flow: the gas/magma volumetric ratio. Flows are characterized by a low ratio consist of isolated bubbles rising with minor interaction between them

Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Table 1  
 Symbols used in the text

Symbol	Meaning
$d$	Average bubble diameter
$D$	Conduit diameter
$L$	Distance between the top of the bubble and the magma/air interface in the initial conditions.
$U$	Terminal slug velocity
$\lambda$	$\lambda = \frac{d}{D}$
$g$	Gravity
$\rho$	Liquid density
$\Delta\rho$	Difference between gas and liquid densities
$\mu$	Dynamic viscosity
$\nu$	Kinematic viscosity $\nu = \frac{\mu}{\rho}$
$\sigma$	Surface tension
$\gamma$	Isothermal expansion ratio
$Re$	Reynolds number $Re = \frac{Ud}{\nu}$
$Fr$	Froude number $Fr = \frac{U}{\sqrt{gD}} \sqrt{\frac{\rho}{\Delta\rho}}$
$Mo$	Morton number $Mo = \frac{g\mu^4\Delta\rho}{\rho^2\sigma^3}$
$Eo$	Eotvos number $Eo = \frac{\Delta\rho g D^2}{\sigma}$
$N_f$	Dimensionless inverse viscosity $N_f = \left(\frac{Eo^3}{Mo}\right)^{\frac{1}{4}}$

and the conduit walls (Fig. 1a). In this regime bubbles rise assuming a shape that depends on their size and on the magma viscosity. Surface tension effect depends on the size of the bubbles and the ascent velocity depends on both the bubble size and the magma viscosity. In the range allowed for basaltic systems, larger bubbles have shapes ranging from dimpled ellipsoidal cap to spherical cap, while smaller bubbles have shapes ranging from spherical to ellipsoidal [7]. For higher gas/magma ratios bubbles begin to interact and to coalesce forming gas slugs occupying most of the conduit diameter (Fig. 1b). The ascent of gas slugs is strongly controlled by the conduit walls. During their passage, the magma is completely mixed and the motion of smaller bubbles is dominated by the fluid vorticity released (in a transitional regime) in their wake. For even higher gas/magma ratios (Fig. 1c) the slugs co-

alesce forming an almost continuous (turbulent) gas flow at the center of the conduit, while walls remains wet with a magma film. The high velocity of the flow produces instabilities on the surface of the magma film pulling out small fragments carried away from the gas.

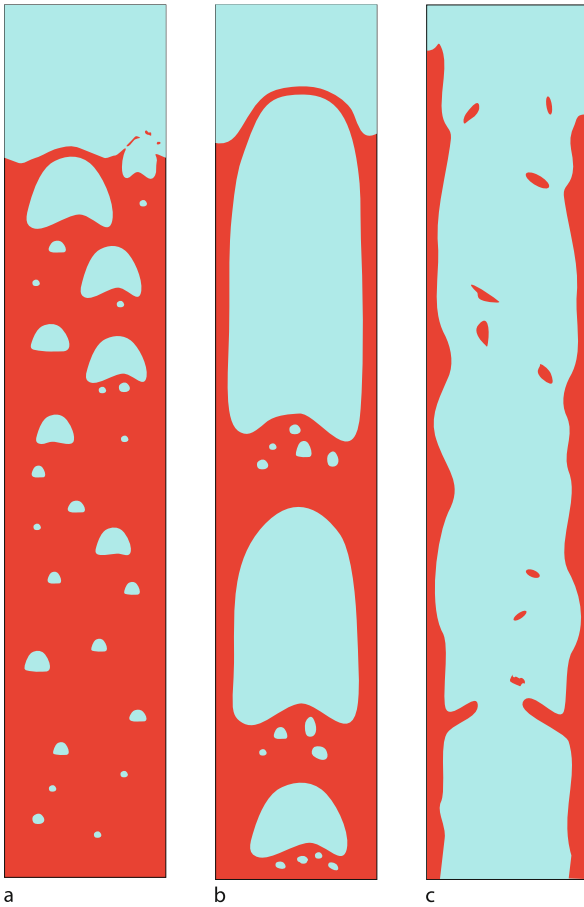
Bubbly flows are related to effusive and to moderate Hawaiian explosive activity. Each single bubble explodes at the magma/air interface ejecting small magma fragments resulting from the shattering of the liquid film around the bubble. The bubble is usually overpressurized because of surface tension, viscous and inertial effects [23]. Slug flows are related instead to intermittent Strombolian activity. Each slug exploding at the magma/air interface generates a jet of gas of short duration (from a few seconds to some tens of seconds) [24]. The main Strombolian explosion can be followed by minor bursts caused by the smaller bubbles trapped behind the slug wake. Annular flows are related to continuous lava fountains with the central gas jet carrying molten magma fragments [11].

The nature of the flow can change along the conduit because of the gas expansion due to the magmastatic pressure decrease. This expansion makes the gas/magma volumetric ratio increase along the conduit. At the base of the conduit, where the gas starts to evolve from the magma there is always a bubbly flow. As the pressure decreases the flow can change into slug flow. This can be made easier by an inclined conduit [4] and by constrictions [12] that force the bubbles to coalesce even at lower gas/magma ra-

Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Table 2

Range of adimensional numbers for flows related to basaltic systems (from [17] and [9])

Parameter	Range
$Fr$	$0.1 \div 0.345$
$Re$	$5 \div 31^3$
$Eo$	$51^5 \div 71^7$
$Mo$	$51^5 \div 101^{10}$
$N_f$	$16 \div 5000$



Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Figure 1

Schematic representation of gas/magma flows in basaltic systems. **a** represents bubbly flow, **b** slug flow and **c** annular flow

tions than those required in a straight vertical conduit. The total gas flow can change in the conduit because of variations in the deep feeding system or because of non-linear instabilities due to the complex shape of the conduit system [11,12].

The expansion of bubbles and slugs is a fundamental factor to take into account when studying Strombolian activity [17]. Assuming a simple ideal isothermal state equation for the gas we get the relative volumetric change for a bubble rising in a magma with density  $\rho$ :

$$\frac{V_0}{V} = \frac{P_{\text{atm}} + \rho gh}{P_{\text{atm}}} \quad (1)$$

It is easy to show that the expansion ratio in the upper few hundreds of meters is more than one order of magnitude.

The flow regime can change also horizontally, for instance in a subvertical dike the flow rate can be higher on

one side, leading to a slug flow, and lower on the other side, leading to a bubbly flow. This explains the coexistence of effusive and Strombolian activity observed during some basaltic eruptions.

### Numerical Modeling

The aim of this work is to investigate the pressure variations induced by a gas bubble rising in a magma-filled volcanic conduit. This phenomenon has been also investigated by means of analogue laboratory models [9,10,12,17].

The major drawback of analogue modeling, in this context, is that it provides only a limited number of pressure time series: one for each sensor. Numerical modeling provides a different point of view giving the full set of scalar (density), vector (velocity) and tensor (pressure) quantities over the whole computational domain. This allows quantitative inferences on the flow regimes and on the elastodynamic wavefield generated (see Sect. “[Seismological Constraints on Numerical Models](#)”).

The modeling of two-phase systems as gas magma is not a simple task in computational fluid dynamics (CFD) [6]. Taking into account surface tension effects at the liquid-gas interface can be done in two different ways. The first consists in explicitly tracking the time evolution of the gas-liquid interface [22]. In situations involving complex flows with extensive occurrences of bubble coalescence and splitting, these methods show a rapid increase in computational effort. Another category of methods model the gas-liquid systems using the diffuse-interface theory. These methods consider two scalar fields, defining the relative local concentrations of the two components and a modified advection-diffusion equation for modeling the physical-chemical interaction between them. This is done using a thermodynamically consistent definition of a free-energy function that takes into account the phase equilibria. This approach leads to smooth interfaces where the concentration of one phase gradually decreases while the other increases. The thickness of these interfaces depends on the numerical method and on the surface tension value [25]. From a numerical point of view this problem can be faced using Lattice Boltzmann Methods (LBM) [20] or by classic CFD [6]. Here we use the latter. Some details about the numerical method are reported in the Appendices.

In our models we have not considered the mechanical interaction between the fluid phases and the elastic conduit walls. The exchange of linear momentum between them is a significant factor in the generation of seismic waves [3] in seismo-volcanic sources. However in

Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Table 3  
Description and adimensional numbers for the six simulations presented in the text

Sim.	Description	Comp. domain (pts.)	$\lambda$	$\gamma$	$Re$	$\log(Mo)$	$Eo$	$Fr$	$N_f$
1A	Bubble ascent (low viscosity)	60 × 600	0.5	5	$1.93 \times 10^3$	0	$9 \times 10^4$	0.37	$5.19 \times 10^3$
1B	Bubble ascent (high viscosity)	60 × 600	0.5	5	$1.90 \times 10^2$	4	$9 \times 10^4$	0.36	$5.19 \times 10^2$
2A	Slug ascent in vertical pipe (low viscosity)	30 × 600	0.8	5	$1.82 \times 10^3$	0	$9 \times 10^4$	0.35	$5.19 \times 10^3$
2B	Slug ascent in vertical pipe (high viscosity)	30 × 600	0.8	5	$1.54 \times 10^2$	4	$9 \times 10^4$	0.30	$5.19 \times 10^2$
3A	Slug ascent in vertical pipe with a flare (low viscosity)	60 × 600	0.8	5.3	$1.83 \times 10^3$	0	$9 \times 10^4$	0.35	$5.19 \times 10^3$
3B	Slug ascent in vertical pipe with a flare (high viscosity)	60 × 600	0.8	5.3	$1.68 \times 10^2$	4	$9 \times 10^4$	0.32	$5.19 \times 10^2$

conduits having ratios between the linear dimension and thickness lower than  $10^1$ – $10^2$  the effect of the motion of the conduit walls does not affect significantly the fluid flow. Furthermore we show in Sect. “**Seismological Constraints on Numerical Models**” that we will not compare the result of the simulation directly with seismograms, but with an equivalent system of forces acting on the conduit walls.

In the following we present results of some elementary two-dimensional conduit models focusing on the slug flow regime which occurs during Strombolian activity. The numerical simulations generate snapshots of the physical quantities (composition, density, pressure and velocity) along all the point of the discretized computational domain. The effective computational domains for each simulation are showed in Table 3. In all the simulations we start from a static fluid with a bubble in the lower part of the conduit and a gas/magma interface in the upper part. The boundary conditions keeps a constant pressure value at the bottom and the top of the model. The no-slip boundary conditions are implemented along the conduit walls.

The adimensional numbers for each simulation are reported in Table 3. Velocities for determining  $Re$  and  $Fr$  are computed when the bubble/slug is moving in a steady state after a transient due to the initial conditions. These velocity values  $U$  are also used for normalizing times:

$$t^* = t \frac{U}{L}, \quad (2)$$

where  $L$  is the distance between the top of the bubble and the magma/air interface. So (2) represents a normalization for the virtual time the bubble needs to reach the surface in an ideal steady motion. We can also normalize distances:

$$x^* = x \frac{1}{D}, \quad (3)$$

where  $D$  is the conduit diameter. We can define then normalized velocities as:

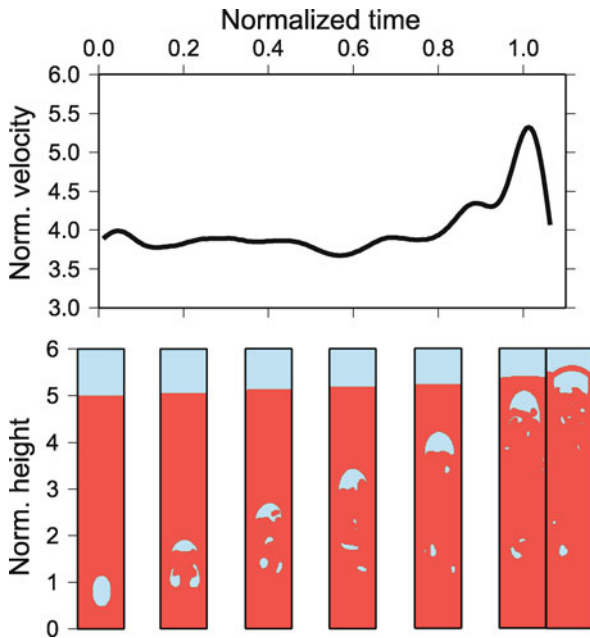
$$v^* = \frac{\partial x^*}{\partial t^*}. \quad (4)$$

In our simulations we have also defined an isothermal expansion ratio  $\gamma$  which is the ratio between the bubble volume at the initial conditions and its volume at (simulated) atmospheric pressure.

### Bubble Ascent

We first consider the ascent of a single bubble in a vertical conduit. In the first simulation (1A in Table 3) the initial shape of the bubble is an ellipse (see Fig. 2). As the simulation evolves the bubble becomes unstable splitting into three smaller bubbles. The two smaller lateral bubbles are embedded in a symmetric vortex structure. As the flow evolves the symmetry is broken and a turbulent wake develops behind the largest bubble. This one rises with an almost steady velocity (Fig. 2 top) with a spherical cap shape. In the final part of the simulation, the bubble suffers an expansion that made the bubble accelerate ( $t > 0.8$ ) and causes the flow to become transitional toward the slug flow regime. Then the bubble reaches the surface and a curved liquid film develops above the bubble just before bursting. This phase is accompanied by a sudden deceleration of the bubble ascent ( $t > 1.0$ ). The pressure variations in the conduit are modest, with a lack of low frequency oscillations. Also the pressure transients generated by the bubble burst are limited (Fig. 3).

The adimensional number for this simulation (Table 3) is in a range related to transitional flow regimes typical of basaltic systems [9,17]. An increase of an order of magnitude of the viscosity (1B in Table 3) makes the flow still in the transitional regime (Table 3). The high values of



Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Figure 2

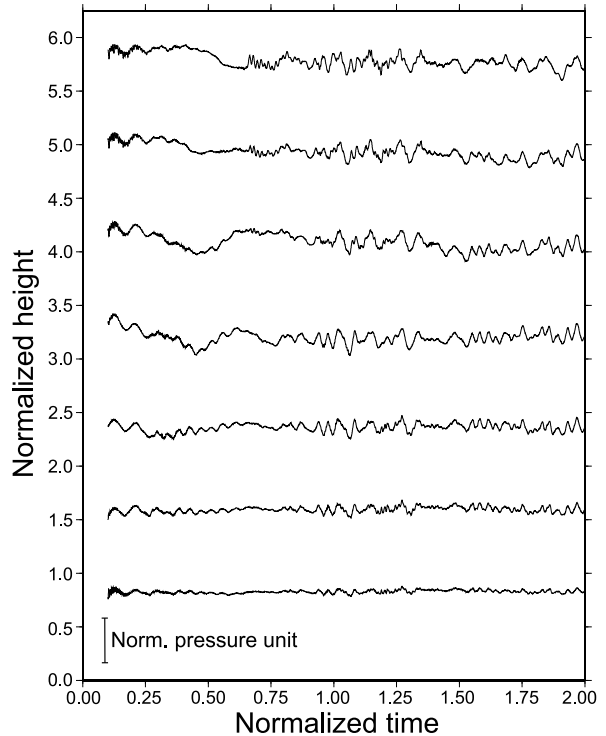
Ascent and bursting of a gas bubble (simulation 1A in Table 3). Time, height and velocity are normalized according to (2), (3) and (4)

the Eotvos number indicates that the surface tension effect is not relevant in the flow dynamics [7] in both cases.

### Slug Ascent in a Vertical Pipe

In these simulations (2A and 2B in Table 3) we model the ascent of a single slug in a vertical pipe. Again we start from an elliptical shape of the bubble. After an initial transient the typical slug shape develops (Fig. 4) and a turbulent wake appears behind the slug. As the slug rises it suffers a volume expansion. In the slug flow regime the liquid is pushed upward causing an overall increase of the hydrostatic pressure in the conduit. After the bursting of the slug the liquid film on the wall falls down to the original hydrostatic level. The slug expansion causes a significant acceleration of the liquid for  $t > 1.0$  (Fig. 4). The velocity drops as the slug reaches the bursting point and the liquid film develops.

The major feature in the pressure pattern is the slow ramp-like increase followed by a drop (Fig. 5). The increase reflects the rise of the hydrostatic head above the slug while the gradual decrease is related to the passage of the slug. The bursting of the slug generates a moderate pressure transient in the conduit (marker “T” in Fig. 5) and damped resonant oscillations in the uppermost part of the pipe, filled with gas (above norm. height 11 in Fig. 5).



Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Figure 3

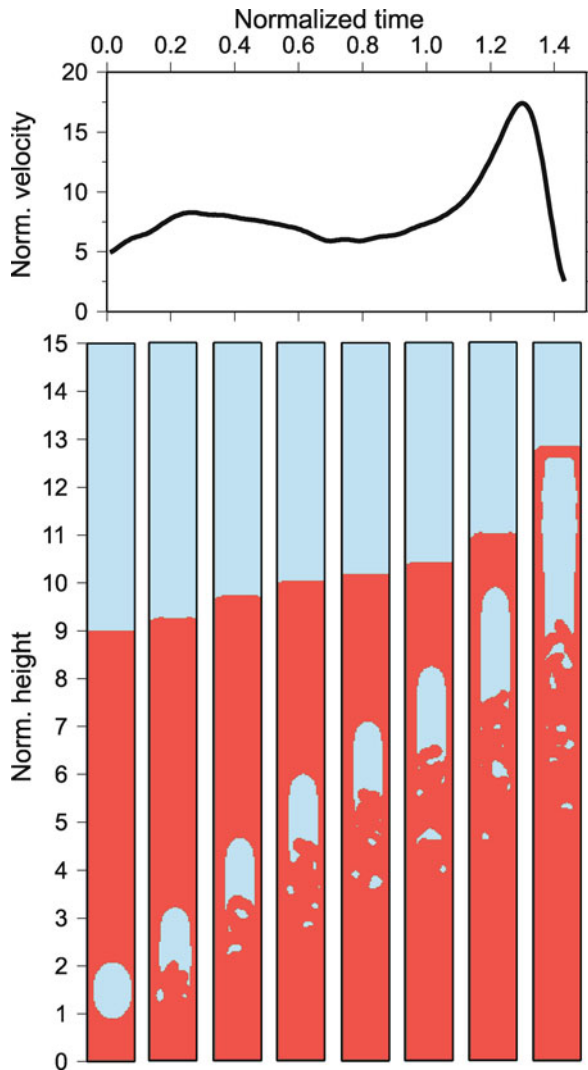
Time series of pressure variations for simulation 1A. The represented pressure values are normalized by the external atmospheric pressure  $P_0$ . Each time series represents the recording of a virtual sensor located at the center of the conduit at the height of its first point (see scale on the left)

The simulated pressure pattern fits very well the observations of analogue modeling [9]. Both the pressure increase linked to the rise of the hydrostatic head and the pressure drop related to the passage of the slug are in good agreement. Together with these long period variations they observe also oscillatory pressure transients just before the slug approaches the surface and after the bursting. The simulations presented here lack this feature because of an intrinsic limit in the numerical method that does not model explicitly the interface (see Appendix).

The overall behavior of simulation 2B, where a tenfold increase of the liquid viscosity (Table 3) is very similar to 2A. The only remarkable difference is the lack of high frequency oscillations in the pressure (as the transient “T” in Fig. 5) due to the greater viscous damping effect.

### Slug Ascent in a Pipe with a Flare

In this set of simulations (3A and 3B in Table 3) we model again the ascent of a single slug in a pipe. However in this

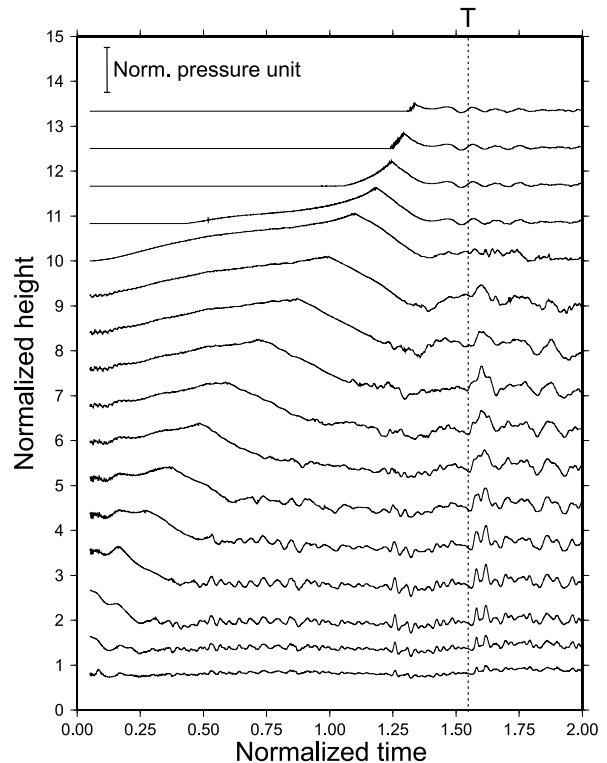


Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Figure 4

Ascent and bursting of a single gas slug in a straight pipe (simulation 2A in Table 3)

case the width of the pipe doubles after a norm. height of 10. In the first part of the simulation, the behavior is similar to the one illustrated in Sect. “[Slug Ascent in a Vertical Pipe](#)”. As the slug nose enters in the flare (Fig. 6) it starts to expand rapidly making the fluid accelerate upward. The expansion of the slug in the flare is followed by its breakup because of the development of strong turbulence. The velocity of the top of the slug increases until the slug passes through the flare then it drops because of the change in the conduit diameter.

The pressure pattern shows a major difference, compared with simulation 2A (Fig. 7). As the slug passes

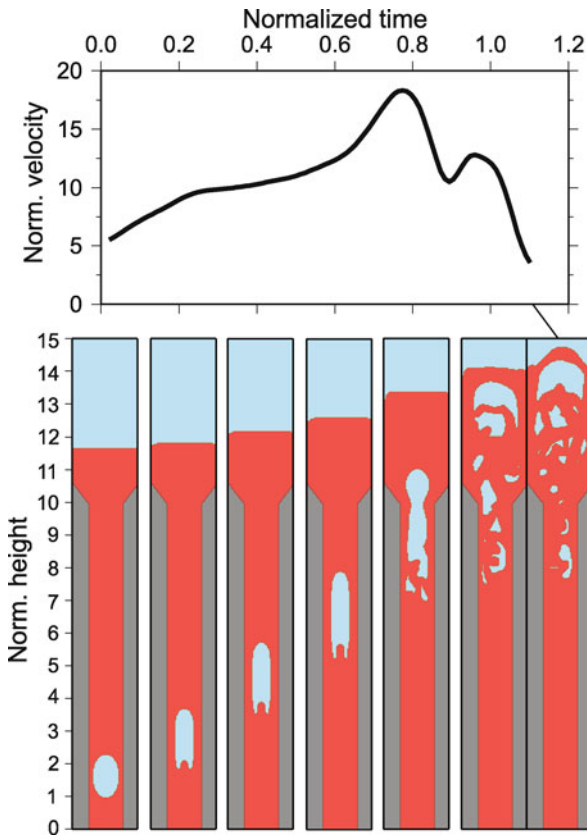


Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Figure 5

Time series of pressure variations for simulation 2A.  $T$  marks the most significant transient in the pressure patterns

through the flare there is a sudden pressure increase recorded along the whole conduit. The pressure then drops when the slug expansion has terminated. This strong pressure pulse is clearly related to the varying fluid flow regime as the slug enters the upper conduit and it deserves a closer analysis. In Fig. 8 we have represented the fluid dynamic regime during the slug expansion together with the related pressure variations. We observe that the pressure rises when most of the slug has passed through the flare. The strong acceleration induces turbulence both in the lower and the upper sections of the conduit. The upward acceleration first causes the disruption of the lower part of the slug leaving behind the main bubble a set of smaller ones, whose motion is driven by the fluid turbulence. When the slug has entered in the upper conduit it suffers another splitting due both to the induced vorticity and to the change in the boundary conditions. The behavior of the slug for  $t^* > 0.8$  is characterized by a flow regime with a smaller value of  $\lambda$  and so it is similar to the initial part ( $t^* < 0.2$ ) of simulation 1A.

Similar fluid dynamics behavior and pressure patterns have been observed in analogue simulations by [10]. In



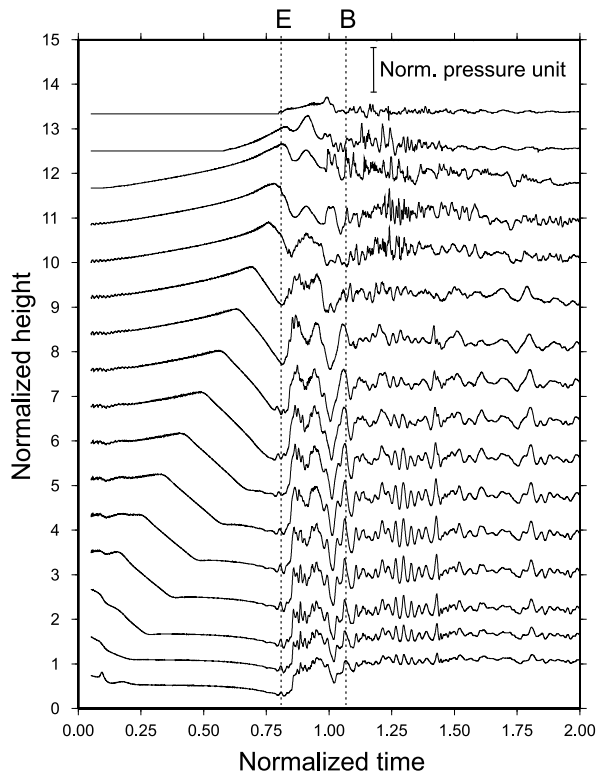
Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Figure 6  
 Ascent and bursting of a single gas slug in a pipe with a flare (simulation 3A in Table 3)

particular, the breakup of the slug and the generation of a positive pressure pulse as it passes through the flare are observed in analogue models within a wide range of fluid viscosities and conduit widenings.

As in the previous case the simulation with higher liquid viscosity (3B in Table 3) shows a similar behavior. Higher viscosities reduce the fluid vorticity leading to less fragmented slugs.

**Seismological Constraints on Numerical Models**

Seismological data analysis is a powerful tool for putting constraints on the geometry and the dynamics of volcanic systems. Non-stationary fluid flow in volcanic conduits generates pressure variations on the conduit walls and then seismic waves propagating toward the Earth’s surface where they are recorded by seismometers. The frequency band of seismic signals recorded in volcanic areas spans a wide range: from the Ultra-Long-Period (ULP)



Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Figure 7  
 Time series of pressure variations for simulation 3A. E marks the entering of the slug in the upper conduit while B marks the bursting of the main bubble at the magma surface

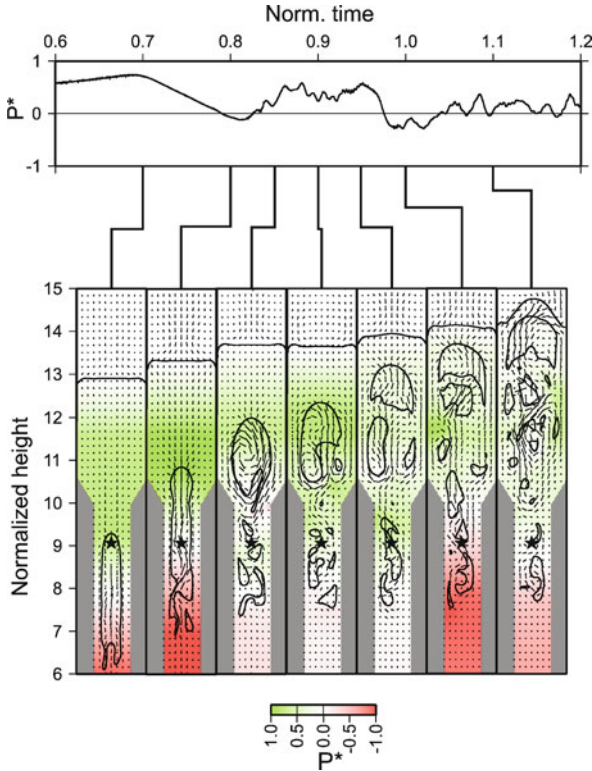
band ( $>10^2$  s), the Very-Long-Period (VLP) band ( $10^2-10^0$  s) and the Long-Period (LP) band ( $10^0-10^1$  s) [5].

From a formal point of view this can be expressed considering the elastodynamic field generated by a general extended source, whose external surface is  $\Sigma$ :

$$u_n(\mathbf{x}, t) = \iint_{\Sigma} [f_q * G_{np} + m_{pq} * G_{np,q}] d\Sigma, \quad (5)$$

where  $u_n(\mathbf{x}, t)$  is the  $n$ th component of the ground displacement recorded at the position  $\mathbf{x}$ ,  $f_q$  is the body-force distribution over  $\Sigma$ ,  $m_{pq}$  is the moment density tensor and  $G_{np}$  are the Green’s functions. In (5) we take into account all the details of the source dynamics. However the density of actual seismic networks is not sufficient for retrieving every minor feature of the seismic sources. Volcano monitoring networks usually have a limited extension ( $10^3-10^4$  m) and the wavelengths associated with ULP and VLP signals are higher. Therefore the analysis of such signals the seismic source (i. e. the part of the volcanic conduit responsible of seismic wave generation) can be represented





Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Figure 8

Detail of the conduit during the slug expansion phase. Each snapshot represents the normalized pressure variations (see the color scale on the bottom) and the local fluid velocity vectors. The thick contours are the bubble interfaces. The pressure values represented above are measured at the virtual sensor indicated by a black star in the snapshots

as a point [5]. Under the assumption of a point source we can express (5) as [5]:

$$u_n(\mathbf{x}, t) = F_q * G_{np} + M_{pq} * G_{np,q}, \quad (6)$$

where:

$$F_q = \iint_{\Sigma} f_q d\Sigma \quad (7)$$

and:

$$M_{pq} = \iint_{\Sigma} m_{pq} d\Sigma. \quad (8)$$

In common earthquake sources the single force component  $F_q$  is null. In volcanic sources, the acceleration of the center of mass of the fluid makes this component noteworthy.

The inversion of the recorder waveforms  $\mathbf{u}(\mathbf{x}, t)$ , after the numerical computation of the Green's functions  $\mathbf{G}$  allows the retrieval of the single force and moment tensor components of (6) [14].  $\mathbf{F}$  and  $\mathbf{M}$  are a synthesis of the force systems acting on volcanic conduits and then can be used for discriminating among numerical models on the basis of their fit with observations.

Numerical simulations provide the pressure tensor field  $\mathbf{P}$  over the whole computational domain. The pressure tensor can be used for computing the forces acting on each point of the conduit walls multiplying it for the unit vector normal to the wall  $\hat{\mathbf{n}}$ :

$$\mathbf{f} = \mathbf{P}\hat{\mathbf{n}}. \quad (9)$$

These values can be integrated numerically using an expression similar to (7) to get the single force component  $\mathbf{F}$ . Then using the definition of moment they can be used for retrieving also the equivalent moment tensor:

$$M_{pq} = \iint_{\Sigma} \left( f_q - \frac{F_q}{\Sigma} \right) r_p d\Sigma, \quad (10)$$

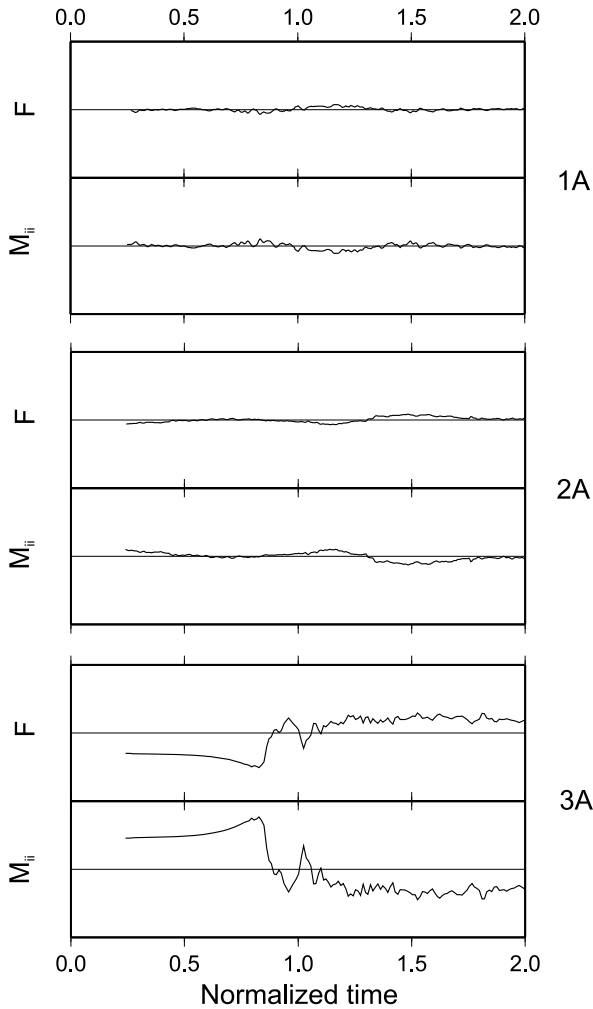
where  $\mathbf{r}$  is the arm formed by the force respect to an arbitrary origin  $\mathbf{O}$ .

In Fig. 9 we have represented the vertical force component and the moment tensor isotropic component (that is the trace  $M_{xx} + M_{zz}$ ) for the three simulations with lower liquid viscosity. The represented values are normalized.

In the first case (1A) we note the absence of significant signals. There are only minor oscillations related to fluid dynamics instability during the bubble ascent. The rise of a bubble smaller than the conduit dimension seems not to be able to generate VLP signals. Therefore a repetitive bursting of small bubbles at the magma/air interface can be responsible for the generation of continuous infrasonic and seismic tremor recorded in some basaltic volcanoes [16].

In the second case, again, the signal amplitudes are quite low. It is evident that there is an anticorrelation between the vertical force and the isotropic moment. The vertical acceleration of the slug ( $1.0 < t^* < 1.3$  in Fig. 4) causes a downward reaction force as well as an increase in the conduit pressure. The bursting ( $t^* > 1.3$ ) causes a downward acceleration of the liquid and a pressure decrease. A similar effect, although with a minor magnitude, is also evident in 1A (Fig. 9).

In the third case the signal amplitude increases dramatically, about five times higher than 2A. The slug entering in the upper, wider portion of the conduit is marked by a downward force and a positive moment. At the end of the expansion phase ( $t^* > 0.8$ ) there is a sudden inversion of the trends in both quantities, followed again by



Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Figure 9

Vertical force and isotropic component for simulations 1A, 2A and 3A of the moment tensor for simulations 1A, 2A and 3A (see Table 3)

a positive downward peak of the force and positive of the momentum ( $1.0 < t^* < 1.1$ ). This peak occurs when two pieces of the original slug coalesce again (Fig. 8) leading to a sudden, but limited in time, upward acceleration. In real cases the dynamics can be even more complex with multiple slugs or slug fragments entering in the upper conduit and interacting with each other. In this case during the main expansion phase, the lower conduit is filled almost exclusively with liquid (Fig. 6). If we suppose the simultaneous presence of many vertically aligned slugs in the lower conduit, then the pressure variations induced by the expansion of the topmost one would influence the lower ones in a complex non-linear mechanism still to investigate.

The signals presented in Fig. 9 need to be scaled to an actual time scale to be compared with real seismic signals source functions. In basaltic volcanoes the dimensions of the upper conduit is of the order of  $10^2$  m while the ascent velocities of the slugs are of the order of  $10^0$ – $10^1$  m/s. This gives a scaling factor of about  $10^1$ – $10^2$  s. Since the simulated transients have a normalized duration of about  $10^{-1}$  the actual simulated signals should have a characteristic period of  $10^0$ – $10^1$  s. These values are within the range of VLP and LP signals [5].

The patterns observed in simulation 3A (Fig. 9) closely matches the results obtained by [4] for the source function of VLP events at the Stromboli volcano. Thus this result, together with analogue simulations [10], strongly supports the hypothesis that VLP events at Stromboli are generated by the passage of a gas slug through a conduit widening and its subsequent expansion and bursting. The long ramp-like signals in both force and moment for  $t^* < 0.8$  in Fig. 9 has a characteristic period longer than the VLP signals recorded and analyzed by [4]. Recordings at Stromboli, using a seismic sensor with a wider frequency range [13] interestingly suggested in some seismic signals a similar ramp having a length of more than 60 s, compatible with our scaling. This long ramp is related to the slow and gradual increase in the magmatic head due to the continuous expansion of the ascending slug.

## Conclusions

The rapid expansion of the gas slugs in the uppermost part of the conduit plays a fundamental role, both in the eruptive dynamics and the seismic wave generation process. We have focused on the role of conduit geometry in the fluid dynamics of gas slug ascent and its implication in the generation of seismic signals. In one of the simulations (3A) we have shown that the system of forces acting on the conduit is able to generate seismic waves with a higher efficiency compared with other cases. This observation can be generalized to more complex geometry such as an alternating of widening and narrowing [10]. The passage of slugs can occur also in very complex conduits [17]. In this case it is possible that pressure transients are generated in different positions along the conduit.

## Future Directions

A deeper understanding of the relationship between slug ascent dynamics and seismic signals generation would require more advanced modeling techniques in various three-dimensional geometries, testing how the effect of changes in slug volumes and magma properties can affect the generation of seismic signals.

These studies are an important step toward more advanced seismic monitoring techniques of active basaltic volcanoes aimed at assigning in real time a volcanological meaning to variations in observed LP and VLP seismic signals.

## Appendix A – Fluid Dynamics of a Two-Phase System

### Definition of a Two-Phase System

A fluid two-phase system can be described using two scalar fields  $n_a$  and  $n_b$ , representing the local molar densities of the two components  $a$  and  $b$ . The actual local density  $\rho$  is then:

$$\rho = m_a n_a + m_b n_b, \quad (11)$$

where  $m_a$  and  $m_b$  are the molecular weight of the two phases. As it will be shown in the following, it is convenient to describe the system using an alternative representation based on the variables:

$$n = n_a + n_b, \quad (12)$$

and

$$\phi = n_a - n_b. \quad (13)$$

$\phi$  defines the local composition of the fluid. It can span the range  $[-n, +n]$  with the value  $-n$  representing a pure  $b$  composition and  $+n$  a pure  $a$  composition. The relation between  $n$ ,  $\phi$  and  $\rho$  is:

$$\rho = \frac{1}{2} [m_a (n + \phi) + m_b (n - \phi)]. \quad (14)$$

The values of  $m_a$  and  $m_b$  are set so that in reference conditions the value of  $n$  is always equal to 1. So for instance, if we consider pure water at ambient conditions  $m_w = 1000$  kg/mol.

### State Equation

We assume that pure phases obey to a simple isothermal ideal gas state equation like:

$$P = \rho c^2, \quad (15)$$

where  $P$  is the pressure and  $c$  is the sound speed. We assume that the pressure is zero in the reference state, so the state equation for the component  $a$  is:

$$P = m_a (n - 1) c_a^2. \quad (16)$$

A similar relation holds for  $b$ . Since along interfaces the composition varies continuously, we should define a state

equation for a mixture that satisfies two requirements: the state equation for pure phases must match (16) and the isobaric contours in the  $(\phi, n)$  plane must be straight lines. This second requirement follows from the consideration that diffusion processes between the components follows straight paths in this domain. In most of the actual Lattice Boltzmann literature, this question is not issued because the proposed implementations are almost isobaric ( $n \simeq 1$ ) [21,25,26]. A suitable choice for the state equation is then:

$$P(n, \phi) = \frac{1}{4} \left[ \Lambda + \sqrt{16m_a m_b c_a^2 c_b^2 (n - 1) + \Lambda^2} \right] \quad (17)$$

with:

$$\Lambda = m_a c_a^2 (n + \phi - 2) + m_b c_b^2 (n - \phi - 2). \quad (18)$$

In Fig. 11 an example contour plot of isobaric lines is shown. We can define another parameter  $\chi$  as:

$$\chi(n, \phi) = -1 + 2 \sqrt{\frac{(\phi + B)^2 + (n - B)^2}{(A + B)^2 + (A - B)^2}}, \quad (19)$$

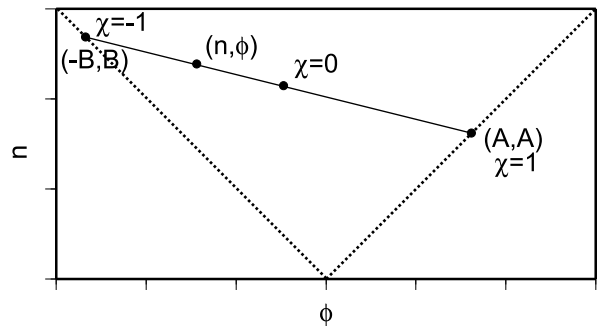
where:

$$A = \frac{P(n, \phi)}{m_a c_a^2} + 1, \quad (20)$$

and

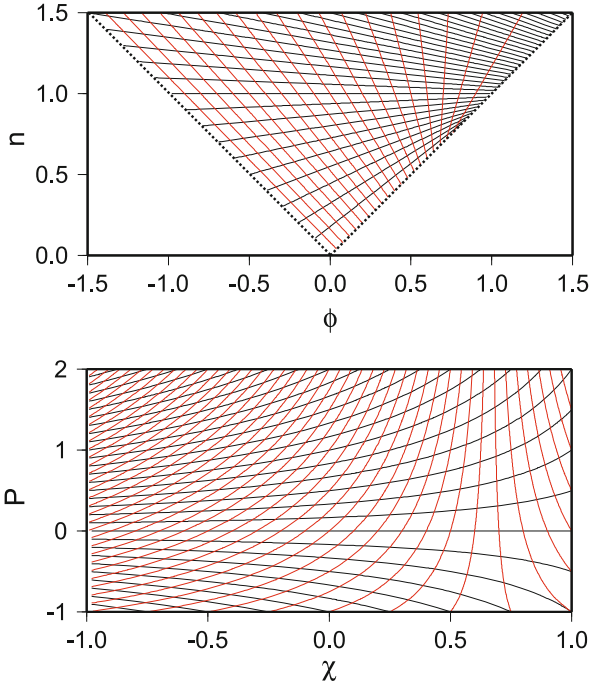
$$B = \frac{P(n, \phi)}{m_b c_b^2} + 1. \quad (21)$$

The actual meaning of  $\chi$  is shown in Fig. 10. It is the rescaled ratio of the distance between the points  $(\phi, n)$  and  $(-B, B)$  and the distance between the points  $(A, A)$  and  $(-B, B)$ . So  $\chi = -1$  along the line  $\phi = -n$  and  $\chi = 1$



Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Figure 10

Definition of  $\chi$ . See text for details



Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Figure 11

In the upper panel, isolines of  $\chi$  (red lines) and  $P$  (black lines) are represented in the  $(n, \phi)$  domain. In the lower panel isolines of  $n$  (black lines) and  $\phi$  (red lines) are represented in the  $(\chi, P)$  domain. These plots represent a two-phase system where  $m_a c_a^2 = 5$  and  $m_b c_b^2 = 1$

along  $\phi = n$ . In practice  $\chi$  is related to the local composition of the fluid. In Fig. 10 the relation between  $\phi, n, P$  and  $\chi$  is represented. This relation can be inverted and the system  $(\chi, P)$  can be used as well as the  $(\phi, n)$  coordinate system. In the following we show that the former is the best choice for describing chemical equilibria.

### Chemical Equilibria and the Diffuse Interface Theory

The chemical equilibrium between the two phases can be described using a physically consistent thermodynamic approach [2,8,26]. We define a Ginzburg–Landau free energy functional for an heterogeneous mixture as [15]:

$$\mathcal{F} = \int \left[ \psi(n, \phi) + \frac{\kappa_n}{2} (\nabla n)^2 + \frac{\kappa_\phi}{2} (\nabla \phi)^2 \right] dV, \quad (22)$$

where  $\kappa_n$  and  $\kappa_\phi$  are parameters related to the surface tension. The system reaches its final thermodynamical equilibrium making the free energy minimum.

We can rewrite (22) in the  $(\chi, P)$  representation:

$$\mathcal{F} = \int \left[ \psi(\chi, P) + \frac{\kappa_\chi}{2} (\nabla \chi)^2 \right] dV. \quad (23)$$

Note that we have set  $\kappa_P = 0$ . This implies that surface tension effects are independent of the absolute pressure. In this paper we set:

$$\psi(\chi, P) = \alpha \left( \frac{\chi^4}{4} - \frac{\chi^2}{2} \right) + P \ln \frac{P}{P_0}, \quad (24)$$

where  $P_0$  is an arbitrary reference pressure value. Using this definition of  $\psi$  we note that this function has two minima for  $\chi = \pm 1$ , corresponding to the two equilibrium compositions. Starting from an heterogeneous mixture, the system evolves by creating domains having a quite homogeneous composition separated by interfaces where the composition varies smoothly. The equilibrium is reached when the chemical potential:

$$\mu = \frac{\delta \mathcal{F}}{\delta \chi}, \quad (25)$$

is everywhere equal to zero [2,26]. So, in our case, using (23) and (24) we obtain the condition of chemical equilibrium for a plane interface orthogonal to the  $x$ -direction:

$$\mu = \frac{\partial \psi}{\partial \chi} - \kappa_\chi \frac{\partial^2 \chi}{\partial x^2}. \quad (26)$$

Assuming the boundary conditions  $\psi(x) = 0$  and  $\frac{\partial \chi}{\partial x} = 0$  for  $x = \pm \infty$  and  $\chi(0) = 0$ , the previous ODE can be integrated giving the expression of the spatial variation of the composition through an equilibrium interface:

$$\chi(x) = \tanh \left( \frac{2x}{\xi} \right), \quad (27)$$

where the interface thickness  $\xi$  is the width where the 96% of variation occurs. Its value is:

$$\xi = 2 \sqrt{-\frac{2\kappa_\chi}{\alpha}}. \quad (28)$$

On the basis of the definition of surface tension  $\sigma$  [25]:

$$\sigma = \int_{-\infty}^{+\infty} \mathcal{F}(x) dx, \quad (29)$$

we can write:

$$\sigma = \frac{2}{3} \sqrt{-2\alpha\kappa_\chi}. \quad (30)$$

Using (28) and (30) we can also retrieve useful inverse relations:

$$\alpha = 3 \frac{\sigma}{\xi}, \quad (31)$$

and

$$\kappa_\chi = \frac{3}{8} \sigma \xi. \quad (32)$$

We should emphasize that in real physical systems expression (27) describes an actual interface having a thickness whose order of magnitude is of molecular scale. In this work we use the diffuse interface theory as a numerical tool for modeling of two-phase systems. In other words, we use unphysical interfaces having a macroscopic thickness (usually  $10^{-3} \div 10^{-1} m$ ).

### Cahn–Hilliard and Mass Conservation Equations

The time evolution under non-equilibrium conditions can be expressed by two Cahn–Hilliard equations [2,26]:

$$\frac{Dn}{Dt} = \Gamma \nabla^2 \mu_n, \quad (33)$$

$$\frac{D\phi}{Dt} = \Gamma \nabla^2 \mu_\phi. \quad (34)$$

These equations are similar to common advection-diffusion equations with  $\Gamma$  being a diffusion coefficient and the operator  $D/Dt$  the substantial derivative. The explicit expressions for the chemical potentials  $\mu_n$  and  $\mu_\phi$  are:

$$\mu_n = \frac{\delta \mathcal{F}}{\delta n} = \frac{\delta \mathcal{F}}{\delta \chi} \frac{\partial \chi}{\partial n} = \mu_\chi \frac{\partial \chi}{\partial n}, \quad (35)$$

$$\mu_\phi = \frac{\delta \mathcal{F}}{\delta \phi} = \frac{\delta \mathcal{F}}{\delta \chi} \frac{\partial \chi}{\partial \phi} = \mu_\chi \frac{\partial \chi}{\partial \phi}. \quad (36)$$

with:

$$\mu_\chi = \frac{\delta \mathcal{F}}{\delta \chi} = \alpha \chi (\chi^2 - 1) - \kappa_\chi \nabla^2 \chi. \quad (37)$$

Then the Cahn–Hilliard Eqs. (33) and (34) can be rewritten as:

$$\frac{D}{Dt} \begin{pmatrix} n \\ \phi \end{pmatrix} = \Gamma \nabla^2 \mu_\chi \begin{pmatrix} \frac{\partial \chi}{\partial n} \\ \frac{\partial \chi}{\partial \phi} \end{pmatrix} \quad (38)$$

On the basis of the definitions of  $n$  (12) and  $\phi$  (13), we can state that the previous equation expresses the mass transfer of components  $a$  and  $b$  because of chemical disequilibria. Since we are dealing with a compressible flow, we should obviously account for this in the mass balances. Then, following basic fluid dynamics [1], we can rewrite (38) in explicit form as:

$$\frac{\partial n}{\partial t} = -\mathbf{v} \nabla n - n \nabla \cdot \mathbf{v} + \frac{\partial \chi}{\partial n} \Gamma \nabla^2 \mu_\chi \quad (39)$$

$$\frac{\partial \phi}{\partial t} = -\mathbf{v} \nabla \phi - \phi \nabla \cdot \mathbf{v} + \frac{\partial \chi}{\partial \phi} \Gamma \nabla^2 \mu_\chi. \quad (40)$$

### Thermodynamic Pressure Tensor

It can be shown, from statistical mechanics that the pressure tensor can be obtained from (23) [8,15]:

$$P_{ij}^{th} = p_0 \delta_{ij} + \kappa_\chi \frac{\partial \chi}{\partial x_i} \frac{\partial \chi}{\partial x_j}, \quad (41)$$

where the isotropic component is:

$$p_0 = P \frac{\delta \mathcal{F}}{\delta P} + \chi \frac{\delta \mathcal{F}}{\delta \chi} - (\psi(\chi, P) + \frac{\kappa_\chi}{2} (\nabla \chi)^2). \quad (42)$$

so:

$$p_0 = P + \alpha \left( \frac{3}{4} \chi^4 - \frac{1}{2} \chi^2 \right) - \kappa_\chi (\nabla^2 \chi) - \frac{\kappa}{2} (\nabla \chi)^2 \quad (43)$$

The tensor  $\mathbf{P}^{th}$  describes the stresses induced by spatial variation in density and composition and has to be added to the viscous stress tensor  $\boldsymbol{\tau}$ .

### Conservation Equations

Let us now apply the results of the previous section for building a system of fluid dynamics equations suitable for a numerical implementation. Together with the mass conservation Eqs. (39) and (40) we need the conservation equation for the linear momentum [1]:

$$\rho \frac{d\mathbf{v}}{dt} = \nabla \cdot \mathbf{P} + \rho \mathbf{g}, \quad (44)$$

where  $\mathbf{P} = -\mathbf{P}^{th} + \boldsymbol{\tau}$  is the full pressure tensor of (41),  $\mathbf{g}$  is the gravity and  $\boldsymbol{\tau}$  is the viscous stress tensor that, for a Newtonian fluid is:

$$\tau_{ij} = \lambda e_{kk} \delta_{ij} + 2\mu e_{ij}, \quad (45)$$

with  $\lambda$  and  $\mu$  the bulk and shear viscosities and  $e_{ij}$  is the strain rate tensor expressed by:

$$e_{ij} = \frac{1}{2} \left( \frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i} \right). \quad (46)$$

## Appendix B – Numerical Implementation

### Transformed Equations

A common problem in the numerical solution of the conservation Eqs. (33) and (34) is the output of values having no physical meaning (for instance negative densities for one of the components) with bad effects on the numerical stability of the code. This is due to the fact that both variables are defined over a limited range of values that are  $n \in [0, +\infty[$  and  $\phi \in [-n, +n]$ . These conditions are

not explicit in the conservation equations, but this problem can be overcome by making simple changes of variable:

$$n = e^l, \quad (47)$$

$$\phi = e^l \tanh q. \quad (48)$$

Using the new variable  $l$  instead of  $n$  the conservation Eq. (39) become:

$$\frac{\partial l}{\partial t} = -\mathbf{v} \cdot \nabla l - \nabla \cdot \mathbf{v} + \frac{1}{e^l} \left( \frac{\partial \chi}{\partial n} \right) \Gamma \nabla^2 \mu. \quad (49)$$

Substituting (48) in (40) we first obtain:

$$\begin{aligned} \frac{\partial q}{\partial t} = & -\mathbf{v} \cdot \nabla q - \frac{1}{2} \sinh 2q \left( \frac{\partial l}{\partial t} + \mathbf{v} \cdot \nabla l + \nabla \cdot \mathbf{v} \right) \\ & + \frac{1}{e^l} \cosh^2 q \left( \frac{\partial \chi}{\partial \phi} \right) \Gamma \nabla^2 \mu. \end{aligned} \quad (50)$$

Using (49) and some algebraic manipulation the previous expression become:

$$\begin{aligned} \frac{\partial q}{\partial t} = & -\mathbf{v} \cdot \nabla q - \frac{1}{e^l} \cosh q \left[ \frac{\partial \chi}{\partial n} \sinh q + \frac{\partial \chi}{\partial \phi} \cosh q \right] \\ & \cdot \Gamma \nabla^2 \mu \end{aligned} \quad (51)$$

### Boundary and Initial Conditions

Equations (49) and (51) have to be solved setting proper boundary and initial conditions.

The no-slip boundary condition, implemented along the conduit walls is:

$$\mathbf{v} = 0. \quad (52)$$

Another boundary condition is set along the walls:

$$\nabla \phi \cdot \mathbf{n} = 0. \quad (53)$$

This is the neutral wetting condition [25] and it is needed in order to avoid the wall to behave sticky respect to one of the phases.

At the conduit top and bottom conditions of constant pressure are implementer. On the top the pressure is kept to a reference atmospheric value, while at the bottom it is kept at the hydrostatic pressure value, computed on the initial conditions.

In the initial conditions we set  $\mathbf{v} = 0$  and the computational domain is composed of domains having a homogeneous composition separated by smooth interfaces,

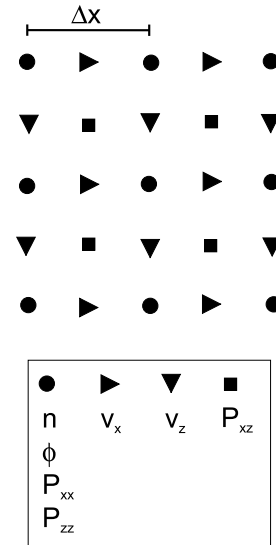
close to the equilibrium solution of (27). At the top of the model a gas domain represents the atmosphere, while the remaining part of the conduit is filled with magma. A small elliptical gas bubble in hydrostatic equilibrium is placed in the lower part of the conduit. The details are specified for each set of simulations.

### Finite Difference Implementation

Equations (49), (51) and (44) are discretized on a regular grid and the differential equations are transformed in finite-difference equations [19]. Scalar, vector and tensor quantities are discretized on staggered grids [6] (Fig. 12). Scalar quantities ( $l$ ,  $q$ ,  $\mu$  and  $P_{ii}$ ) are defined at integer grid steps ( $x_i$ ,  $z_j$ ). Velocities are staggered half grid step along  $x$  and  $z$  directions, that is  $v_x$  is defined at  $(x_{i+1/2}, z_j)$  and  $v_z$  at  $(x_i, z_{j+1/2})$ . The deviatoric part of the pressure tensor  $P_{xz}$  is defined at grid points  $(x_{i+1/2}, z_{j+1/2})$ . This allows a second-order accuracy in the computation of spatial derivative.

We apply also a staggering in time. At the time step  $k$  we first solve (49) and (51). Then the current values of  $l$  and  $q$  are updated:

$$l_{i,j}^{(k)} = l_{i,j}^{(k-1)} + \frac{dl^{(k)}}{dt} \Delta t, \quad (54)$$



Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Figure 12

**Spatial scheme of the staggered grid used in the computation.** Circle represents the isotropic grid (density, composition and isotropic pressure), rightward triangles are the grids for  $v_x$  while downward triangles are for  $v_z$ . Squares are the grids used for discretizing  $P_{xz}$

$$q_{i,j}^{(k)} = q_{i,j}^{(k-1)} + \frac{dq}{dt}^{(k)} \Delta t. \quad (55)$$

Using these new values the full pressure tensor  $\mathbf{P}^{(k)}$  is computed using the discretized version of (41), (42) and (43). At the time step  $k + 1/2$  (44) is solved and the values of the local velocities are updated.

The computation of  $\frac{dl}{dt}^{(k)}$  and  $\frac{dq}{dt}^{(k)}$  requires the values of  $l^{(k-1)}$ ,  $q^{(k-1)}$  and  $v^{(k-1/2)}$ . On the other hand the computation of  $v^{(k+1/2)}$  requires the values of  $l^{(k)}$ ,  $q^{(k)}$  and  $v^{(k-1/2)}$ .

The time step  $\Delta t$  is chosen in order to satisfy a stability condition. Since we are dealing with isothermal flows with a very low Mach number, the condition can be expressed by:

$$\Delta x < \Delta t c_{\max}, \quad (56)$$

where  $c_{\max}$  is the highest sound speed among the two components. In our simulations we have set:

$$\Delta x = 6 \Delta t c_{\max}, \quad (57)$$

## Bibliography

### Primary Literature

1. Aris R (1962) Vectors, tensors and the basic equation of fluid mechanics. Dover Publications
2. Cahn JW, Hilliard JE (1958) Free energy of a nonuniform system. i. interfacial free energy. *J Chem Phys* 28(2):258–267
3. Chouet B (1986) Dynamics of a fluid-driven crack in three dimensions by the finite difference method. *J Geophys Res* 91:13967–13992
4. Chouet B, Dawson P, Ohminato T, Martini M, Saccorotti G, Giudicepietro F, De Luca G, Milana G, Scarpa R (2003) Source mechanisms of explosions at stromboli volcano, italy, determined from moment-tensor inversions of very-long-period data. *J Geophys Res* 108(B1)
5. Chouet BA (1996) New methods and future trends in seismological volcano monitoring. In: Scarpa R, Tilling RI (eds) Monitoring and mitigation of volcano hazards. Springer
6. Chung TJ (2002) Computational fluid dynamics. Cambridge University Press, Cambridge
7. Clift R, Grace JR, Weber ME (1978) Bubbles, drops and particles. Dover Publications
8. Evans R (1979) The nature of the liquid-vapour interface and other topics in the statistical mechanics of non-uniform, classical fluids. *Adv Phys* 28(2):143–200
9. James MR, Lane SJ, Chouet B, Gilbert JS (2004) Pressure changes associated with the ascent and bursting of gas slugs in liquid-filled vertical and inclined conduits. *J Volc Geotherm Res* 129:61–82
10. James MR, Lane SJ, Chouet BA (2006) Gas slug ascent through changes in conduit diameter: Laboratory insight into a vol-

cano-seismic source process in low-viscosity magmas. *J Geophys Res* 111

11. Jaupart C (2000) Magma ascent at shallow levels. In: Sigurdsson H (ed) Encyclopedia of Volcanoes. Academic Press
12. Jaupart C, Vergnolle S (1988) Laboratory models of hawaiian and strombolian eruptions. *Nature* 331:58–60
13. Kirchdorfer M (1999) Analysis and quasistatic fe modeling of long period impulsive events associated with explosions at stromboli volcano (italy). *Annali di Geofisica* 42(3):379–390
14. Ohminato T, Chouet BA, Dawson P, Kedar S (1998) Waveform inversion of very long period impulsive signals associated with magmatic injection beneath kilauea volcano, hawaii. *J Geophys Res* 103(B10):23839–23862
15. Pooley CM, Kuksenok O, Balazs AC (2005) Convection-driven pattern in phase-separating binary fluids. *Phys Rev E* 71:030501(R)
16. Ripepe M, Poggi P, Braun T, Gordeev E (1996) Infrasonic waves and volcanic tremor at stromboli. *Geophys Res Lett* 23(2):181–184
17. Seyfried R, Freundt A (2000) Experiments on conduit flow and eruption behaviour of basaltic volcanic eruptions. *J Geophys Res* B10(B10):23727–23740
18. Sparks RSJ (1978) The dynamics of bubble formation and growth in magmas: a review and analysis. *J Volc Geotherm Res* 3:137–186
19. Strikwerda JC (2004) Finite difference schemes and partial differential equations. SIAM, Philadelphia
20. Succi S (2001) The Lattice Boltzmann Equation for fluid dynamics and beyond. In: Numerical mathematics and scientific computation. Oxford University Press, Oxford
21. Swift MR, Orlandini E, Osborn WR, Yeomans JM (1996) Lattice Boltzmann simulations of liquid-gas and binary fluid system. *Phys Rev E* 54(5):5041–5052
22. Tomiyama A, Takagi S, Matsumoto Y (1999) Numerical simulation of bubble flows using interface tracking and bubble tracking method. *Trans Model Simul* 23
23. Vergnolle S, Brandeis G (1996) Strombolian explosions 1. a large bubble breaking at the surface of a lava column as a source of sound. *J Geophys Res* 101(B9):20433–20447
24. Vergnolle S, Mangan M (2000) Hawaiian and strombolian eruptions. In: Sigurdsson H (ed) Encyclopedia of Volcanoes. Academic Press
25. Xu A, Gonnella G, Lamura A (2003) Phase-separating binary fluids under oscillatory shear. *Phys Rev E* 67
26. Yue P, Feng JJ, Liu C, Shen J (2004) A diffuse-interface method for simulating two-phase flows of complex fluids. *J Fluid Mech* 515:293–317

### Books and Reviews

- Brennen CE (2005) Fundamentals of multiphase flow. Cambridge University Press, Cambridge
- Scarpa R, Tilling RI (eds) (1996) Monitoring and mitigation of volcano hazards. Springer
- Sigurdsson H, Bruce Houghton BF, McNutt SR, Rymer H, Stix J (eds) (2000) In: Encyclopedia of Volcanoes. Academic Press
- Tannehill JC, Anderson DA, Pletcher RH (1997) Computational fluid mechanics and heat transfer, second edn. Taylor and Francis, London
- Zobin V (2003) Introduction to volcanic seismology. Elsevier Science

# Solitons, Tsunamis and Oceanographical Applications of

M. LAKSHMANAN

Center for Nonlinear Dynamics, Bharathidasan University, Tiruchirapalli, India

## Article Outline

Glossary

Definition of the Subject

Introduction

Shallow Water Waves and KdV Type Equations

Deep Water Waves and NLS Type Equations

Tsunamis as Solitons

Internal Solitons

Rossby Solitons

Bore Solitons

Future Directions

Bibliography

## Glossary

**Soliton** A class of nonlinear dispersive wave equations in (1+1) dimensions having a delicate balance between dispersion and nonlinearity admit localized solitary waves which under interaction retain their shapes and speeds asymptotically. Such waves are called solitons because of their particle like elastic collision property. The systems include Korteweg–de Vries, nonlinear Schrödinger, sine-Gordon and other nonlinear evolution equations. Certain (2+1) dimensional generalizations of these systems also admit soliton solutions of different types (plane solitons, algebraically decaying lump solitons and exponentially decaying dromions).

**Shallow and deep water waves** Considering surface gravity waves in an ocean of depth  $h$ , they are called shallow-water waves if  $h \ll \lambda$ , where  $\lambda$  is the wavelength (or from a practical point of view if  $h < 0.07\lambda$ ). In the linearized case, for shallow water waves the phase speed  $c = \sqrt{gh}$ , where  $g$  is the acceleration due to gravity. Water waves are classified as deep (practically) if  $h > 0.28\lambda$  and the corresponding wave speed is given by  $c = \sqrt{g/k}$ ,  $k = \frac{2\pi}{\lambda}$ .

**Tsunami** Tsunami is essentially a long wavelength water wave train, or a series of waves, generated in a body of water (mostly in oceans) that vertically displaces the water column. Earthquakes, landslides, volcanic eruptions, nuclear explosions and impact of cosmic bodies can generate tsunamis. Propagation of tsunamis is in many cases in the form of shallow wa-

ter waves and sometimes can be of the form of solitary waves/solitons. Tsunamis as they approach coastlines can rise enormously and savagely attack and inundate to cause devastating damage to life and property.

**Internal solitons** Gravity waves can exist not only as surface waves but also as waves at the interface between two fluids of different density. While solitons were first recognized on the surface of water, the commonest ones in oceans actually happen underneath, as internal oceanic waves propagating on the pycnocline (the interface between density layers). Such waves occur in many seas around the globe, prominent among them being the Andaman and Sulu seas.

**Rossby solitons** Rossby waves are typical examples of quasigeostrophic dynamical response of rotating fluid systems, where long waves between layers of the atmosphere as in the case of the Great Red Spot of Jupiter or in the barotropic atmosphere are formed and may be associated with solitonic structures.

**Bore solitons** The classic bore (also called mascaret, poroca and aeger) arises generally in funnel shaped estuaries that amplify incoming tides, tsunamis or storm surges, the rapid rise propagating upstream against the flow of the river feeding the estuary. The profile depends on the Froude number, a dimensionless ratio of inertial and gravitational effects. Slower bores can take on oscillatory profile with a leading dispersive shock-wave followed by a train of solitons.

## Definition of the Subject

Surface and internal gravity waves arising in various oceanographic conditions are natural sources where one can identify/observe the generation, formation and propagation of solitary waves and solitons. Unlike the standard progressive waves of linear dispersive type, solitary waves are localized structures with long wavelengths and finite energies and propagate without change of speed or form and are patently nonlinear entities. The earliest scientifically recorded observation of a solitary wave was made by John Scott Russel in August 1834 in the Union Canal connecting the Scottish cities of Glasgow and Edinburgh. The theoretical formulation of the underlying phenomenon was provided by Korteweg and de Vries in 1895 who deduced the now famous Korteweg–de Vries (KdV) equation admitting solitary wave solutions. With the insightful numerical and analytical investigations of Martin Kruskal and coworkers in the 1960s the KdV solitary waves have been shown to possess the remarkable property that under collision they pass through each other without change of shape or speed except for a phase shift and so they are



solitons. Since then a large class of soliton possessing nonlinear dispersive wave equations such as the sine-Gordon, modified KdV (MKdV) and nonlinear Schrödinger (NLS) equations occurring in a wide range of physical phenomena have been identified.

Several important oceanographic phenomena which correspond to nonlinear shallow water wave or deep water wave propagation have been identified/interpreted in terms of soliton propagation. These include tsunamis, especially earthquake induced ones like the 1960 Chilean or 2004 Indian Ocean earthquakes, internal solitary waves arising in stratified stable fluids such as the ones observed in Andaman or Sulu seas, Rossby waves including the Giant Red Spot of Jupiter and tidal bores occurring in estuaries of rivers. Detailed observations/laboratory experiments and theoretical formulations based on water wave equations resulting in the nonlinear evolution equations including KdV, Benjamin–Ono, Intermediate Long Wave (ILW), Kadomtsev–Petviashvili (KP), NLS, Davey–Stewartson (DS) and other equations clearly establish the relevance of soliton description in such oceanographic events.

## Introduction

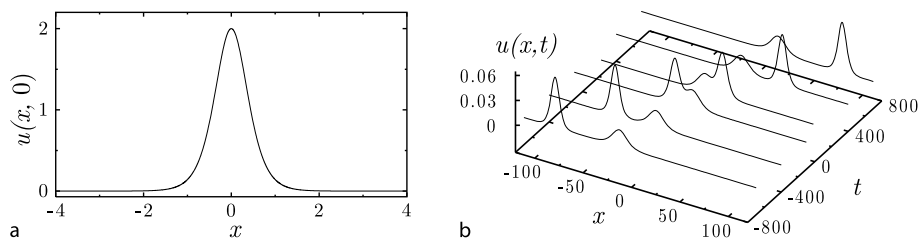
Historically, the remarkable observation of John Scott Russel [1,2] of the solitary wave in the Union Canal connecting the cities of Edinburgh and Glasgow in the month of August 1834 may be considered as the precursor to the realization of solitons in many oceanographic phenomena. While riding on a horse back and observing the motion of boat drawn by a pair of horses which suddenly stopped but not so the mass of water it had set in motion, the wave (which he called the ‘Great Wave of Translation’) in the form of large solitary heap of water surged forward and travelled a long distance without *change of form or diminution of speed*. The wave observed by Scott Russel is nothing but a solitary wave having a remarkable staying power and a patently nonlinear entity. Korteweg and de Vries in 1895,

starting from the basic equations of hydrodynamics and considering unidirectional shallow water wave propagation in rectangular channels, deduced [3] the now ubiquitous KdV equation as the underlying nonlinear evolution equation. It is a third order nonlinear partial differential equation in (1+1) dimensions with a delicate balance between dispersion and nonlinearity. It admits elliptic function cnoidal wave solutions and in a limiting form exact solitary wave solution of the type observed by John Scott Russel thereby vindicating his observations and putting to rest all the controversies surrounding them.

It was the many faceted numerical and analytical study of Martin Kruskal and coworkers [4,5] which firmly established by 1967 that the KdV solitary waves have the further remarkable feature that they are solitons having elastic collision property (Fig. 1).

It was proved decisively that the KdV solitary waves on collision pass through each other except for a finite phase shift, thereby retaining their forms and speeds asymptotically as in the case of particle like elastic collisions. The inverse scattering transform (IST) formalism developed for this purpose clearly shows that the KdV equation is a completely integrable infinite dimensional nonlinear Hamiltonian system and that it admits multisoliton solutions [6,7,8,9]. Since then a large class of nonlinear dispersive wave equations such as the sine-Gordon (s-G), modified Korteweg–de Vries (MKdV), NLS, etc. equations in (1+1) dimensions modeling varied physical phenomena have also been shown to be completely integrable soliton systems [6,7,8,9]. Interesting (2+1) dimensional versions of these systems such as Kadomtsev–Petviashvili (KP), Davey–Stewartson (DS) and Nizhnik–Novikov–Veselov (NNV) equations have also been shown to be integrable systems admitting basic nonlinear excitations such as line (plane) solitons, algebraically decaying lump solitons and exponentially localized dromion solutions [7,8].

It should be noted that not all solitary waves are solitons while the converse is always true. An example of a solitary wave which is not a soliton is the one which oc-



Solitons, Tsunamis and Oceanographical Applications of, Figure 1

KdV equation: **a** One soliton solution (solitary wave) at a fixed time, say  $t = 0$ , **b** Two soliton solution (depicting elastic collision)

curs in double-well  $\lambda\phi^4$  wave equation which radiates energy on collision with another such wave. However even such solitary waves having finite energies are sometimes referred to as solitons in the condensed matter, particle physics and fluid dynamics literature because of their localized structure.

As noted above solitary waves and solitons are abundant in oceanographic phenomena, especially involving shallow water wave and deep water wave propagations. However, these events are large scale phenomena very often difficult to measure experimentally, rely mostly on satellite and other indirect observations and so controversies and differing interpretations do exist in ascribing exact solitonic or solitary wave properties to these phenomena. Yet it is generally realized that many of these events are closely identifiable with solitonic structures. Some of these observations deserve special attention.

### Tsunamis

When large scale earthquakes, especially of magnitude 8.0 in Richter scale and above, occur in seabeds at appropriate geological faults tsunamis are generated and can propagate over large distances as small amplitude and long wavelength structures when shallowness condition is satisfied. Though the tsunamis are hardly felt in the mid-sea, they take monstrous structures when they approach land masses due to conservation of energy. The powerful Chilean earthquake of 1960 [10] led to tsunamis which propagated for almost fifteen hours before striking Hawaii islands and a further seven hours later they struck the Japanese islands of Honshu and Hokkaido. More recently the devastating Sumatra–Andaman earthquake of 2004 in the Indian Ocean generated tsunamis which not only struck the coastlines of Asian countries including Indonesia, Thailand, India and Srilanka but propagated as far as Somalia and Kenya in Africa, killing more than a quarter million people. There is very good sense in ascribing soliton description to such tsunamis.

### Internal Solitons

Peculiar striations, visible on satellite photographs of the surface of the Andaman and Sulu seas in the far east (and in many other oceans around the globe), have been interpreted as secondary phenomena accompanying the passage of “internal solitons”. These are solitary wavelike distortions of the boundary layer between the warm upper layer of sea water and cold lower depths. These internal solitons are travelling ridges of warm water, extending hundreds of meters down below the thermal boundary, and carry enormous energy. Osborne and Burch [11]

investigated the underwater currents which were experienced by an oil rig in the Andaman sea, which was drilling at a depth of 3600 ft. One drilling rig was apparently spun through ninety degrees and moved one hundred feet by the passage of a soliton below.

### Rosby Waves and Solitons

Rosby waves [12] are long waves between layers of the atmosphere, created by the rotation of the planet. In particular, in the atmosphere of a rotating planet, a fluid particle is endowed with a certain rotation rate, determined by its latitude. Consequently its motion in the north-south direction is constrained by the conservation of angular momentum as in the case of internal waves where gravity inhibits the vertical motion of a density stratified fluid. There is an analogy between internal waves and Rossby waves under suitable conditions. The KdV equation has been proposed as a model for the evolution of Rossby solitons [13] and NLS equation for the evolution of Rossby wave packets. The Great Red Spot of the planet of Jupiter is often associated with a Rossby soliton.

### Bore Solitons

Rivers which flow to the open oceans are very often affected by tidal effects, tsunamis or storm surges. Tidal motions generate intensive water flows which can propagate upstream on tens of kilometers in the form of step-wise perturbation (hydraulic jumps) analogous to shock waves in acoustics. This phenomenon is known as a bore or mascaret (in French). Examples of such bores include the tidal bore of Seine river in France, Hooghly bore of the Ganges in India, the Amazon river bore in Brazil and Hangzhou bore in China. Bore disintegration into solitons is a possible phenomenon in such bores [14].

Besides these there are other possible oceanographic phenomena such as capillary wave solitons [15], resonant three-wave or four wave interaction solitons [16], etc., where also soliton picture is useful.

In this article we will first point out how in shallow channels the long wavelength wave propagation is described by KdV equation and its generalizations (Sect. “**Shallow Water Waves and KdV Type Equations**”). Then we will briefly point out how NLS family of equations arise naturally in the description of deep water waves (Sect. “**Deep Water Waves and NLS Type Equations**”). Based on these details, we will point how the soliton picture plays a very important role in the understanding of tsunami propagation (Sect. “**Tsunamis as Solitons**”), generation of internal solitons (Sect. “**Internal Solitons**”), formation of Rossby solitons (Sect. “**Rosby Solitons**”) and

disintegration of bores into solitons (Sect. “Bore Solitons”).

**Shallow Water Waves and KdV Type Equations**

Kortweg and de Vries [3] considered the wave phenomenon underlying the observations of Scott Russel from first principles of fluid dynamics and deduced the KdV equation to describe the unidirectional shallow water wave propagation in one dimension.

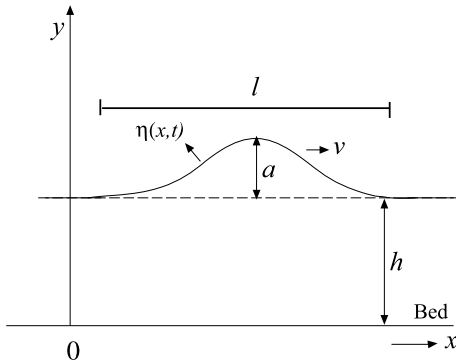
Consider the one-dimensional ( $x$ -direction) wave motion of an incompressible and inviscid fluid (water) in a shallow channel of height  $h$ , and of sufficient width with uniform cross-section leading to the formation of a solitary wave propagating under gravity. The effect of surface tension is assumed to be negligible. Let the length of the wave be  $l$  and the maximum value of its amplitude,  $\eta(x, t)$ , above the horizontal surface be  $a$  (see Fig. 2).

Then assuming  $a \ll h$  (shallow water) and  $h \ll l$  (long waves), one can introduce two natural small parameters into the problem  $\epsilon = a/h$  and  $\delta = h/l$ . Then the analysis proceeds as follows [3,6].

**Equation of Motion: KdV Equation**

The fluid motion can be described by the velocity vector  $V(x, y, t) = u(x, y, t)\mathbf{i} + v(x, y, t)\mathbf{j}$ , where  $\mathbf{i}$  and  $\mathbf{j}$  are the unit vectors along the horizontal and vertical directions, respectively. As the motion is irrotational, we have  $\nabla \times V = 0$ . Consequently, we can introduce the velocity potential  $\phi(x, y, t)$  by the relation  $V = \nabla\phi$ .

**Conservation of Density** The system obviously admits the conservation law for the mass density  $\rho(x, y, t)$  of the fluid,  $d\rho/dt = \rho_t + \nabla \cdot (\rho V) = 0$ . As  $\rho$  is a constant,



Solitons, Tsunamis and Oceanographical Applications of, Figure 2 One-dimensional wave motion in a shallow channel

we have  $\nabla \cdot V = 0$ . Consequently  $\phi$  obeys the Laplace equation

$$\nabla^2 \phi(x, y, t) = 0. \tag{1}$$

**Euler’s Equation** As the density of the fluid  $\rho = \rho_0 = \text{constant}$ , using Newton’s law for the rate of change of momentum, we can write  $dV/dt = \partial V/\partial t + (V \cdot \nabla)V = -\frac{1}{\rho_0} \nabla p - g\mathbf{j}$ , where  $p = p(x, y, t)$  is the pressure at the point  $(x, y)$  and  $g$  is the acceleration due to gravity, which is acting vertically downwards (here  $\mathbf{j}$  is the unit vector along the vertical direction). Since  $V = \nabla\phi$  we obtain (after one integration)

$$\phi_t + \frac{1}{2} (\nabla\phi)^2 + \frac{p}{\rho_0} + gy = 0. \tag{2}$$

**Boundary Conditions** The above two Eqs. (1) and (2) for the velocity potential  $\phi(x, y, t)$  of the fluid have to be supplemented by appropriate boundary conditions, by taking into account the fact (see Fig. 2) that (a) the horizontal bed at  $y = 0$  is hard and (b) the upper boundary  $y = y(x, t)$  is a free surface. As a result

(a) the vertical velocity at  $y = 0$  vanishes,  $v(x, 0, t) = 0$ , which implies

$$\phi_y(x, 0, t) = 0. \tag{3}$$

(b) As the upper boundary is free, let us specify it by  $y = h + \eta(x, t)$  (see Fig. 2). Then at the point  $x = x_1, y = y_1 \equiv y(x, t)$ , we can write  $\frac{dy_1}{dt} = \frac{\partial \eta}{\partial t} + \frac{\partial \eta}{\partial x} \frac{dx_1}{dt} = \eta_t + \eta_x u_1 = v_1$ . Since  $v_1 = \phi_{1y}, u_1 = \phi_{1x}$ , we obtain

$$\phi_{1y} = \eta_t + \eta_x \phi_{1x}. \tag{4}$$

(c) Similarly at  $y = y_1$ , the pressure  $p_1 = 0$ . Then from (2), it follows that

$$u_1 t + u_1 u_{1x} + v_1 v_{1x} + g\eta_x = 0. \tag{5}$$

Thus the motion of the surface of water wave is essentially specified by the Laplace Eq. (1) and Euler Eq. (2) along with one fixed boundary condition (3) and two variable nonlinear boundary conditions (4) and (5). One has to then solve the Laplace equation subject to these boundary conditions.

**Taylor Expansion of  $\phi(x, y, t)$  in  $y$**  Making use of the fact  $\delta = h/l \ll 1, h \ll l$ , we assume  $y (= h + \eta(x, t))$  to be small to introduce the Taylor expansion

$$\phi(x, y, t) = \sum_{n=0}^{\infty} y^n \phi_n(x, t). \tag{6}$$

Substituting the above series for  $\phi$  into the Laplace Eq. (1), solving recursively for  $\phi_n(x, t)$ 's and making use of the boundary condition (4),  $\phi_y(x, 0, t) = 0$ , one can show that

$$u_1 = \phi_{1x} = f - \frac{1}{2}y_1^2 f_{xx} + \text{higher order in } y_1, \quad (7)$$

$$v_1 = \phi_{1y} = -y_1 f_x + \frac{1}{6}y_1^3 f_{xxx} + \text{higher order in } y_1, \quad (8)$$

where  $f = \partial\phi_0/\partial x$ . We can then substitute these expressions into the nonlinear boundary conditions (4) and (5) to obtain equations for  $f$  and  $\eta$ .

**Introduction of Small Parameters  $\epsilon$  and  $\delta$**  So far the analysis has not taken into account fully the shallow nature of the channel ( $a/h = \epsilon \ll 1$ ) and the solitary nature of the wave ( $a/l = a/h \cdot h/l = \epsilon\delta \ll 1, \epsilon \ll 1, \delta \ll 1$ ), which are essential to realize the Scott Russel phenomenon. For this purpose one can stretch the independent and dependent variables in the above equations through appropriate scale changes, but retaining the overall form of the equations. To realize this one can introduce the natural scale changes

$$x = lx', \quad \eta = a\eta', \quad t = \frac{l}{c_0}t', \quad (9)$$

where  $c_0$  is a parameter to be determined. Then in order to retain the form of (7) and (8) we require

$$\begin{aligned} u_1 &= \epsilon c_0 u'_1, \quad v_1 = \epsilon \delta c_0 v'_1, \quad f = \epsilon c_0 f', \\ y_1 &= h + \eta(x, t) = h(1 + \epsilon \eta'(x', t')). \end{aligned} \quad (10)$$

Then

$$u'_1 = f' - \frac{1}{2}\delta^2(1 + \epsilon\eta')^2 f'_{x't'} = f' - \frac{1}{2}\delta^2 f'_{x't'}, \quad (11)$$

where we have omitted terms proportional to  $\delta^2\epsilon$  as small compared to terms of the order  $\delta^2$ . Similarly from (8), we obtain

$$v'_1 = -(1 + \epsilon\eta')f'_{x'} + \frac{1}{6}\delta^2 f'_{x't'}. \quad (12)$$

Now considering the nonlinear boundary condition (4) in the form  $v_1 = \eta_t + \eta_x u_1$ , it can be rewritten as

$$\eta'_{t'} + f'_{x'} + \epsilon\eta' f'_{x'} + \epsilon f' \eta'_{x'} - \frac{1}{6}\delta^2 f'_{x't'x'} = 0. \quad (13)$$

Similarly considering the other boundary condition (5) and making use of the above transformations, it can be rewritten, after neglecting terms of the order  $\epsilon^2\delta^2$ , as

$$f'_{t'} + \epsilon f' f'_{x'} + \frac{ga}{\epsilon c_0^2} \eta'_{x'} - \frac{1}{2}\delta^2 f'_{x't't'} = 0. \quad (14)$$

Now choosing the arbitrary parameter  $c_0$  as  $c_0^2 = gh$  so that  $\eta'_{x'}$  term is of order unity, (14) becomes

$$f'_{t'} + \eta'_{x'} + \epsilon f' f'_{x'} - \frac{1}{2}\delta^2 f'_{x't't'} = 0. \quad (15)$$

(Note that  $c_0 = \sqrt{gh}$  is nothing but the speed of the water wave in the linearized limit). Omitting the primes for convenience, the evolution equation for the amplitude of the wave and the function related to the velocity potential reads

$$\eta_t + f_x + \epsilon\eta f_x + \epsilon f \eta_x - \frac{1}{6}\delta^2 f_{xxx} = 0, \quad (16)$$

$$f_t + \eta_x + \epsilon f f_x - \frac{1}{2}\delta^2 f_{xxt} = 0. \quad (17)$$

Note that the small parameters  $\epsilon$  and  $\delta^2$  have occurred in a natural way in (16), (17).

**Perturbation Analysis** Since the parameters  $\epsilon$  and  $\delta^2$  are small in (16), (17), we can make a perturbation expansion of  $f$  in these parameters:

$$f = f^{(0)} + \epsilon f^{(1)} + \delta^2 f^{(2)} + \text{higher order terms}, \quad (18)$$

where  $f^{(i)}$ ,  $i = 0, 1, 2, \dots$  are functions of  $\eta$  and its spatial derivatives. Note that the above perturbation expansion is an asymptotic expansion. Substituting this into Eqs. (16) and (17) and regrouping and comparing different powers proportional to  $(\epsilon, \delta^2)$  and solving them successively one can obtain (see for example [6] for further details) in a self consistent way,

$$f^{(0)} = \eta, \quad f^{(1)} = -\frac{1}{4}\eta^2, \quad f^{(2)} = \frac{1}{3}\eta_{xx}. \quad (19)$$

Using these expressions into (18) and substituting it in (16) and (17), we ultimately obtain the KdV equation in the form

$$\eta_t + \eta_x + \frac{3}{2}\epsilon\eta\eta_x + \frac{\delta^2}{6}\eta_{xxx} = 0, \quad (20)$$

describing the unidirectional propagation of long wavelength shallow water waves.

### The Standard (Contemporary) Form of KdV Equation

Finally, changing to a moving frame of reference,  $\xi = x - t$ ,  $\tau = t$ , and introducing the new variables  $u = (3\epsilon/2\delta^2)\eta$ ,  $\tau' = (6/\delta^2)\tau$  and redefining the variables  $\tau'$  as  $t$  and  $\xi$  as  $x$ , we finally arrive at the ubiquitous form of the KdV equation as

$$u_t + 6uu_x + u_{xxx} = 0. \quad (21)$$

The Korteweg–de Vries Eq. (21) admits cnoidal wave solution and in the limiting case solitary wave solution as well. This form can be easily obtained [6,17] by looking for a wave solution of the form  $u = 2f(\xi)$ ,  $\xi = (x - ct)$ , and reducing the KdV equation into a third order nonlinear ordinary differential equation of the form

$$-c \frac{\partial f}{\partial \xi} + 12f \frac{\partial f}{\partial \xi} + \frac{\partial^3 f}{\partial \xi^3} = 0. \tag{22}$$

Integrating Eq. (22) twice and rearranging, we can obtain

$$\left(\frac{\partial f}{\partial \xi}\right)^2 = -4f^3 + cf^2 - 2df - 2b \equiv P(f), \tag{23}$$

where  $b$  and  $d$  are integration constants. Calling the three real roots of the cubic equation  $P(f) = 0$  as  $\alpha_1$ ,  $\alpha_2$  and  $\alpha_3$  such that

$$\left(\frac{\partial f}{\partial \xi}\right)^2 = -4(f - \alpha_1)(f - \alpha_2)(f - \alpha_3), \tag{24}$$

the solution can be expressed in terms of the Jacobian elliptic function as

$$f(\xi) = f(x - ct) = \alpha_3 - (\alpha_3 - \alpha_2) \operatorname{sn}^2 \left[ \sqrt{\alpha_3 - \alpha_1} (x - ct), m \right], \tag{25a}$$

where

$$(\alpha_1 + \alpha_2 + \alpha_3) = \frac{c}{4}, \quad m^2 = \frac{\alpha_3 - \alpha_2}{\alpha_3 - \alpha_1}, \quad \delta = \text{constant}. \tag{25b}$$

Here  $m$  is the modulus parameter of the Jacobian elliptic function. Eq. (25a) represents in fact the so called cnoidal wave (because of its elliptic function form). In the limiting case  $m = 1$ , the form (25a) reduces to

$$f = \alpha_2 + (\alpha_3 - \alpha_2) \operatorname{sech}^2 \left[ \sqrt{\alpha_3 - \alpha_1} (x - ct) \right]. \tag{26}$$

Choosing now  $\alpha_1 = 0$ ,  $\alpha_2 = 0$ , and using (25b) we have

$$f = \frac{c}{4} \operatorname{sech}^2 \left[ \frac{\sqrt{c}}{2} (x - ct) \right]. \tag{27}$$

Then the solitary wave solution to the KdV Eq. (21) can be written in the form

$$u(x, t) = \frac{c}{2} \operatorname{sech}^2 \frac{\sqrt{c}}{2} (x - ct + \delta), \quad \delta : \text{constant} \tag{28}$$

Note that the velocity of the solitary wave is directly proportional to the amplitude: larger the wave the higher is the speed. More importantly, the KdV solitary wave is

a soliton: it retains its shape and speed upon collision with another solitary wave of different amplitude, except for a phase shift, see Fig. 1 [6,7,8]. In fact for an arbitrary initial condition, the solution of the Cauchy initial value problem consists of  $N$ -number of solitons of different amplitudes in the background of small amplitude dispersive waves. All these results ultimately lead to the result that the KdV equation is a completely integrable, infinite dimensional, nonlinear Hamiltonian system. It possesses [6,7,8]

- (i) a Lax pair of linear differential operators and is solvable through the so called inverse scattering transform (IST) method,
- (ii) infinite number of conservation laws and associated infinite number of involutive integrals of motion,
- (iii)  $N$ -soliton solution,
- (iv) Hirota bilinear form,
- (v) Hamiltonian structure

and a host of other interesting properties (see for example [6,7,8,9]).

### KdV Related Integrable and Nonintegrable NLEEs

Depending on the actual physical situation, the derivation of the shallow water wave equation can be suitably modified to obtain other forms of nonlinear dispersive wave equations in (1+1) dimensions as well as in (2+1) dimensions relevant for the present context. Without going into the actual derivations, some of the important equations possessing solitary waves are listed below [6,7,8,9].

1. Boussinesq equation [7]

$$u_t + uu_x + g\eta_x - \frac{1}{3}h^2u_{txx} = 0, \tag{29a}$$

$$\eta_t + [u(h + \eta)]_x = 0 \tag{29b}$$

2. Benjamin–Bona–Mahoney (BBM) equation [18]

$$u_t + u_x + uu_x - u_{xxt} = 0 \tag{30}$$

3. Camassa–Holm equation [19]

$$u_t + 2\kappa u_x + 3uu_x - u_{xxt} = 3u_x u_{xx} + uu_{xxx} \tag{31}$$

4. Kadomtsev–Petviashvili (KP) equation [7]

$$(u_t + 6uu_x + u_{xxx})_x + 3\sigma^2 u_{yy} = 0 \tag{32}$$

( $\sigma^2 = -1$ : KP-I,  $\sigma^2 = +1$ : KP-II).

In the derivation of the above equations, generally the bottom of water column or fluid bed is assumed to be flat. However in realistic situations the water depth varies as a function of the horizontal coordinates. In this situation, one often encounters inhomogeneous forms of the above wave equations. Typical example is the variable coefficient KdV equation [14]:

$$u_t + f(x, t)uu_x + g(x, t)u_{xxx} = 0, \quad (33)$$

where  $f$  and  $g$  are functions of  $x, t$ . More general forms can also be deduced depending upon the actual situations, see for example [14].

### Deep Water Waves and NLS Type Equations

Deep water waves are strongly dispersive in contrast to the weakly dispersive nature of the shallow water waves (in the linear limit). Various authors (see for details [8]) have shown that nonlinear Schrödinger family of equations models the evolution of a packet of surface waves in this case. There are several oceanographic situations where such waves can arise [8]:

(i) A localized storm at sea can generate a wide spectrum of waves, which then propagates away from the source region in all horizontal directions. If the propagating waves have small amplitudes and encounter no wind away from the source region, these waves can eventually sort themselves into nearly one-dimensional packets of nearly monochromatic waves. For appropriately chosen scales, the underlying evolution of each of these packets can be shown to satisfy the nonlinear Schrödinger equation and its generalizations in (2+1) dimensions.

(ii) Nearly monochromatic, nearly one-dimensional waves can cover a broad range of surface waves in the sea that results due to a steady wind of long duration and fetch. Then the generalization of the NLS equation in (2+1) dimensions can describe the waves that result in when the wind stops.

In all the above situations one looks for the solution of the equations of motion (1)-(5) but generalized in three dimensions which consists mainly in the form of a small amplitude, nearly monochromatic, nearly one-dimensional wave train. Assuming that this wave train travels in the  $x$ -direction with a mean wave number  $\kappa = (k, l)$  with a characteristic amplitude ' $a$ ' of the disturbance and a characteristic variation  $\delta k$  in  $k$ , one can deduce the NLS equation in (2+1) dimensions under the following conditions:

- (i) small amplitudes such that  $\hat{\epsilon} = \epsilon\delta \equiv \kappa a \ll 1$
- (ii) slowly varying modulations,  $\frac{\delta k}{\kappa} \ll 1$

- (iii) nearly one dimensional waves,  $\frac{|l|}{k} \ll 1$
- (iv) balance of all three effects,  $\frac{\delta k}{\kappa} = \frac{|l|}{k} \approx O(\hat{\epsilon})$
- (v) for finite and deep water waves,  $(kh)^2 \gg \hat{\epsilon}$

In the lowest order approximation (linear approximation) of water waves, the prediction is that a localized initial state will generally evolve into wave packets with a dominant wave number  $\kappa$  and corresponding frequency  $\omega$ , given by the dispersion relation  $\omega = (g\kappa + \sigma\kappa^3)\tanh\kappa h$ ,  $\kappa = |\kappa| = \sqrt{k^2 + l^2}$  within which each wave propagates with the phase speed  $c = \omega/k$ , while the envelope propagates with the group velocity  $c_g = d\omega/d\kappa$ . After a sufficiently long time the wave packet tends to disperse around the dominant wave number.

This tendency for dispersion can be offset by cumulative nonlinear effects. In the absence of surface tension, the outcome for unidirectional waves can be shown to be describable by the NLS equation. If the surface wave in the lowest order is  $\phi \approx A \exp i(kx - \omega t) + c.c.$ , where  $\phi$  is the velocity potential, then to leading order the wave amplitude evolves as

$$i(A_t + c_g A_x) + \frac{1}{2}\lambda A_{xx} + \mu |A|^2 A = 0. \quad (34)$$

The coefficients here are given by

$$\begin{aligned} \lambda &= \frac{\partial^2 \omega}{\partial k^2}, \\ \mu &= -\frac{\omega k^2}{16S^4}(8C^2S^2 + g - 2T^2) \\ &\quad + \frac{\omega}{8C^2S^2} \frac{(2\omega C^2 + kc_g)^2}{(gh - c_g^2)}, \end{aligned} \quad (35)$$

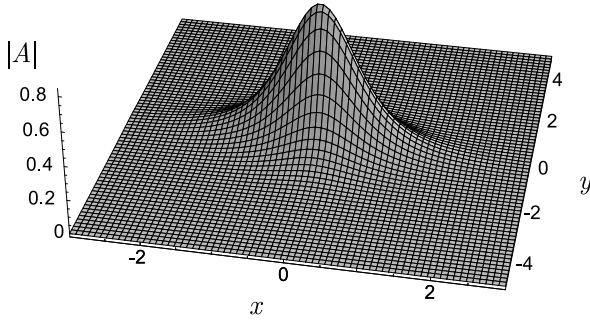
where  $C = \cosh(kh)$ ,  $S = \sinh(kh)$ ,  $T = \frac{S}{C}$ . Equation (34) has been obtained originally by Zakharov in 1968 for deep water waves [20] and by Hasimoto and Ono for waves of finite depth in 1972 [21].

The NLS equation is also a soliton possessing integrable system and is solvable by the IST method [6,7,8,9]. For  $\lambda > 0$  (focusing case), the envelope solitary (soliton) wave solution (also called bright solitons in the optical physics context) is given by

$$A(x, t) = a \operatorname{sech} \gamma(x - c_g t) \exp(-i\Omega t), \quad (36)$$

where  $\mu a^2 = \lambda \gamma^2$ ,  $\Omega = +\frac{1}{2}\mu a^2$ .

When the effects of modulation in the transverse  $y$ -direction are taken into account, so that the wave amplitude is now given by  $A(x, y, t)$ , the NLS equation is replaced by the Benney-Roskes system [22] also popularly known as the Davey-Stewartson equations [23],



Solitons, Tsunamis and Oceanographical Applications of, Figure 3

Exponentially localized dromion solution of the Davey–Stewartson equation at a fixed time ( $t = 0$ ) for suitable choice of parameters

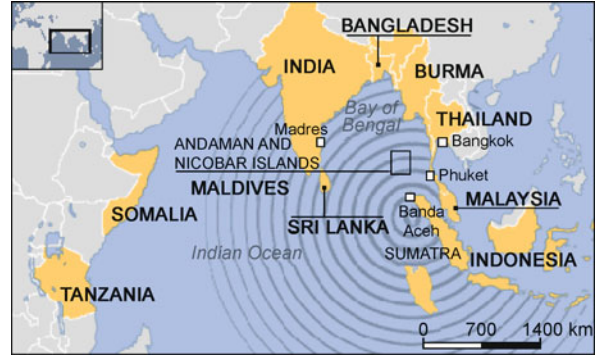
$$i(A_t + c_g A_x) + \frac{1}{2} \lambda A_{xx} + \frac{1}{2} \delta A_{yy} + \mu |A|^2 A + UA = 0, \tag{37a}$$

$$\alpha U_{xx} + U_{yy} + \beta (|A|^2)_{yy} = 0, \tag{37b}$$

where  $\delta = \frac{c_g}{k}$ ,  $\alpha = 1 - \left(\frac{c_g^2}{gh}\right)$ ,  $gh\beta = \frac{\omega}{8C^2g^2}(2\omega C^2 + kc_g)^2$ . Here  $U(x, y, t)$  is the wave induced mean flow. In the deep water wave limit,  $kh \rightarrow \infty$ , and  $U \rightarrow 0$ ,  $\beta \rightarrow 0$  and one has the nonintegrable (2+1) dimensional NLS equation. On the other hand in the shallow water limit, one has the integrable Davey–Stewartson (DS) equations. For details see [7,8]. The DS-I equation admits algebraically decaying lump solitons and exponentially decaying dromions [24] besides the standard line solitons for appropriate choices of parameters. A typical dromion solution of DS-I equation is shown in Fig. 3.

### Tsunamis as Solitons

The term ‘tsunami’ (tsu:harbour, nami:wave in Japanese) which was perhaps an unknown word even for scientists in countries such as India, Srilanka, Thailand, etc. till recently has become a house-hold word since that fateful morning of December 26, 2004. When a powerful earthquake of magnitude 9.1–9.3 on the Richter scale, epicentered off the coast of Sumatra, Indonesia, struck at 07:58:53, local time, described as the 2004 Indian Ocean earthquake or Sumatra–Andaman earthquake (Fig. 4), it triggered a series of devastating tsunamis as high as 30 meters that spread throughout the Indian Ocean killing about 275,000 people and inundating coastal communities across South and Southeast Asia, including parts of Indonesia, Srilanka, India and Thailand and even reaching as far as the east coast



Solitons, Tsunamis and Oceanographical Applications of, Figure 4

26 December 2004 Indian Ocean tsunami (adapted from the website [www.blogaid.org.uk](http://www.blogaid.org.uk) with the courtesy of Andy Budd)

of Africa [25]. The catastrophe is considered to be one of the deadliest disasters in modern history.

Since this earthquake and consequent tsunamis, several other earthquakes of smaller and larger magnitudes keep occurring off the coast of Indonesia. Even as late as July 17, 2006 an earthquake of magnitude 7.7 on the Richter scale struck off the town of Pandering at 15.19 local time and set off a tsunami of 2m high which had killed more than 300 people.

These tsunamis, which can become monstrous tidal waves when they approach coastline, are essentially triggered due to the sudden vertical rise of the seabed by several meters (when earthquake occurs) which displaces massive volume of water. The tsunamis behave very differently in deep water than in shallow water as pointed out below. By no means the tsunami of 2004 and later ones are exceptional; More than two hundred tsunamis have been recorded in scientific literature since ancient times. The most notable earlier one is the tsunami triggered by the powerful earthquake (9.6 magnitude) off southern Chile on May 22, 1960 [10] which traveled almost 22 hours before striking Japanese islands.

It is clear from the above events that the tsunami waves are fairly permanent and powerful ones, having the capacity to travel extraordinary distances without practically diminishing in size or speed. In this sense they seem to have considerable resemblance to shallow water nonlinear dispersive waves of KdV type, particularly solitary waves and solitons. It is then conceivable that tsunami dynamics has close connection with soliton dynamics.

### Basics of Tsunami Waves

As noted above tsunami waves of the type described above are essentially triggered by massive earthquakes which

lead to vertical displacement of a large volume of water. Other possible reasons also exist for the formation and propagation of tsunami waves: underwater nuclear explosion, larger meteorites falling into the sea, volcano explosions, rockslides, etc. But the most predominant cause of tsunamis appear to be large earthquakes as in the case of the Sumatra–Andaman earthquake of 2004. Then there are three major aspects associated with the tsunami dynamics [26]:

1. Generation of tsunamis
2. Propagation of tsunamis
3. Tsunami run up and inundation

There exist rather successful models to approach the generation aspects of tsunamis when they occur due to the earthquakes [27]. Using the available seismic data it is possible to reconstruct the permanent deformation of the sea bottom due to earthquakes and simple models have been developed (see for example, [28]). Similarly the tsunami run up and inundation problems [29] are extremely complex and they require detailed critical study from a practical point of view in order to save structures and lives when a tsunami strikes.

However, here we will be more concerned with the propagation of tsunami waves and their possible relation to wave propagation associated with nonlinear dispersive waves in shallow waters. In order to appreciate such a possible connection, we first look at the typical characteristic properties of tsunami waves as in the case of 2004 Indian Ocean tsunami waves or 1960 Chilean tsunamis.

### The Indian Ocean Tsunami of 2004

Considering the Indian Ocean 2004 tsunami, satellite observations after a couple of hours after the earthquake establish an amplitude of approximately 60 cms in the open ocean for the waves. The estimated typical wavelength is about 200 kms [30]. The maximum water depth  $h$  is between 1 and 4 kms. Consequently, one can identify in an average sense the following small parameters ( $\epsilon$  and  $\delta^2$ ) of roughly equal magnitude:

$$\epsilon = \frac{a}{h} \approx 10^{-4} \ll 1, \quad \delta^2 = \frac{h^2}{l^2} \approx 10^{-4} \ll 1 \quad (38)$$

As a consequence, it is possible that a nonlinear shallow water wave theory where dispersion (KdV equation) also plays an important role (as discussed in Sect. “**Shallow Water Waves and KdV Type Equations**”) has considerable relevance [26]. However, we also wish to point out here that there are other points of view: Constantin and Johnson [31] estimate  $\epsilon \approx 0.002$  and  $\delta \approx 0.04$

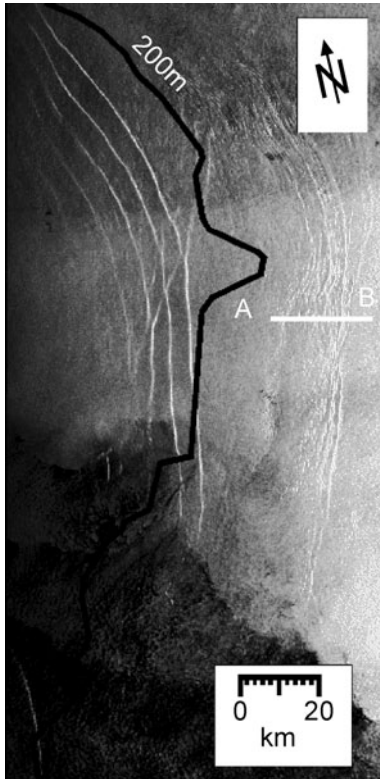
and conclude that for both nonlinearity and dispersion to become significant the quantity  $\delta\epsilon^{-3/2} \times \text{wavelength}$  estimated as 90,000 kms is too large and shallow water equations with variable depth (without dispersion) should be used. However, it appears that these estimates can vary over a rather wide range and with suitable estimates it is possible that the range of 10,000–20,000 kms could be also possible ranges and hence taking into account the fact that both the Indian Ocean 2004 and Chilean 1960 tsunamis have traveled over 10 hours or more (in certain directions) before encountering land mass appears to allow for the possibility of nonlinear dispersive waves as relevant features for the phenomena. Segur [32] has argued that in the 2004 tsunamis, the propagation distances from the epicenter of the earthquake to India, Srilanka, or Thailand were too short for KdV dynamics to develop. In the same way one can argue that the waves that hit Somalia and Kenya in the east coast of Africa (or as in the case of Chilean earthquake see also [32]) have traveled sufficiently long distance for KdV dynamics to become important [33]. In any case one can conclude that at least for long distance tsunami propagation solitary wave and soliton picture of KdV like equations become very relevant.

### Internal Solitons

For a long time seafarers passing through the Strait of Malacca on their journeys between India and the Far East have noticed that in the Andaman sea, between Nicobar islands and the north east coast of Sumatra, often bands of strongly increased surface roughness (rippings or bands of choppy water) occur [11,34]. Similar observations have been reported in other seas around the globe from time to time. In recent times there has been considerable progress in understanding these kind of observations in terms of internal solitons in the oceans [7]. These studies have been greatly facilitated by photographs taken from satellites and space-crafts orbiting the earth, for example by synthetic aperture radar (SAR) images of ERS-1/2 satellites [34,35].

Peculiar striations of 100 km long, separated by 6 to 15 km and grouped in packets of 4 to 8, visible on satellite photographs (see Fig. 5) of the surface of the Andaman and Sulu seas in the Far East, have been interpreted as secondary phenomena accompanying the passage of ‘internal solitons’, which are solitary wavelike distortions of the boundary layer between warm upper layer of sea water and cold lower depths. These internal solitons are traveling edges of warm water, extending hundreds of meters down below the thermal boundary [7]. They carry enormous energy with them which is perhaps the reason for unusually strong underwater currents experienced by deep-sea





Solitons, Tsunamis and Oceanographical Applications of, Figure 5

SAR image of a 200 km  $\times$  200 km large section of the Andaman Sea acquired by the ERS-2 satellite on April 15, 1996 showing sea surface manifestations of two internal solitary wave packets, see [34]. Figure reproduced from ESA website [www.earth.esa.int/workshops/ers97/papers/alpers3](http://www.earth.esa.int/workshops/ers97/papers/alpers3) with the courtesy of European Space Agency and W. Alpers

drilling rigs. Thus these internal solitons are potentially hazardous to sub-sea oil and gas explorations. The ability to predict them can improve substantially the cost effectiveness and safety of offshore drilling.

A systematic study of the underwater currents experienced by an oil rig in the Andaman sea which was drilling at a depth of 3600 ft was carried out by Osborne and Burch in 1980 [11]. They spent four days measuring underwater currents and temperatures. The striations seen on satellite photographs turned out to be kilometer-wide bands of extremely choppy water, stretching from horizon to horizon, followed by about two kilometers of water “as smooth as a millpond”. These bands of agitated water are called “tide rips”, they arose in packets of 4 to 8, spaced about 5 to 10 km apart (they reached the research vessel at approximately hourly intervals) and this pattern was repeated with the regularity of tidal phenomenon.

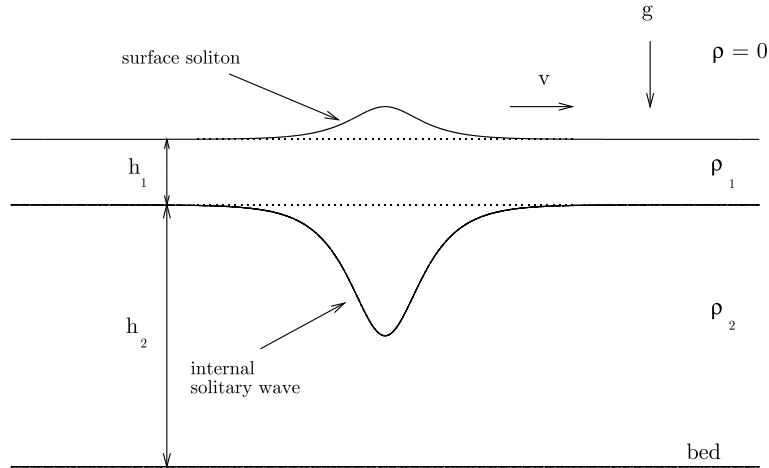
As described in [7], Osborne and Burch found that the amplitude of each succeeding soliton was less than the previous one, is precisely what is expected for solitons (note that the velocity of a solitary wave solution of KdV equation increases with amplitude, vide Eq. (22)). Thus if a number of solitons are generated together, then we expect them eventually to be arranged in an ordered sequence of decreasing amplitude. From the spacing between successive waves in a packet and the rate of separation calculated from the KdV equation, Osborne and Burch were able to estimate the distance the packet had traveled from its source and thus identify possible source regions [7]. They concluded that the solitons are generated by tidal currents off northern Sumatra or between the islands of the Nicobar chain that extends beyond it and that their observations have good general agreement with the predictions for internal solitons as given by the KdV equation. Numerous recent observations and predictions of solitons in the Andaman sea have clearly established that it is a site where extraordinarily large internal solitons are encountered [34,35].

Further, Apel and Holbrook [36] undertook a detailed study of internal waves in the Sulu sea. Satellite photographs had suggested that the source of these waves was near the southern end of the Sulu sea and their research ship followed one wave packet for more than 250 miles over a period of two days – an extraordinary coherent phenomenon [7]. These internal solitons travel at speeds of about 8 kilometers per hour (5 miles per hour), with amplitude of about 100 meters and wavelength of about 1700 meters.

Similar observations elsewhere have confirmed the presence of internal solitons in oceans including the strait of Messina, the strait of Gibraltar, off the western side of Baja California, the Gulf of California, the Archipelago of La Maddalena and the Georgia strait [7]. There has also been a number of experimental studies of internal solitons in laboratory tanks in the last few decades [37]. These experiments provide detailed quantitative information usually unavailable in the field conditions, and are also an efficient tool for verifying various theoretical models.

As a theoretical formulation of internal solitons [7], consider two incompressible, immiscible fluids, with densities  $\rho_1$  and  $\rho_2$  and depths  $h_1$  and  $h_2$  respectively such that the total depth  $h = h_1 + h_2$ . Let the lighter fluid of height  $h_1$  be lying over a heavier fluid of height  $h_2$ , in a constant gravitational field (Fig. 6). The lower fluid is assumed to rest on a horizontal impermeable bed, and the upper fluid is bounded by a free surface.

Then as in Sect. “Shallow Water Waves and KdV Type Equations”, we denote the characteristic amplitude



Solitons, Tsunamis and Oceanographical Applications of, Figure 6  
 Formation of internal soliton (note that under suitable conditions small amplitude surface soliton can also be formed)

of wave by ‘ $a$ ’ and the characteristic wavelength  $l = k^{-1}$ . Then the various nonlinear wave equations to describe the formation of internal solitons follow by suitable modification of the formulation in Sect. “Shallow Water Waves and KdV Type Equations”, assuming viscous effects to be negligible. Each of these equations is completely integrable and admits soliton solutions [7].

(a) **KdV equation (Eq. (21))** follows when

- (i) the waves are of long wavelength  $\delta = \frac{h}{l} \ll 1$ ,
- (ii) the amplitude of the waves are small,  $\epsilon = \frac{a}{h} \ll 1$ , and
- (iii) the two effects are comparable  $\delta^2 = O(\epsilon)$

(b) **Intermediate-Long-Wave (ILW) equation [38]**

$$u_t + u_x + 2uu_x + Tu_{xx} = 0, \tag{39}$$

where  $Tu$  is the singular integral operator

$$(Tf)(x) = \frac{1}{2l} \int_{-\infty}^{\infty} \coth \left\{ \frac{\pi}{2l}(y-x) \right\} f(y) dy \tag{40}$$

with  $\int_{-\infty}^{\infty}$  the Cauchy principal value integral is obtained under the assumption that

- (a) there is a thin (upper) layer,  $\epsilon = \frac{h_1}{h_2} \ll 1$ ,
- (b) the amplitude of the waves is small,  $a \ll h_1$ ,
- (c) the above two effects balance,  $\frac{a}{h_1} = O(\epsilon)$ ,
- (d) the characteristic wavelength is comparable to the total depth of the fluid,  $l = kh = O(1)$  and
- (e) the waves are long waves in comparison with the thin layer,  $kh_1 \ll 1$ .

(c) **Benjamin-Ono equation [39]**

$$u_t + 2uu_x + Hu_{xx} = 0, \tag{41}$$

where  $Hu$  is the Hilbert transform

$$(Hf)(x) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{f(y)}{y-x} dy, \tag{42}$$

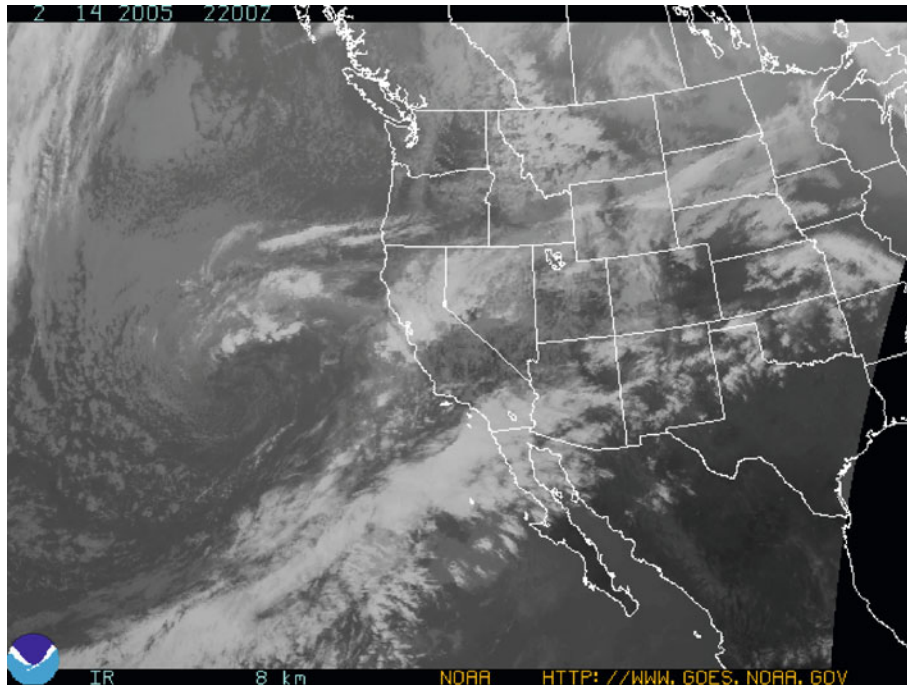
is obtained under the assumption

- (a) there is a thin (upper) layer  $h_1 \ll h_2$ ,
- (b) the waves are long waves in comparison with the thin layer,  $kh_1 \ll 1$ ,
- (c) the waves are short in comparison with the total depth of the fluid,  $kh \gg 1$  and
- (d) the amplitude of the waves is small,  $a \ll h_1$ .

It may be noted that in the shallow water limit, as  $\delta \rightarrow 0$ , the ILW equation reduces to the KdV equation, while the Benjamin-Ono equation reduces to it in the deep water wave limit as  $\delta \rightarrow \infty$ . Each of these equations have their own ranges of validity and admit solitary wave and soliton solutions to represent internal solitons of the oceans.

**Rossby Solitons**

The atmospheric dynamics is an important subject of human concern as we live within the atmosphere and are continuously affected by the weather and its rather complex behavior. The motion of the atmosphere is intimately connected with that of the ocean with which it exchanges



Solitons, Tsunamis and Oceanographical Applications of, Figure 7

A coupled high/low pressure systems in the form of a Rossby soliton formed off the coast of California/Washington on Valentine's Day 2005. The low pressure front hovers over Los Angeles dumping 30 inches of rain on the city in two weeks. The high pressure system lingers off the coast of Washington state providing unseasonably warm and sunny weather. This particular Rossby soliton proved exceptionally stable because of its remarkable symmetry. In an accompanying animated gif in this website one can watch the very interesting phenomenon as the jet stream splits around the soliton suctioning warm wet air from the off the coast of Mexico to Arizona, leaving behind a welcome drenching of rain. The figure and caption have been adapted from the website [http://mathpost.la.asu.edu/~rubio/rossby\\_soliton/rs.html](http://mathpost.la.asu.edu/~rubio/rossby_soliton/rs.html) with the courtesy of the National Oceanic and Atmospheric Administration (NOAA) and Antonio Rubio

fluxes of momentum, heat and moisture [40]. Then the atmospheric dynamics is dominated by the rotation of the earth and the vertical density stratification of the surrounding medium, leading to newer effects. Similar effects are also present in other planetary dynamics as well.

In the atmosphere of a rotating planet, a fluid particle is endowed with a certain rotation rate (Coriolis frequency), determined by its latitude. Consequently its motion in the north-south direction is inhibited by conservation of angular momentum. The large scale atmospheric waves caused by the variation of the Coriolis frequency with latitude are called *Rossby waves*. In Sect. “**Internal Solitons**” we saw that KdV equation and its modifications model internal waves. Since there is a resemblance between internal waves and Rossby waves, it is expected that KdV like equations can model Rossby waves as well [8]. Under the simplest assumptions like long waves, incompressible fluid,  $\beta$ -plane approximations, etc. Benney [41] had derived the KdV equation as a model for Rossby

waves in the presence of east-west zonal flow. Boyd [42] had shown that long, weakly nonlinear, equatorial Rossby waves are governed either by KdV or MKdV equation. Recently it has been shown by using the eight years of Topex/Poseidon altimeter observations [43] that a detailed characterization of major components of Pacific dynamics confirms the presence of equatorial Rossby solitons.

Also observational and numerical studies of propagation of nonlinear Rossby wave packets in the barotropic atmosphere by Lee and Held [44,45] have established their presence, notably in the Northern Hemisphere as storm tracks, but more clearly in the Southern Hemisphere (see Fig. 7). They also found that the wavepackets both in the real atmosphere and in the numerical models behave like the envelope solitons of the nonlinear Schrödinger equation (Fig. 7).

An interesting application of KdV equation to describe Rossby waves is the conjecture of Maxworthy and Redekopp [46] that the planet Jupiter's Great Red Spot

might be a solitary Rossby wave. Photographs taken by the spacecraft Voyager of the cloud pattern show that the atmospheric motion on Jupiter is dominated by a number of east-west zonal currents, corresponding to the jet streams of the earth's atmosphere [8]. Several oval-shaped spots are also seen, including the prominent Great Red Spot in the southern hemisphere. The latter one has been seen approximately this latitude for hundreds of years and maintains its form despite interactions with other atmospheric objects. One possible explanation is that the Red Spot is a solitary wave/soliton of the KdV equation, deduced from the quasigeostrophic form of potential vorticity equation for an incompressible fluid. A second test of the model is the combined effect of interaction of the Red Spot and Hollow of the Jupiter atmosphere on the South Tropical Disturbance which occurred early in the 20th century and lasted several decades. Maxworthy and Redekopp [46] have interpreted this interaction as that of two soliton collision of KdV equation, with the required phase shift [6,7,8].

The above connection between the KdV solitary wave and the rotating planet may be seen more concretely by the following argument as detailed in [8]. One can start with the quasigeostrophic form of the potential vorticity equation for an incompressible fluid in the form

$$\left\{ \left( \frac{\partial}{\partial t} + v \frac{\partial}{\partial x} \right) + \epsilon \left( \frac{\partial \psi}{\partial y} \right) \left( \frac{\partial}{\partial x} - \frac{\partial \psi}{\partial x} \frac{\partial}{\partial y} \right) \right\} \left\{ \mu^2 \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial}{\partial z} \left( K^2 \frac{\partial}{\partial z} \right) \right\} \psi + (\beta - U'') \frac{\partial \psi}{\partial x} = 0, \quad (43)$$

where  $x$ ,  $y$  and  $z$  represent the east, north and vertical directions, while the function  $U$  is related to the horizontal stream function  $\Phi$  as  $\Phi(x, y, z, t) = \int_y^y U(\eta) d\eta + \epsilon \psi(x, y, z, t)$  in terms of a zonal shear flow and a perturbation. In (43), in the  $\beta$ -plane approximation, the Coriolis parameter is given as  $f = 2\Omega \sin \theta_0 + \beta y$  and the function  $K(z)$  is given by  $K(z) = 2\Omega \sin \theta_0 l_2 / N(z) d$ , where  $N(z)$  is the Brunt-Väisälä frequency and  $l_2$  is the characteristic length scales in the north-south direction while  $d$  is the length scale in the vertical direction. Note that  $K$  compares the effects of rotation and density of variation. Then  $\mu = l_2 / l_1$  represents the ratio of the length scales in the north-south and east-west directions.

In the linear ( $\epsilon \rightarrow 0$ ), long wave ( $\mu \rightarrow 0$ ) limit, the potential function  $\psi$  has the form  $\psi = \sum_n A_n(x - c_n t) \phi_n(y) p_n(z)$ , where  $c_n$  is deduced by solving two related eigenvalue problems [8],  $(K^2 p_n')' + k_n^2 p_n = 0$ ,  $p_n(0) = p_n(1) = 0$  and  $\phi_n'' - k_n^2 \phi_n + [(\beta - U'') / (U - c_n)] \phi_n = 0$ ,

$\phi_n(y_s) = \phi_n(y_N) = 0$ . If the various modes are separable on a short time scale, one can then deduce an evolution equation for the individual modes, by eliminating secular terms at higher order in the expansion. Depending on the nature and existence of a stable density of stratification, which is characterized by the function  $N(z)$  mentioned above, either KdV or mKdV equation can be deduced for a given mode and thereby establishing the soliton connection in the present problem.

In spite of the complexity of the phenomenon underlying Rossby solitons there is clear evidence of the significance of solitonic picture.

### Bore Solitons

Rivers which flow into the open oceans are usually affected by tidal flows, tsunami or storm surge. For a typical estuary as one moves towards the mouth of the river, the depth increases and width decreases. When a tidal wave, tsunami or storm surge hits such an estuary, it can be seen as a hydraulic jump (step-wise perturbations like a shock-wave) in the water height and speed which will propagate upstream [47]. Far less dangerous but very similar is the bore (mascaret in French), a tidal wave which can propagate in a river for considerable distances.

Typical examples of bores occur in Seine river in France and the Hooghli river in West Bengal in India. A bore existed in the Seine river upto 1960 and disappeared when the river was dredged. The tide amplitude here is one of the largest in the world. The Hughli river is a branch of the Ganges that flows through Kolkata where a bore of 1m is present, essentially due to the shallowness of the river. Another interesting example is that at the time of the 1983 Japan sea tsunami, waves in the form of a bore ascended many rivers [48]. In some cases, bores had the form of one initial wave with a train of smaller waves and in other cases only a step with flat water surface behind was observed (Fig. 8). The Hangzhou bore in China is a tourist attraction. Other well known bores occur in the Amazon in Brazil and in Australia.

Another interesting situation where bore solitons were observed was in the International H<sub>2</sub>O Project (IHOP), as a density current (such as cold air from thunderstorm) intrudes into a fluid of lesser density that occurs beneath a low level inversion in the atmosphere. A spectacular bore and its evolution into a beautiful amplitude-ordered train of solitary waves were observed and sampled during the early morning of 20 June 2002 by the Leandre-II abroad the P-3 aircraft. The origin of this bore was traceable to a propagating cold outflow boundary from a mesoscale convective system in extreme western Kansas [49].



Solitons, Tsunamis and Oceanographical Applications of, Figure 8

Tidal bore at the mouth of the Araguari River in Brazil. The bore is undular with over 20 waves visible. (Adapted from the book "Gravity Currents in the Environment and the Laboratory" by John E. Simpson, Cambridge University Press with the courtesy of D.K. Lynch, John E. Simpson and Cambridge University Press)

In the process of propagation the bore undergoes dissipation, dispersive disintegration, enhancement due to decrease of the river width and depth, influence of nonlinear effects and so on. The profile depends on the Froude number which is a dimensionless ratio of inertial and gravitational effects. Theoretical models have been developed to study these effects based on KdV and its generalizations [14]. For example, bore disintegration into solitons in channel of constant parameters can be studied in signal coordinates in terms of the KdV like equation for the perturbation of water surface,

$$\eta_x + \frac{1}{c_0}(1 - \alpha\eta)\eta_t - \beta\eta_{ttt} = 0, \quad (44)$$

where  $c_0 = \sqrt{gh}$ ,  $\alpha = 3/2h$ ,  $\beta = h^2/6c_0^3$ ,  $h$  being the depth of the river, with the bore represented by a Heaviside step function as the boundary condition at  $x = 0$ . Other effects then can be incorporated into a variable KdV equation of the form (26).

### Future Directions

We have indicated a few of the most important oceanographical applications of solitons including tsunamis, internal solitons, Rossby solitons and bore solitons. There are other interesting phenomena like capillary wave solitons [15], resonant three and four wave interaction solitons [16], etc. which are also of considerable interest depending on whether the wave propagation corresponds to shallow, intermediate or deep waters. Whatever be the situation, it is clear that experimental observations as

well as their theoretical formulation and understandings are highly challenging complex nonlinear evolutionary problems. The main reason is that the phenomena are essentially large scale events and detailed experimental observations require considerable planning, funding, technology and manpower. Often satellite remote sensing measurements need to be carried out as in the case of internal solitons and Rossby solitons. Events like tsunami propagation are rare and time available for making careful measurements are limited and heavily dependent on satellite imaging, warning systems and after event measurements. Consequently developing and testing theoretical models are extremely hazardous and difficult. Yet the basic modeling in terms of solitary wave/soliton possessing nonlinear dispersive wave equations such as the KdV and NLS family of equations and their generalizations present fascinating possibilities to understand these large scale complex phenomena and opens up possibilities of prediction. Further understanding of such nonlinear evolution equations, both integrable and nonintegrable systems particularly in higher dimensions, can help to understand the various phenomena clearly and provide means of predicting events like tsunamis and damages which occur due to internal solitons and bores. Detailed experimental observations can also help in this regard. It is clear that what has been understood so far is only the qualitative aspects of these phenomena and much more intensive work is needed to understand the quantitative aspects to predict them.

### Bibliography

#### Primary Literature

1. Russel JS (1844) Reports on Waves. 14th meeting of the British Association for Advancement of Science. John Murray, London, pp 311–390
2. Bullough RK (1988) The Wave Par Excellence. The solitary progressive great wave of equilibrium of fluid: An early history of the solitary wave. In: Lakshmanan M (ed) Solitons: Introduction and Applications. Springer, Berlin
3. Korteweg DJ, de Vries G (1895) On the change of form of long waves advancing in a rectangular canal, and on a new type of long stationary waves. Philos Mag 39:422–443
4. Zabusky NJ, Kruskal MD (1965) Interactions of solitons in a collisionless plasma and recurrence of initial states. Phys Rev Lett 15:240–243
5. Gardner CS, Greene JM, Kruskal MD, Miura RM (1967) Method for solving the Korteweg–de Vries equation. Phys Rev Lett 19:1095–97
6. Lakshmanan M, Rajasekar S (2003) Nonlinear Dynamics: Integrability and Chaos. Springer, Berlin
7. Ablowitz MJ, Clarkson PA (1991) Solitons, Nonlinear Evolution Equations and Inverse Scattering. Cambridge University Press, Cambridge

8. Ablowitz MJ, Segur H (1981) *Solitons and Inverse Scattering Transform*. Society for Industrial and Applied Mathematics, Philadelphia
9. Scott AC (1999) *Nonlinear Science: Emergence and Dynamics of Coherent Structures*. Oxford University Press, New York
10. Dudley WC, Miu L (1988) *Tsunami!* University of Hawaii Press, Honolulu
11. Osborne AR, Burch TL (1980) Internal solitons in the Andaman Sea. *Science* 258:451–460
12. Rossby GG (1939) Relation between variations in the intensity of the zonal circulation of the atmosphere. *J Mar Res* 2:38–55
13. Redekopp L (1977) On the theory of solitary Rossby waves. *J Fluid Mech* 82:725–745
14. Caputo JG, Stepanyants YA (2003) Bore formation and disintegration into solitons in shallow inhomogeneous channels. *Nonlinear Process Geophys* 10:407–424
15. Longuet-Higgins MS (1993) Capillary gravity waves of solitary type and envelope solitons in deep water. *J Fluid Mech* 252:703–711
16. Philips OM (1974) Nonlinear dispersive waves. *Ann Rev Fluid Mech* 6:93–110
17. Helal MA, Molines JM (1981) Nonlinear internal waves in shallow water. A theoretical and experimental study. *Tellus* 33:488–504
18. Benjamin TB, Bona JL, Mahoney JJ (1972) Model equations for long waves in nonlinear dispersive systems. *Philos Trans A Royal Soc* 272:47–78
19. Camassa R, Holm D (1992) An integrable shallow water equation with peaked solitons. *Phys Rev Lett* 71:1661–64
20. Zakharov VE (1968) Stability of periodic waves of finite amplitude on the surface of a deep fluid. *J Appl Mech Tech Phys* 2:190–194
21. Hasimoto H, Ono H (1972) Nonlinear modulation of gravity waves. *J Phys Soc Jpn* 33:805–811
22. Benney DJ, Roskes G (1969) Wave instabilities. *Stud Appl Math* 48:377–385
23. Davey A, Stewartson K (1974) On three dimensional packets of surface waves. *Proc Royal Soc Lond A* 338:101–110
24. Fokas AS, Santini PM (1990) Dromions and a boundary value problem for the Davey–Stewartson I equation. *Physica D* 44:99–130
25. Kundu A (ed) (2007) *Tsunami and Nonlinear Waves*. Springer, Berlin
26. Dias F, Dutykh D (2007) Dynamics of tsunami waves. In: Ibrahimbegovic A, Kozar I (eds) *Extreme man-made and natural hazards in dynamics of structures*. NATO security through Science Series. Springer, Berlin, pp 201–224
27. Dutykh D, Dias F (2007) Water waves generated by a moving bottom. In: Kundu A (ed) *Tsunami and Nonlinear Waves*. Springer, Berlin, pp 65–94
28. Okada Y (1992) Internal deformation due to shear and tensile faults in a half space. *Bull Seism Soc Am* 82:1018–1040
29. Carrier GF, Wu TT, Yeh H (2003) Tsunami runup and drawdown on a plane beach. *J Fluid Mech* 475:79–99
30. Banerjee P, Politz FF, Burgman R (2005) The size and duration of the Sumatra–Andaman earthquake from far-field static offsets. *Science* 308:1769–1772
31. Constantin A, Johnson RS (2006) Modelling tsunamis. *J Phys A39*:L215–L217
32. Segur H (2007) Waves in shallow water, with emphasis on the tsunami of (2004). In: Kundu A (ed) *Tsunami and Nonlinear Waves*. Springer, Berlin, pp 3–29
33. Lakshmanan M (2007) Integrable nonlinear wave equations and possible connections to tsunami dynamics. In: Kundu A (ed) *Tsunami and Nonlinear Waves*. Springer, Berlin, pp 31–49
34. Alpers W, Wang-Chen H, Cook L (1997) Observation of internal waves in the Andaman Sea by ERS SAR. *IEEE Int* 4:1518–1520
35. Hyder P, Jeans DRG, Cauquill E, Nerzic R (2005) Observations and predictability of internal solitons in the northern Andaman Sea. *Appl Ocean Res* 27:1–11
36. Apel JR, Holbrook JR (1980) The Sulu sea internal soliton experiment, 1. Background and overview. *EOS Trans AGU* 61:1009
37. Ostrovsky LA, Stepanyants YA (2005) Internal solitons in laboratory experiments: Comparison with theoretical models. *Chaos* 15(1–28):037111
38. Joseph RI (1977) Solitary waves in a finite depth fluid. *J Phys A10*:L225–L227
39. Benjamin TB (1967) Internal waves of permanent form in fluids of great depth. *J Fluid Mech* 29:559–592
40. Kundu PK, Cohen IM (2002) *Fluid Mechanics*, Second Edition. Academic Press, San Diego
41. Benney DJ (1966) Long nonlinear waves in fluid flows. *Stud Appl Math* 45:52–63
42. Boyd JP (1980) Equatorial solitary waves. Part I: Rossby solitons. *J Phys Oceanogr* 10:1699–1717
43. Susanto RD, Zheng Q, Xiao-Hai Y (1998) Complex singular value decomposition analysis of equatorial waves in the Pacific observed by TOPEX/Poseidon altimeter. *J Atmospheric Ocean Technol* 15:764–774
44. Lee S, Held I (1993) Baroclinic wave packets in models and observations. *J Atmospheric Sci* 50:1413–1428
45. Tan B (1996) Collision interactions of envelope Rossby solitons in a barotropic atmosphere. *J Atmospheric Sci* 53:1604–1616
46. Maxworthy T, Redekopp LG (1976) Theory of the Great Red Spot and other observed features of the Jovian atmosphere. *Icarus* 29:261–271
47. Caputo JG, Stepanyants YA (2007) Tsunami surge in a river: a hydraulic jump in an inhomogeneous channel. In: Kundu A (ed) *Tsunami and Nonlinear Waves*. Springer, Berlin, pp 97–112
48. Tsuji T, Yanuma T, Murata I, Fujiwara C (1991) Tsunami ascending in rivers as an undular bore. *Nat Hazard* 4:257–266
49. Koch SE, Pagowski M, Wilson JW, Fabry F, Flamant C, Feltz W, Schwemmer G, Geerts B (2005) The structure and dynamics of atmospheric bores and solitons as determined from remote sensing and modelling experiments during IHOP, AMS 32nd Conference on Radar Meteorology, Report JP6J.4

## Books and Reviews

- Lamb H (1932) *Hydrodynamics*. Dover, New York
- Miles JW (1980) Solitary waves. *Ann Rev Fluid Mech* 12:11–43
- Stoker JJ (1957) *Water Waves*. Interscience, New York
- Dauxois T, Peyrard M (2006) *Physics of Solitons*. Cambridge University Press, Cambridge
- Johnson RS (1997) *An Introduction to the Mathematical Theory of Water Waves*. Cambridge University Press, Cambridge
- Hammack JL (1973) A note on tsunamis: their generation and propagation in an ocean of uniform depth. *J Fluid Mech* 60:769–800
- Drazin PG, Johnson RS (1989) *Solitons: An Introduction*. Cambridge University Press, Cambridge

- Mei CC (1983) *The Applied Dynamics of Ocean Surface Waves*. Wiley, New York
- Scott AC et al (1973) The soliton: a new concept in applied science. *Proc IEEE* 61:1443–83
- Scott AC (ed) (2005) *Encyclopedia of Nonlinear Science*. Routledge, New York
- Sulem C, Sulem P (1999) *The Nonlinear Schrödinger Equation*. Springer, Berlin
- Helfrich KR, Melville WK (2006) Long nonlinear internal waves. *Ann Rev Fluid Mech* 38:395–425
- Bourgault D, Richards C (2007) A laboratory experiment on internal solitary waves. *Am J Phys* 75:666–670

## Submarine Landslides and Slow Earthquakes: Monitoring Motion with GPS and Seafloor Geodesy

BENJAMIN A. BROOKS<sup>1</sup>, JAMES H. FOSTER<sup>1</sup>,  
JEFFREY J. MCGUIRE<sup>2</sup>, MARK BEHN<sup>2</sup>

<sup>1</sup> School of Ocean and Earth Science and Technology,  
University of Hawaii, Honolulu, USA

<sup>2</sup> Department of Geology and Geophysics, Woods Hole  
Oceanographic Institution, Woods Hole, USA

### Article Outline

Glossary

Definition of the Subject

Introduction

Monitoring Motion:

Subaerial and Submarine Geodetic Methods

Data Analysis and Inversion

Discussion: Slow Earthquake

and Submarine Landslide Process

Future Directions: Slow Earthquakes

and Submarine Landslide Monitoring

Acknowledgments

Bibliography

### Glossary

**Submarine landslide** A gravitational mass failure feature on the seafloor.

**Slow earthquake** A discrete slip event that produces millimeter to meter-scale displacements identical to those produced during earthquakes but without the associated seismic shaking.

**GPS** The Global Positioning System consists of a constellation of at least 24 medium earth orbiting satellites transmitting two or more microwave frequencies for use in precise positioning.

**Seafloor geodesy** The application of geodetic methods (studies of the change in the shape of the earth's surface) applied to a submarine environment.

### Definition of the Subject

The term 'submarine landslide' encompasses a multitude of gravitational mass failure features at areal scales from square meters to thousands of square kilometers. Here, we concentrate on the large end of that spectrum, namely the submarine landslides that, when they move either in contained slip events or catastrophically, can generate surface

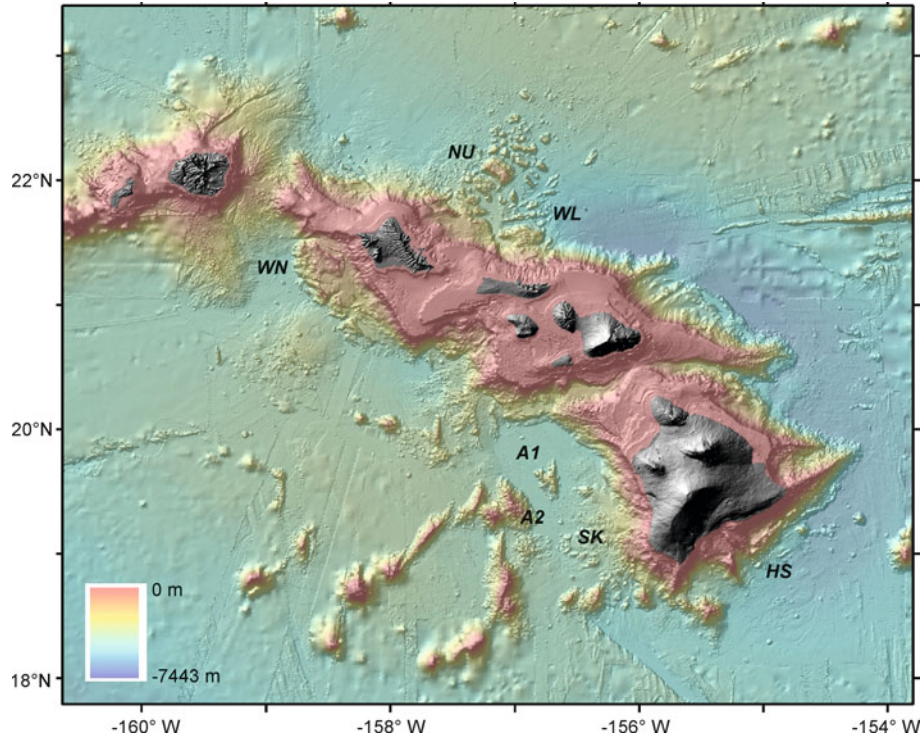
displacements equivalent to  $> M6$  earthquakes and/or hazardous tsunami.

The term 'slow earthquake' describes a discrete slip event that produces millimeter to meter-scale displacements identical to those produced during earthquakes but without the associated seismic shaking. Slow earthquakes, primarily associated with tectonic fault zones, have been recognized and studied with increasing frequency in the past decade largely due to the decreasing cost and proliferation of Global Positioning System (GPS) geodetic networks capable of detecting the ground motion [1,2,3]. Recently, one such GPS network on the south flank of Kilauea volcano, has recorded multiple slow earthquakes on the subaerial portion of a large landslide system that occurs primarily in the submarine environment [4,5,6]. Because the bathymetric charts surrounding the Hawaiian islands are littered with the remnants of massive, catastrophically emplaced submarine landslides (Fig. 1) it is natural to wonder if a slow-slipping submarine landslide is a precursory stage of one that will ultimately fail catastrophically.

We see two principal reasons why monitoring submarine landslides and slow earthquakes associated with them is important. First, because catastrophic failure of submarine landslides can cause tsunami they represent significant hazards to coastal zones. Understanding and monitoring how slow slip may lead to accelerated slip and catastrophic failure is, therefore, very important in terms of hazard mitigation. Second, submarine landslide systems can be some of the most active as well as spatially confined deforming areas on earth and so they represent excellent targets of study for furthering our understanding of the general fault failure process. For instance a pertinent question for which we do not yet have an answer is: are fault frictional properties homogeneous enough that the occurrence of slow earthquakes on a detachment fault plane underlying a landslide could relieve stress on the fault or do the slow earthquakes in one region load a neighboring seismogenic patch bringing it closer to a large sudden failure (i. e. an earthquake)?

While installation and operation of GPS networks on land is now relatively routine and somewhat inexpensive, the in situ monitoring of submarine landslide motion represents a significant technical challenge with accordingly higher costs. Submarine geodesy e. g. [7], however, is a nascent and rapidly evolving field with relative and absolute positioning techniques being intensely studied and developed. The near future is sure to see many advances in our monitoring and understanding of submarine landslides and slow earthquakes due to the application of submarine geodetic techniques.





Submarine Landslides and Slow Earthquakes: Monitoring Motion with GPS and Seafloor Geodesy, Figure 1  
 Topographic and bathymetric map of the Hawaiian Islands (data from <http://geopubs.wr.usgs.gov/i-map/i2809/>). Studied submarine landslides are indicated: HS, Hilina Slump; SK, South Kona; A1, Alika 1; A2, Alika 2; WL, Wailau; NU, Nu'uuanu; WN, Wai'anae

## Introduction

### Submarine Landslides

The last 30 years have seen a dramatic increase in the recognition of submarine landslides world-wide, due largely to the increased prevalence and capability of swath and side-looking sonar mapping systems and systematic submarine mapping programs. For instance, the side-scan sonar mapping of the Hawaiian exclusive economic zone in the 1980s resulted in the discovery that massive submarine landslides are spatially distributed along the entire Hawaiian Ridge [8] (Fig. 1). These features are some of the largest landslides on the planet, with more than 70 attaining lengths greater than 20 km and some having lengths greater than 200 km and total volumes exceeding 5000 km<sup>3</sup>. Since then, submarine landslides associated with other volcanic islands e.g. [9,10,11], mid-ocean ridges [12], and continental margins [13] have also been studied. Of these, the slopes flanking volcanic islands, especially when they are in their steeper-sloped shield-building stage [14,15], tend to be particularly susceptible to landslide instability and so they have been the focus of much recent research in the Canary Islands [16,17] and

especially in the Hawaiian Islands [18,19,20,21,22,23]. For instance, most of the Hawaiian submarine landslides are thought to be inactive, except for those on the flanks of the Big Island e.g. [21].

The morphology of submarine landslide features is similar to their subaerial counterparts. In map view they generally exhibit lobate and hummocky bathymetry. In cross-section, a wedge-shaped region of deformed material thins down-slope and is underlain by a gently-sloping planar dislocation surface (sometimes referred to as a basal detachment or 'decollement') separating the deformed carapace from the underlying undeformed substratum (Fig. 2). An upslope extensional head-scarp region transitions into a contractional fold belt towards the toe. The normal faults in the upslope region typically intersect the surface at high angles (> 45 degrees) and are manifested as scalloped-shaped scarps at the head of the slide that separate regions of differentially tilted fault blocks; in the sub-surface they may continue at high angles until they intersect the basal decollement, or they may sole with depth either into the decollement or into another sub-horizontal slip-surface [24]. The contractional regions are characterized by folded and bulging layers, closed de-

pressions, and steep toes [20,21,25]. Moore et al. [8] separated the Hawaiian submarine landslides into two principal types: ‘slumps’ and ‘debris avalanches’. The slumps are wide (up to  $\sim 100$  km), deep-seated ( $\sim 10$  km thick), and have surface inclinations of up to 3 degrees while the debris avalanches are long (up to  $\sim 230$  km), shallowly-seated (50 m–2 km thick), and have surface slopes generally less than 3 degrees.

Concomitant with the mapping of the landslide features has been the increasing recognition that sudden submarine landslide movement can cause tsunamis with destructive implications for coastal societies e.g. [26]. A particularly well-known example of this scenario is the 1929 Grand Banks failure [27]. Moreover, in the 1990s alone workers have attributed at least 5 tsunami events to catastrophic landslide failure sources: (1) 1992 Flores Island, Indonesia [28]; (2) 1994 Mindoro, Phillipines [29]; (3) 1998 Papua New Guinea e.g. [30]; (4) Kocaeli, Turkey [31]; (5) 1999 Pentecost Island, Vanuatu [32]. In Hawaii, the  $M_w$  7.7 1975 Kalapana earthquake, most likely due to slip of the fault surface underlying an active submarine landslide [33,34,35], caused local loss of life and damage in Southern California.

From a hazards standpoint it is particularly important to understand how tsunamis are generated by submarine landslides because the propagation time between tsunami generation to runup is typically on the order of minutes. For instance, a catastrophic failure of the west side of the island of Hawaii would likely send tsunami waves around the Hawaiian islands that would reach the densely populated areas of Oahu’s Waikiki beaches in less than an hour and more likely  $\sim 30$  min (G. Fryer, personal communication, 2007). Simulating waves generated by a sudden, chaotic disturbance of the seafloor, as expected from a submarine landslide, is quite complicated e.g. [36,37] and not all workers agree on approaches or results. Murty [38], however, stressed that parameters such as slide angle, water depth, density, speed, duration of the slide are second order, while instantaneously displaced volume is likely the most important parameter controlling tsunami generation.

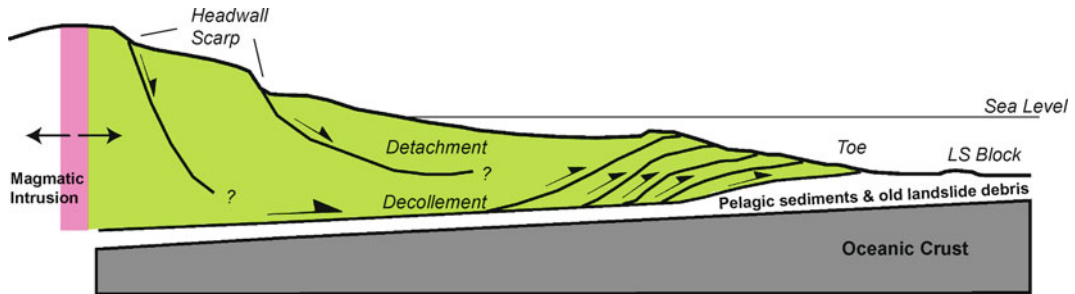
Despite the increasing awareness of their hazard, little is known about how submarine landslides actually move, largely because of the challenge of installing instruments and retrieving data from the submarine environment. Moore et al. [8] recognized that while slumps more likely move relatively slowly, debris avalanches could be deposited very rapidly based on, for instance, uphill flow of material in the distal portions of the landslide deposits. In agreement with this, the estimated downhill velocities from the 1929 Grand Banks event was 60–100 km/h [39]

whereas many studies have documented  $\sim 6$ –10 cm/yr horizontal and vertical velocities associated with a submarine landslide flanking the Island of Hawaii’s Kilauea volcano, the Hilina Slump [4,5,6,40,41,42]. Recently, GPS data from the Hilina Slump have elucidated that not only does the slump move at the above-stated, fairly smooth, background velocities, but also that the slump will occasionally deform in discrete accelerated cm-scale motions equivalent to  $M_6$  earthquakes but without the shaking [4,5,6]. These ‘slow earthquakes’ last hours and accommodate cm’s of ground displacement (Fig. 3). It is not currently known, however, how the occurrence of a slow earthquake in a submarine landslide system will affect its future movement by either making it more or less probable of failing catastrophically.

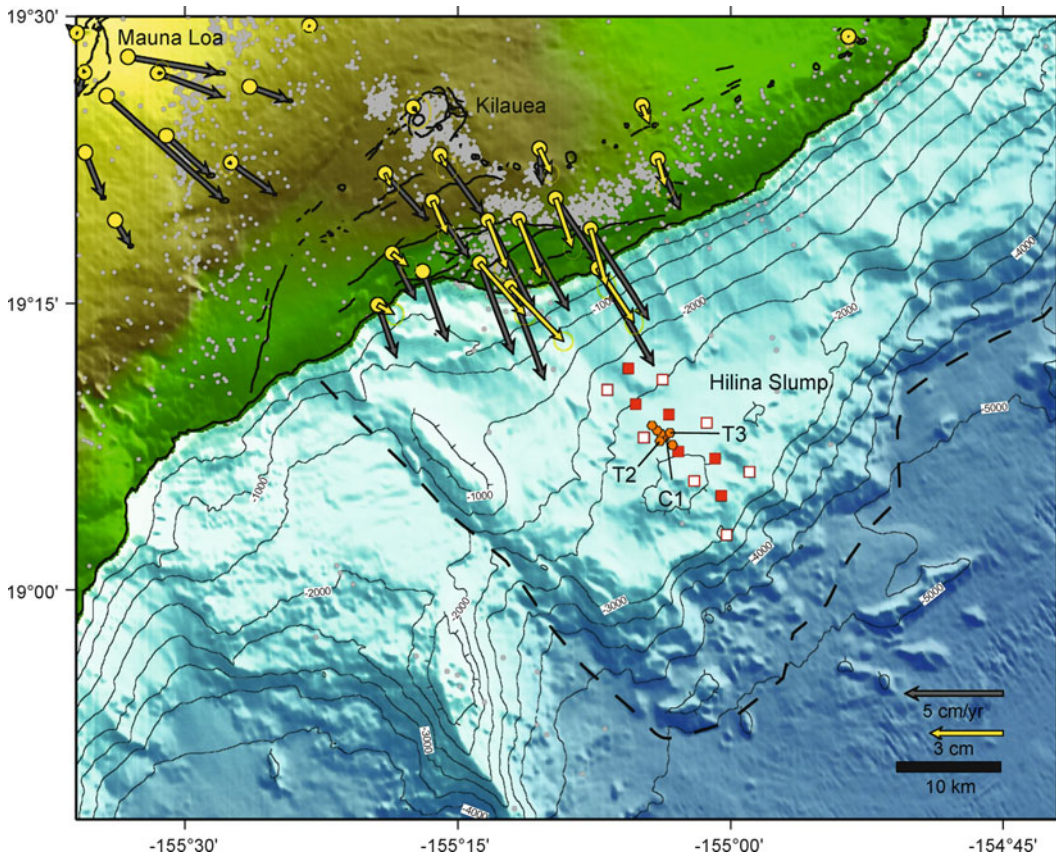
### Slow Earthquakes

The term ‘slow earthquake’ has been used to describe a variety of transient aseismic deformation phenomena including slow precursor events preceding large earthquakes [43,44,45,46,47,48], afterslip following earthquakes [49,50,51], variable fault creep rates [52,53,54,55], certain subduction zone thrust events with unusually long durations and large amplitude tsunamis for their size [56], and discrete fault-slip events that do not produce detectable seismic shaking but are accompanied by ground displacements very similar to those produced during earthquakes [1,2,3,4,5,6,57]. Hereafter, when we use ‘slow earthquake’ we will refer to this latter description although the term was first used in the modern literature to describe a slow precursor to the great Chilean 1960 earthquake [43].

In contrast to slow precursor events whose ground motions, like traditional earthquakes, are measured in seconds or minutes, slow earthquake (SE) displacements usually accrue over time periods ranging from hours to days and so they have been typically sensed with geodetic methods. In the last decade, the proliferation of continuous GPS (CGPS) networks has led to numerous SE observations and the discovery of some very rich behavior. In certain regions SEs have occurred with very regular periods [2,4,58], they are often associated with non-volcanic tremor [59,60] and, apparently, SEs follow very different scaling laws (moment vs. duration) than traditional earthquakes [61]. Explanations for SE slip behavior has varied to date. For subduction zones, the combination of deep ( $> 35$  km) SE sources and their association with tremor (a phenomenon initially thought to be caused by forced fluid flow [62] but more recently also explained in terms of shear failure [63,64]) led to one current hypothesis that SE



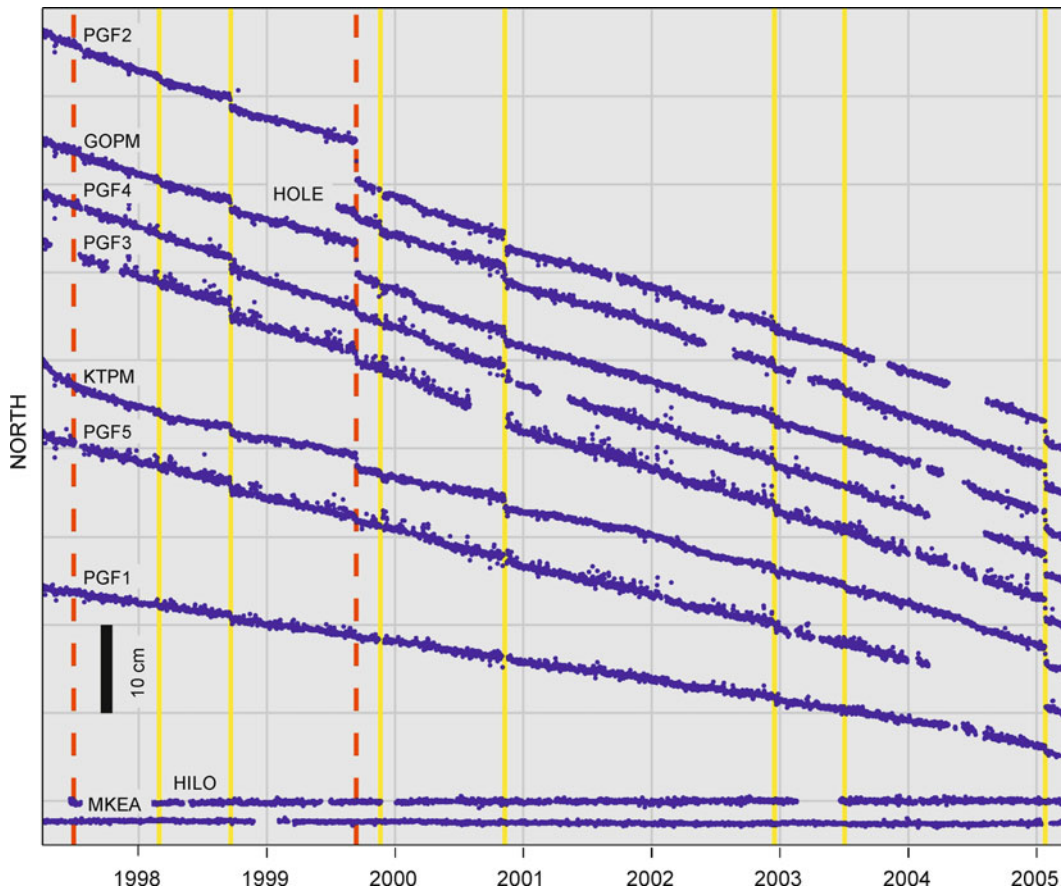
Submarine Landslides and Slow Earthquakes: Monitoring Motion with GPS and Seafloor Geodesy, Figure 2  
Schematic cross-section of a submarine landslide flanking an active ocean-island volcano (after [21])



Submarine Landslides and Slow Earthquakes: Monitoring Motion with GPS and Seafloor Geodesy, Figure 3  
Map of geodetic networks on the Hilina Slump overlain on topographic/bathymetric map. Yellow circles on land are CGPS sites operated jointly by the USGS Hawaiian Volcano Observatory, University of Hawaii, and Stanford University. Red squares offshore (open and filled) are seafloor geodetic sites operated by Scripps Institution of Oceanography. Orange circles offshore are acoustic extensometer sites deployed by our group with locations of transponders (T2,T3) and transceiver (C1) indicated. Grey vectors are average horizontal velocities from 1997–2005. Yellow vectors are horizontal motions from the January 2005 slow earthquake. Grey dots are earthquakes from the HVO catalog for the period May 2004–2005

mechanics are controlled by water released during metamorphic phase changes at the interface between a subducting and overriding plate [65,66]. Other explanations have invoked rate- and state-variable frictional behavior to sug-

gest that SEs occur preferentially at transitions between velocity strengthening and weakening regimes on a fault plane [67,68,69] and that temporally varying climatic load changes could help explain SE periodicity [58].



Submarine Landslides and Slow Earthquakes: Monitoring Motion with GPS and Seafloor Geodesy, Figure 4

Time series of north component of motion for selected GPS stations. Offsets of 7 slow earthquakes are identified by yellow lines. Red dashed lines are magmatic diking events

Due, in part, to their large magnitude deformation signal and the high concentration of CGPS networks focused on them, subduction zones have dominantly been the location of the most SEs to date [1,2,57,59,70,71,72]. Recently, the CGPS network on Kilauea volcano's mobile south flank has recorded multiple SEs [4,5,6] (Fig. 4) and it is through the Kilauea events that SEs have come to be associated with submarine landslides.

### Monitoring Motion: Subaerial and Submarine Geodetic Methods

Not surprisingly, much more is known about the motion of subaerial than submarine landslides. Geodetic measurements on land combined with contemporaneous measurements of other properties (pore-water pressures, strength of materials, etc.) have allowed, in some cases, a very thorough understanding of how landslide motion is related to driving forces such as gravitational stresses and rainfall.

For instance, Baum and Reid [73] instrumented a slow-moving submarine landslide in Honolulu's Manoa valley with extensometers recording at 15 minute intervals and rain gauges and found a direct correlation between rain fall and deformation events. Similarly, Malet et al. [74] showed that GPS-measured surface velocities increased to as high as 20 cm/day following periods of higher rainfall during May 1999 at the Super-Sauze earthflow in the French Alps. At a slightly different scale, Hilley et al. [75] used InSAR to simultaneously map deformation of multiple landslides in California's Berkely Hills at  $\sim$  monthly intervals and found that landslide motion correlated with times of high precipitation and that during the 1997–1998 El Nino event displacement rates doubled from the background rate of  $\sim$  27–38 mm/year, albeit with a  $\sim$  3 month time lag between the onset of motion and the high precipitation.

Much of the current knowledge that we have about the motion of submarine landslides comes from their easier-to-monitor subaerial portions, such as at Kilauea's Hilina

slump (Fig. 3). In the case of the Hilina Slump, fully 3/4 of the feature resides offshore at depths greater than 2000 m and the deformation monitoring network is necessarily concentrated on the down-dropped blocks near the head-wall scarp of the entire system. Recently, however, submarine methods have started to provide geodetic information from the ocean floor itself e.g. [7], e.g. [76,77,78,79]. As these methods become more cost-effective and widespread they will surely yield much insight into submarine landslide kinematics and, eventually, be employed in operational hazard monitoring/mitigation scenarios.

Whether a network is solely subaerial, submarine, or some combination of the two, it is important to consider that monitoring strategies can vary substantially depending on the time duration of the expected signal, the desired threshold of detection, and the desired time latency for individual solutions. For instance, simple detection and warning of catastrophic landslide failure needs the most rapid solution latency but coarse detection thresholds (meter rather than millimeter level, for instance) are appropriate. Conversely, if the goal is to detect small, potentially precursory motions such as slow earthquakes, then detection threshold must be as sensitive as possible along with solution latency being low.

### Subaerial Geodesy: GPS

A multitude of geodetic techniques, from mechanically- to electromagnetically-based, are currently employed on landslides to measure motion caused by a range of deformational phenomena spanning opening of small surface cracks to the motion of kilometer sized blocks e.g. [74]. Data from Global Positioning System (GPS) networks has increasingly contributed to the library of observations associated with landslide motion. In particular, for submarine landslides that are large enough so that their motion causes earthquakes or slow earthquakes [4,6,80], a technique which is suitable for inter-station distances measured in kilometers is most appropriate and so below, we concentrate on the use of GPS with submarine landslides.

**GPS** GPS networks capable of sub-cm to mm-scale 3-dimensional ground motion detection are now deployed in many of Earth's most actively deforming zones (see for example, [http://sps.unavco.org/crustal\\_motion/dxdt/](http://sps.unavco.org/crustal_motion/dxdt/)) and readers are referred to thorough reviews of the general technique and its application for geodynamic studies [81,82,83].

Crustal motion GPS networks are usually divided into two types: those that record data continuously (CGPS) and those whose individual monuments are occupied less

frequently in survey mode (SGPS). Depending on a variety of factors including the modernity of the receiver, the bandwidth of telemetry networks (should they exist), and the storage capacity of the archival center, CGPS sampling rate generally varies between once every 30 and 1 s (though most modern receivers are capable of sampling at frequencies higher than 1 Hz). SGPS sampling is more varied though it usually comprises re-occupation of sites at intervals ranging from months to years with occupation times of hours to days and sampling rates similar to CGPS. Accordingly, SGPS networks are more useful for wider ranging spatial characterization of deformation phenomena rather than for the rapid detection of motion or for tracking temporal evolution during transient events.

High rates of sampling alone, however, do not guarantee that CGPS network positional solutions will achieve their highest precision and/or accuracy. Assuming that all sites within a CGPS network have stable monuments and high-grade geodetic antennae and dual frequency receivers, the most important components of its error budget for deformation monitoring are: (1) integer ambiguity resolution; (2) orbital estimation; (3) atmospheric delay estimation; (4) antenna multipath; (5) satellite constellation geometry, and (6) intra-network baseline length. For networks monitoring landslides with spatial scales on the order of kms or tens of kms, however, the baselines are short enough that errors scaling with baseline lengths are small contributors to the error budget. In addition, as absolute positioning in a global reference frame is not essential, precise orbital estimation is less important. For the other error sources, freely available software packages such as GAMIT, GIPSY, and BERNESE (<http://facility.unavco.org/software/processing/processing.html>), and precise orbit processing centers such as the IGS (International GNSS Service, <http://igs.cb.jpl.nasa.gov/>) or the Scripps Orbit and Permanent Array Center (SOPAC, <http://sopac.ucsd.edu/>) usually allow good enough mitigation of these errors so that daily GPS solutions have resolution on the order of 2–5 mm in the horizontal and  $\sim 2$ –3 times worse in the vertical e.g. [83]. This resolution rule-of-thumb, however, generally applies to post-processed data with occupation times exceeding  $\sim 6$ –8 h, use of precise orbits, and the best atmospheric estimation techniques e.g. [84]. Because of the hours-to-days delay needed for estimating various grades of precise orbits and the computational time needed to estimate all the parameters for all sites with this much data this standard rule-of-thumb cannot necessarily be applied to a real-time or near real-time solution.

**Real- and Near-Real Time GPS Processing** There are a variety of processing techniques that may be suitable

for real-time or near-real time monitoring applications for the subaerial counterparts to submarine landslides. For instance, the well-known real time kinematic (RTK) positioning technique frequently employed by the surveying community applies differential corrections sent via radio link between stations to yield site position estimates with cm-scale precision over baselines up to  $\sim 10$  km in real-time [85]. The technique works best for baseline distances less than  $\sim 10$  km typically, because the assumption of correlated errors between stations is not necessarily valid and differential corrections cannot be accurately applied for larger baselines. Thus, RTK may be a suitable monitoring technique if a more spatially contained portion of a landslide, such as a fault zone, is a particularly good indicator of motion for the overall larger unit.

It is the number of measurement epochs required to resolve the integer-cycle phase ambiguity inherent with GPS data, however, that is the principal factor limiting the temporal latency of high resolution GPS positioning solutions. Kinematic techniques typically require minutes worth of GPS data (when sampled at 1–30 s) in order to resolve integer ambiguities for initialization and reinitialization if a cycle slip or loss of phase lock occurs during measurement, although progress is being made on mitigating ionospheric effects and reducing time of integer ambiguity resolution e.g. [86]. Regardless, if GPS station locations are chosen with relatively good sky-view then loss of lock due to poor satellite visibility should be minimal.

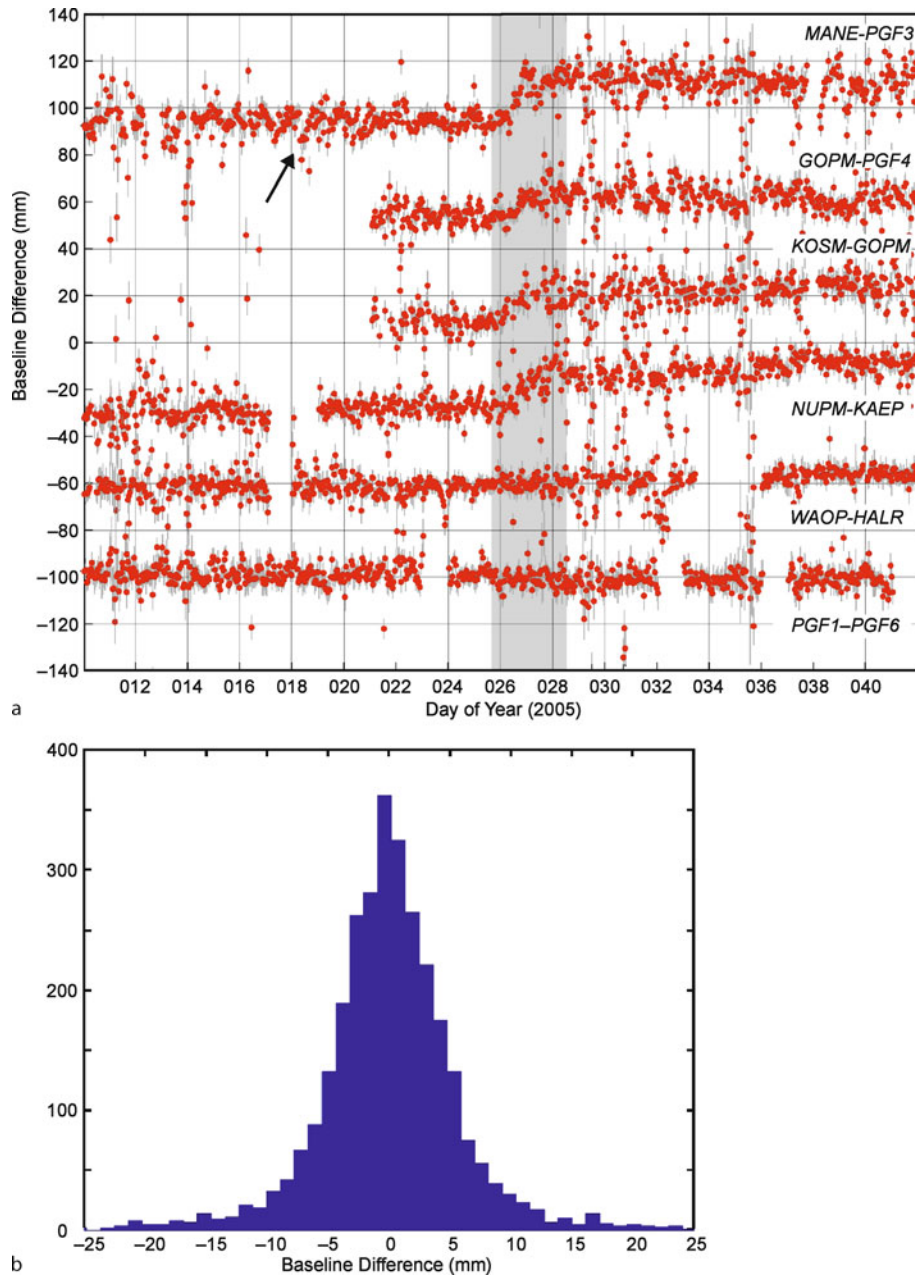
Recently, Bock et al. [87] developed a method of resolving integer ambiguities from a single epoch of dual frequency GPS phase and pseudo range data that provides independent epoch-by-epoch position estimates over baselines as large as  $\sim 40$  km. Their method allowed instantaneous positioning resolution of 1.5 cm in the horizontal and 7–8 times worse in the vertical coordinates. Langbein and Bock [88] applied the technique to determine offsets in positional time series due to slip events recorded by the Parkfield, California CGPS network which has similar spatial scales to a typical large submarine landslide. They found that offset sensitivity was  $\sim 5$  mm for a 2 s sampling window and this decreased to  $\sim 2$  mm when a 60 s window was used [88]. To the best of our knowledge this represents the current state-of-the art in terms of real-time detection of motion over spatial and temporal scales typical of submarine landslides.

At Kilauea, we have developed an automated *near* real-time processing strategy for our monitoring efforts at the Hilina Slump (Figs. 3 and 5) [89]. The CGPS data are telemetered hourly by radio modem back to the USGS Hawaii Volcano Observatory (HVO) where they are re-

trieved via FTP. We use a sliding window approach, collecting all data available within the most current two-hour period and performing a network solution using the PAGES [90] processing software. We use the 30 s, ionosphere free phase combination as the observable and use and hold fixed the IGS ultra-rapid orbits and apply the NGS antenna phase calibration patterns for each site [91]. We also apply the IERS standard solid Earth tide [92] and Schwiderski ocean tide loading corrections [93,94]. Every half hour we estimate a piece-wise, linear neutral atmosphere (troposphere) correction and one set of N–S and E–W neutral atmospheric gradient corrections for each site. We apply a ‘weak’ constraint of 10 cm to the atmospheric corrections based on examination of previous adjustments.

Figure 5a shows an example of baseline change time series from indicative stations derived from this approach and re-run in a simulated near real-time mode for the time period bracketing the slow earthquake from January 26–28, 2005 at Kilauea. Generally, the hourly baseline change noise levels are on the order of  $\pm 10$  mm, although because of some particularly large outliers, the standard deviation of the baseline changes (not including the time period of the SE) is closer to 20 mm (Fig. 5b). Some of the large excursions in the time series not associated with the SE are due to the high amplitude, strongly spatially and temporally varying atmospheric water vapor gradients often associated with tropical islands such as Hawaii e.g. [95]. For instance for the excursion in the MANE-PGF3 baseline in the middle of day 18 (Fig. 5a), the time series takes a sharp upward bend until the start of day 19 when it again begins to oscillate about its mean value from present days. This baseline change gradient is very similar to the onset of the SE near the beginning of day 26. It is clear from the remainder of the time series after day 27 that the  $\sim 25$  mm offset is permanent, however, and so indicative of a real deformation event.

In a real-time monitoring scenario, because of the 2 h temporal latency and the atmospheric noise levels on the order of 1/3 the maximum signal levels in the baseline change plots, it would probably require on the order of 12–24 h before this event could be definitively classified as a deformational event. This could certainly be useful to hazard mitigators for being aware that a slow event such as an SE was occurring, but not in the event of a rapidly accelerating catastrophic collapse where the detection and warning time must be on the order of minutes. For more rapid warning, more sophisticated filtering techniques could be employed (such as monument motion, and network-wide coherence assessment as described below for the Network Inverse Filter [96]) although regions



Submarine Landslides and Slow Earthquakes: Monitoring Motion with GPS and Seafloor Geodesy, Figure 5  
**a** Baseline difference (position – median position) and  $2\sigma$  errors for near-real time processing of selected sites at the Hilina Slump during early 2005. See text for description of processing. *Grey shaded region* indicates the time interval of the slow earthquake. *Black arrow* indicates an example of an anomalous trend due to atmospheric delays discussed in text. **b** *Histogram* of baseline difference values for all of the sites in a, excluding the time period of the slow earthquake

of large atmospheric gradient will always be hampered by low signal-to-noise ratios unless more sophisticated atmospheric mitigation techniques such as tomographic mapping [97] are employed in a real-time manner. Accord-

ingly, the small detection threshold levels of the epoch-by-epoch processing at Parkfield described above [88] must be taken in the context of the relatively low atmospheric delay environment present there.

## Submarine Geodesy

Subaerial geodetic methods have, until now, provided essentially the entirety of geodetic evidence for submarine landslide-related deformation including slow earthquakes. Our understanding of submarine landslide motion and slow earthquake mechanics stands to increase dramatically in the coming decade, however, owing to the rapidly advancing field of submarine geodesy which can now reliably provide in situ measurements of seafloor deformation. The submarine environment is particularly challenging for geodesy techniques especially if they require the propagation of energy across a medium (ocean water) that can exhibit highly spatially and temporally variable material properties.

**Direct Path and Indirect Acoustic Approach** One of the submarine geodesy techniques most frequently employed to date is direct path acoustic measurement of baseline length changes using acoustic transponders. This method has been used to make in situ observations of tectonic motions on the seafloor at the Juan de Fuca ridge where one study found cm-scale motion occurring over a period of a few days associated with an on-axis eruption [77] and another found no detectable motion over a period of a few years at a quiescent ridge segment [76]. By interrogating each other, pairs of instruments use two-way travel time measurements to constrain the inter-station distance. After correcting for variations in sound speed that result from changes in water temperature, salinity, pressure, and local tilt of the monument itself, these systems yield sub-cm precision for individual measurements. Operating frequencies typically range from 7.5 to 108 kHz and in deeper isothermal waters, the upward refraction of sound waves prevents signals from one transponder being received by the other at distances greater than  $\sim 1$  km [79]. Travel-time is determined simply by correlating the transmitted and received signals and picking the peak of the resultant correlogram. In quiet operating environments this can usually be done at the scale of  $\sim 5 \mu\text{s}$ , resulting in a  $\sim 4$  mm range error [7].

For the greater depth ranges, Sweeney et al. [79] devised an approach that allowed  $\pm 2$  cm resolution measurements for baselines up to 10 km at 2500–2600 m depths at a stable site on the Juan de Fuca plate. Sweeney et al. [79] suspended an acoustic interrogator hundreds of meters above the seafloor in a position acoustically visible to an array of seafloor-mounted instruments. In a manner similar to Spiess [98] they estimated the relative positions of the stations by moving the interrogator and collecting acoustic range data multiple locations.

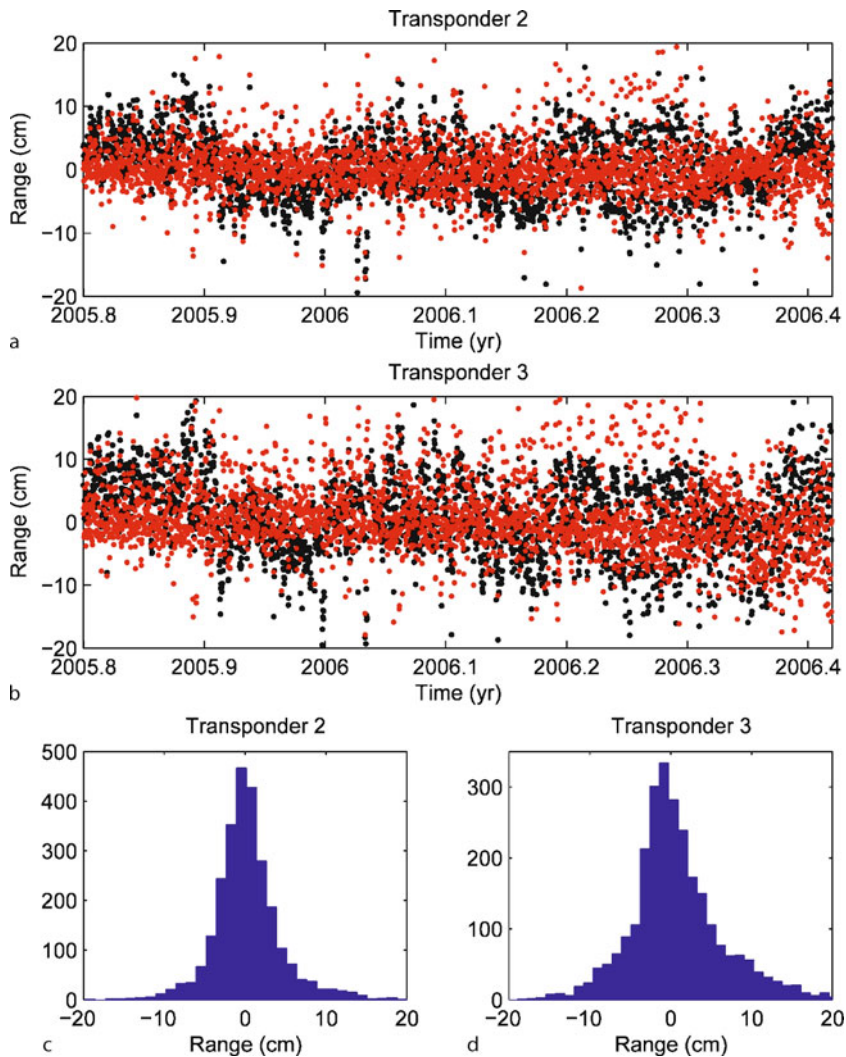
The focus of these initial deployments of the direct and indirect acoustic approach systems was on measuring annual tectonic rates, and not necessarily on capturing transient events such as SEs that occur over hours or days. Accordingly sampling rates in current seafloor geodesy projects typically do not exceed a few times per day, largely because providing adequate power to seafloor instruments is still financially prohibitive. However with cabled oceanographic observatories scheduled to come online in 2007–2010 (<http://www.neptunecanada.ca/>; <http://www.orionprogram.org/>) the power delivery problem could be solved at high priority seafloor sites.

From October, 2005 through June, 2006 we deployed seven 10 kHz Linkquest transponders mounted  $\sim 3$  m above steel tripods and spaced over a distance of  $\sim 3$  km on the Hilina Slump (Fig. 3). One of the goals of this pilot project was to evaluate the performance of the transponders for use in submarine landslide monitoring at significant depth; each unit operated at a depth between 2640 and 2690 m. Battery power restrictions for the  $\sim 8$  month duration of the project meant that the instruments ranged to one another 12 times per day.

In Fig. 6 we show range change time series from two transponder-pair baselines ( $\sim 530$  and 683 m respectively) for which data recovery was complete over the experiment's duration. For the other transponder pairs data recovery was not as complete because either: (1) the unit was knocked over by local mass-wasting events or (2) the baseline distance was too long and data dropouts occurred when the transponders lost sync with one another. The two way travel times were picked from the peak of the correlation function between the outgoing and received waveforms at one end of the baseline. The black dots in Fig. 6a and b show the raw measurements, while the red dots show the distance measurements corrected for the variations in sound speed due to temperature changes, which were measured by an external conductivity and temperature sensor mounted on the transponder frames. The raw time series show significant long-term trends due to temperature variations that are effectively removed using just the temperature measurements at the transponders. We did not correct for salinity variations because all of our conductivity sensors provided contaminated data due to clogging by the local mass wasting events.

For these baseline pairs which were oriented approximately perpendicular to the maximum expected motion of the Hilina Slump we expect essentially no motion for such short baselines and during such a short measurement period. At the  $1\sigma$  level individual measurement noise is  $\sim 4.1$  and 5.7 cm respectively, though it is clear that smaller, cm-scale changes would be detectable given a long





Submarine Landslides and Slow Earthquakes: Monitoring Motion with GPS and Seafloor Geodesy, Figure 6  
**a,b** Baseline difference (position – median position) for transponder-transceiver pairs (T2-C1, T3-C1 in Fig. 3) pairs from October 2005 – June 2006. See text for processing description. *Black dots*, raw baseline difference uncorrected for temperature. *Red dots*, baseline differences corrected for temperature. **c,d** Histogram of baseline difference values from **a** and **b**, respectively

enough time period. For instance, daily estimates of baseline length have approximately a 1 cm standard deviation. For rapid event detection, baseline changes would need to be on the order of 10 or more cm or measurement frequency would need to be increased.

**GPS-Acoustic Method** The acoustic measurements described above provide only relative measurements of baseline length changes; however, for submarine landslide monitoring efforts it may be desirable to place submarine and subaerial measurements in a single regional or global reference frame. Largely with the aim of furthering

seafloor active tectonics studies, a research group spearheaded by the late Fred Spiess conceived of, and have begun implementing integrated GPS and acoustic seafloor geodesy (GPS-A) studies, described in great detail by Spiess et al. [7] and references therein.

Briefly, GPS-A positions seafloor geodetic markers in a global reference frame by combining kinematic GPS positioning of a sea-surface platform (ship or buoy), precision underwater sound travel time measurement over km-scale path lengths, and a strategy for eliminating the large errors in travel time measurement arising from a spatially and temporally varying sound speed structure in the near

surface portion of the water column. The error mitigation strategy is based on the recognition that the location of the sea surface midpoint above the center of a triangle of seafloor transponders (themselves positioned via a near-bottom acoustic survey) is the point at which the three sea surface-to-seafloor travel times are identical and so, independent of sound speed.

Spieß et al. [7] reported repeatabilities of  $\pm 0.8$  cm and  $\pm 3.9$  cm in the north and east components, respectively, of seafloor measurements on the Juan de Fuca plate from 1994–1995. More recently, Gagnon et al. [78] reported velocities with  $\sim 5.5$  cm magnitudes and 0.6–0.8 cm  $1\sigma$  errors from surveys between 2001–2003 on the updip portion of the Nazca–South America subduction zone in the Peru trench.

**Optical Path Length** Zumberge [99] combined commercial subaerial surveying technology (an EDM) with an optical fiber strainmeter in order to devise a low-power, cost-effective system for the harsh seafloor operating environment. EDMs (electronic distance meters) typically can measure distances with 1–2 mm precision over several km by transmitting intensity modulated infra-red energy and measuring travel time from far away reflective surfaces. Zumberge [99] modified the EDM for seafloor geodesy by focusing the transmitted beam into the fiber's core and gathering the reflected light from the fiber's far-end into the EDM receiver optics.

The composite instrument comprises an optical fiber (enclosed in a hermetically sealed stainless steel casing) that stretches between two anchors, one 'active' and one 'passive', separated by several hundred meters on the seafloor. The active anchor houses the EDM and all supporting electronics and the passive anchor's purpose is to fix the fiber to the seafloor. In a test of their instrument, Zumberge et al. [100], report 1 mm scatter of distance measurements over a 500-m-long fiber during a 50 day long period.

One of the obvious advantages of the optical path length technique is the high precision, isolation from sea water-borne errors, and the continuous nature of the measurement. The disadvantage, however, is that the technique is limited to short baselines ( $< \sim 1$  km) and so the precise location of straining regions must be known a priori. Nonetheless, for monitoring purposes with very well-defined targets such as the headwall portion of an active landslide, the technique holds much promise.

**Pressure Sensor Vertical Deformation** Phillips et al. [40] recently developed a new technique that used pressure sensors in campaign-mode repeat surveys to mea-

sure vertical deformation rates of the seafloor at depths exceeding 2000 m at the offshore portion of the Hilina Slump. Depth and pressure are related to one another in a straightforward manner through the hydrostatic equation, though for measuring seafloor deformation rates at the cm/yr level many other reductions must be performed (Phillips et al. [40] – also her thesis). The pressure sensor method is technologically challenging, requires much ship time, and requires during each site visit both a geodetic monument and a pressure sensor (lowered from a ship) to be operating in close proximity to one another at depth on the seafloor. Because, in each revisit it was not practical to collocate the pressure sensor in a repeatable fashion on the benchmark, Phillips et al. [40] found it necessary to measure the vertical offset between the pressure sensor and the benchmark via acoustic ranging.

Important error sources for this technique are poor knowledge of  $\alpha$  the specific volume of the water column (the reciprocal of density) and the speed of sound in depth. Due to changing tides, secular and seasonal barometric changes it is also necessary to utilize a common, stable reference, such as mean sea level (MSL) which, in turn, requires estimation of the geopotential anomaly.

Ultimately, Phillips et al. [40] conclude that their data show significant vertical deformation ( $9 \pm 2.4$  cm/yr) in the mid-section of the HS and negligible deformation towards the outer bench (just inboard of the toe of the entire landslide feature). These data are consistent with dislocation models delimiting the potential amount ( $25.0\text{--}28.1 \pm 7.3$  cm/yr), spatial extent ( $24.8\text{--}27.0 \pm 0.5$  km seaward of Kilauea's East rift zone), and depth (7 km) of the principal slipping surface below the landslide. This is a significant advance not only because it represents the first data set that allows a glimpse of how strain is partitioned in a massive submarine landslide such as the Hilina Slump, but also because it represents an ongoing monitoring project at a submarine landslide, albeit at temporally sparse sampling rate.

## Data Analysis and Inversion

In addition to the challenge of acquiring geodetic data from submarine or even subaerial landslides is the added challenge of inferring sub-surface fault geometry and/or deformation processes, with attendant realistic error bounds on estimated parameters, from a data set of earth surface displacements. We identify two principal types of analysis modes: (1) process-based and (2) hazards-based. Process-based analysis focuses on deriving the most accurate and, if possible, complete assessment of the factors involved in the observed landslide motion. Hazards-based

analysis has two primary components: (a) rapid warning and (b) long-term hazards estimation. Rapid warning comprises, with the smallest temporal latency possible, providing information regarding the current state of the landslide system and how it relates to current, short- and long-term dangers to life and infrastructure. Long-term estimation comprises making probabilistic statements about the components and potential future behavior of the system and so temporal latency is not a limiting consideration.

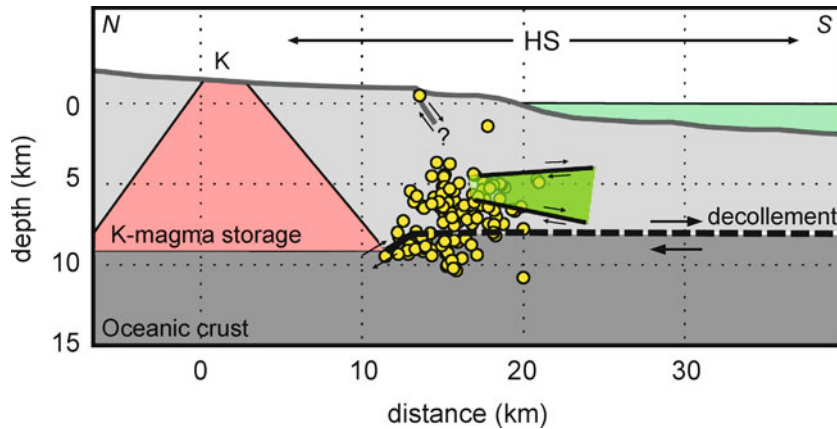
For either of the two types of analysis modes, the first objective is to usually try to relate the measured displacements  $d$ , of monuments at the surface of the earth to the source parameters,  $m$ , of the buried feature causing the deformation (usually considered a fault) through  $G(m)$ , a model of the deformation process. For instance, one very common model formalism, the dislocation in an elastic half-space, is a non-linear problem whose model vector  $m$  contains 9 parameters describing the location, orientation, and relative displacement of the dislocation approximating a faulting source [101,102].

Determining the most likely components of  $m$  and associated errors falls under the wide category of geophysical inversion for which there are a multitude of techniques e. g. [103]. In addition to the computational power available and the total amount of time allotted to the inversion, the choice of inversion method depends upon the nature of the model (for instance, linear vs. non-linear), the number of parameters in  $m$ , and the computational cost associated with individual model realizations and misfit assessments. Furthermore, it must be decided if the goal of an inversion is to obtain a rapid solution (a solution that satisfies some predetermined misfit level), an optimal solution (i. e. one 'best-fitting' solution), or rather to probe parameter space as thoroughly as possible with the goal of estimating posterior probability densities for each of the parameters for their value as indicators of resolution and uncertainty e. g. [104]. For instance, the most robust inversion method possible is the direct or 'grid' search, where the misfit for every possible permutation of  $m$  in parameter space is calculated. For more complicated or non-linear  $G(m)$  however the computational cost associated with each forward model run can make a grid-search time prohibitive, even for a process-based analysis. For these types of more difficult problems, inversion based on Monte Carlo sampling, which collects pseudo-random samples from multidimensional parameter space as a proxy for the problem's true posterior probability density,  $\sigma(m)$ , has been found to be quite successful [105,106]. Indeed, it is common practice in the tectonic geodesy community to use the Okada model combined with some type

of Monte-Carlo method to derive parameters of earthquakes and slow earthquakes from surface displacements observed with GPS [80,104,107].

The recent evolution of the work on the SEs at Kilauea demonstrates how important the inversion results are to the overall analysis either from a process-based or long-term hazards-based perspective. In their initial recognition of the November 2000 SE, Cervelli et al. [5] used a simulated annealing optimization routine [80] and the Okada model to invert the GPS observations and conclude that the most likely dislocation source for the SE occurred on a gently landward-dipping thrust fault plane at 5–6 km depth. Combined with the fact that the SE post-dated by 9 days a burst of local rainfall of nearly 1 m and reasonable estimates of local hydrologic parameters, Cervelli et al. [5] suggested that the increased pore pressure due to deeply percolating rainwater triggered SE motion by inducing a  $\sim 2$  MPa pressure decrease of the effective normal stress on their preferred fault plane. Brooks et al. [4], however, showed from Gibbs Sampling inversion [104] results of GPS data on three additional SEs that posterior distributions of estimated fault parameters allow a wide family of equivalently plausible solutions, ranging from deeper seated decollement solutions at  $\sim 8$  km depth to the more shallow fault plane favored by Cervelli et al. [5] (Fig. 7). Moreover, they showed that the other SEs were not associated with anomalous rainfall. Segall et al. [6] then used the same inversion method of Cervelli et al. [5] for the GPS data of the additional SEs relocations of high-frequency earthquakes triggered by the January 2005 SE, seismicity rate theory [108], and Coulomb stress modeling [109] to conclude that the SE (and other similar events) likely occurred at a depth of  $\sim 8 \pm 1$  km on the main decollement plane below the Hilina Slump.

In this case, the crucial addition to the source inversion was the added constraint of the triggered microearthquakes which were relocated to depths near the decollement [6]. In their earthquake relocations, however, Segall et al. [6] did not use the full waveform data; rather, they used a double-difference-derived mapping with manual picks, assuming a 1D velocity model, between triggered events and previous high-precision relocations and tomography from elsewhere at Kilauea [110]. Two other studies, [111,112] performed high precision relocations using waveform cross correlation data and found different depths for the same cluster of events as Segall et al. [6]. Got and Okubo [111] suggested that their relocated events (including those triggered by the 1998 SE) do not illuminate a sub-horizontal fault plane but rather a deeper, steeply south-dipping reverse fault. Wolfe et al. [112] found that the triggered seismicity from the four SEs identified by



Submarine Landslides and Slow Earthquakes: Monitoring Motion with GPS and Seafloor Geodesy, Figure 7

North-south cross-section of the Hilina Slump after Got and Okubo [111]. K, Kilauea. HS, Hilina slump. Black line below K is the seismic (solid) and aseismic (dashed) position of the decollement from Got and Okubo [111]. Two thicker black lines with green shaded area, are the family of geologically plausible dislocation solutions equivalently supported by inversion of GPS data. Yellow colored circles, locations of microearthquakes from HVO catalog for  $\pm 5$  days around each slow earthquake event

Brooks et al. [4] consistently relocates on distinct map-view clusters aligned in the direction of the SE displacements themselves and in a subhorizontal band with depths of  $\sim 5$  km. This is a solution consistent with Morgan et al. [21] who used seismic reflection data to identify moderately landward-dipping fault planes at similar depths. While the epicenters are well constrained in the Wolfe et al. [112] study, however, they were concerned that poor station geometry as well as near source velocity heterogeneity may bias the absolute depth of the relocations and the analyses may not be capable of distinguishing between shallow and deep fault zones. Thus, while the decollement is certainly the prime candidate on which a SE would occur (in agreement with the Segall et al. [6] suggestion), the analyses remain somewhat equivocal. This, in turn affects our best understanding of the most important hazard issue, the propensity of the decollement for a catastrophic failure. In order to more definitively answer these questions the above-mentioned research groups in conjunction with the Hawaiian Volcano Observatory of the US Geological Survey (the group charged with hazards analysis and mitigation for the region) have teamed up in a joint seismological and geodetic research project aimed at definitively constraining the depths of triggered seismicity, and, hopefully, the depth of SE sources in the region.

While the above example focuses mostly on process and long-term hazards assessment, recent developments focused on deriving the time-dependent history of slip during a deformation event could also potentially be used for automated real-time inversion and event detec-

tion [96,113]. The Network Inverse Filter (NIF) [96] is a recursive Kalman filter algorithm that operates on either processed position estimates or raw phase data (rather than derived displacement rates) from an entire network of geodetic stations, includes a stochastic description of local benchmark motion, and finds a non-parametric description of slip rate on a fault plane as a function of time. In current implementations, the NIF employs Green's functions relating slip to surface displacement computed from the analytical solutions for a dislocation in an elastic half-space [101] although there is no reason other deformation models could not be employed [113]. Application of the NIF to the 1999 Cascadia SE allowed McGuire and Segall [113], for instance, to determine that slip rate on the slipping fault plane took up to  $\sim 20$  days to reach its peak and that the southern portion of the fault had finished slipping before the northern portion began to slip. Cervelli et al. [5] and Segall et al. [6] also applied the NIF to the Kilauea SEs to derive source-time functions for the events. One caveat, however, is that the fault plane is not solved for by the NIF, rather, it must be known and held fixed a priori. For subduction zones with greater spatial geodetic coverage above the slipping portion of the fault this condition may be satisfied more satisfactorily than at submarine landslides.

#### Discussion: Slow Earthquake and Submarine Landslide Process

Largely because of CGPS subaerial measurements at Kilauea there are now a suite of observations of slow earth-

quakes related to submarine landslide motion. As it seems that SEs are a fairly general fault slip phenomenon, occurring at a variety of subduction zone locales globally, it stands to reason that SEs may be common to submarine landslide-related deformation, at least on the flanks of ocean island volcanoes. Clearly, as seafloor geodesy projects become more common we will learn if this is the case or not. It is also clear that the search for a theoretical understanding of the slow earthquake phenomena will be coupled to further understanding the motion of submarine landslides.

Currently, explanations for slow earthquake slip behavior focus on rock mechanic theory that follows an empirically derived rate- and state-variable frictional constitutive law based on laboratory experiments e. g. [114]. (The description below follows the summary and notation in Scholz [114]). The Dieterich–Ruina or ‘slowness’ law is expressed as:

$$\tau = [\mu_0 + a \ln(V/V_0) + b \ln(V_0\theta/\zeta)]\sigma,$$

where  $\tau$  is shear stress,  $\mu_0$  is the steady-state friction,  $\sigma$  is effective normal stress,  $V$  is slip velocity,  $V_0$  is a reference velocity,  $a$  and  $b$  are frictional material properties,  $\zeta$  is the critical slip distance and  $\theta$  is a state variable that evolves according to:

$$d\theta/dt = 1 - \theta V/\zeta.$$

The frictional stability of the system, then, depends on  $(a - b)$ , the velocity dependence of steady-state friction, defined:

$$a - b = \delta\mu/\delta(\ln(V)).$$

In the context of a simple spring-slider model approximating a slipping fault, the boundary between the stable and unstable frictional regimes will occur at a critical value of effective normal stress,  $\sigma_c$ , given by:

$$\sigma_c = k\zeta / (a - b),$$

where  $k$  is the stiffness. When  $(a - b) > 0$  the material is said to follow ‘velocity strengthening’ behavior and the system is stable – earthquakes cannot nucleate in this regime and earthquakes propagating into such regions will be abruptly terminated. When  $(a - b) < 0$  the material is said to follow ‘velocity weakening behavior and the system is unstable for  $\sigma \geq \sigma_c$  – earthquakes will nucleate in this regime. When  $\sigma \leq \sigma_c$  the system exhibits oscillatory behavior and is said to be ‘conditionally stable’, it is stable under quasi-static loading but requires a discrete velocity

perturbation in order for earthquakes to nucleate. Others interpret these relations slightly differently suggesting that nucleation occurs when matrix stiffness (scale and stress-rate dependent) drops to a critical value over a spatial scale large enough to promote rupture [68,115].

It may be in the boundary of this stability transition where slow earthquake slip behavior arises. For instance, a common observation of most of the subduction zone SEs is that they occur down-dip of the ‘locked’ zone or near the base of the seismogenic zone where earthquakes nucleate [1,57,58,69,71]. At these depths in subduction zones temperatures are close to the 450°C temperature at which feldspar starts to exhibit plastic behavior and so conditionally stable behavior would be expected [114], although it has recently been shown that transient oscillatory behavior may also arise naturally from system dynamics alone [116]. Faults in the conditionally stable regime, under steady-state loading, slip aseismically unless a perturbation to the system is large enough to push the fault across the stability boundary triggering an earthquake [117]. If the perturbation, however, is not quite large enough to push the fault across the stability regime then a period of stable sliding at increased velocity will occur (i. e. a slow earthquake) as the fault evolves back to steady-state [117]. It is not yet clear, however, how the explanations for subduction-zone related SEs may translate to the submarine landslide environment. While many subduction zone SEs apparently occur down-dip of the ‘locked’ zone where earthquakes nucleate along the subduction interface, at Kilauea, SEs apparently occur up-dip of the zone where the majority of earthquakes occur [4,6].

Recent studies have more explicitly focused on helping to explain SEs in terms of the rate- and state-variable formalism. For instance, Kato [67] used rate- and state friction to simulate the effect that ‘asperities’, velocity weakening regions surrounded by velocity strengthening regions, have on slip behavior and found that episodic SEs occur when the velocity-weakening patch size is close to the critical size of earthquake nucleation. Similarly, Liu and Rice [68] found that when they applied along-strike variations in the frictional  $(a - b)$  parameter in models of subduction zone processes, that SE-type transient deformation events emerged spontaneously near the down-dip end of the seismogenic zone. They suggested that the down-dip end of the seismogenic zone is likely to be in the conditionally stable boundary between unstable- and stable regimes and that this could allow the SE behavior. More recently, Lowry [58] employed further theoretical implications of rate- and state-variable friction under resonant loading conditions [115,118] to suggest that Earth’s response to climatic redistribution of atmospheric, hydro-

spheric, and cryospheric loads could lead to resonant fault slip behavior that could explain observed periodic slow earthquakes on the Cocos-North America plate boundary at Guerrero, Mexico.

Shibazaki and Shimamoto [119] recently proposed an alternate model of slow earthquakes that was motivated by laboratory experiments on the velocity dependence of frictional stability. They introduced a cutoff velocity to the rate-state formulation to mimic laboratory results where fault surfaces transition from velocity weakening at low sliding velocities to velocity strengthening at higher velocities. Implementing the laboratory values for these transitions reproduced the slip ( $10^{-7}$  m/s) and rupture propagation velocities (km/day) seen in subduction zone environments. Their model predicts a linear relationship between these two velocities that can in principle be inferred from high quality continuous geodetic measurements of slow earthquakes. This relationship may be an important way to relate observations of slow earthquakes to mechanical models, particularly in regions such as Kilauea where the fault planes are shallow enough ( $< 10$  km) to allow the details of the rupture to be resolved with high quality instrumentation. We note additionally that other recent model parameterizations [116] yield recurrence intervals and propagation velocities similar to observed events and Lowry [58] also suggested means of relating geodetic observations and model parameters.

Two additional factors that may modulate the frictional properties of faults in submarine landslide environments include local hydrologic and magmatic forces (in regions of active volcanism) [23]. For instance, the Nov. 2000 SE at Kilauea post-dated by 9 days an intense rain storm of nearly 1 m at the southeastern Big Island [5]. Cervelli et al. [5] estimated permeability using regionally appropriate hydrologic parameters for porosity and fault zone diffusivity and suggested that the 1 m of rain could have triggered SE motion by inducing a  $\sim 2$  MPa pressure decrease of the effective normal stress on a gently landward dipping fault at  $\sim 5$  km depth. Brooks et al. [4], however, showed that 3 other SEs at Kilauea were not preceded by anomalous rainfall and that other periods of anomalous rainfall were not accompanied by SEs. Additionally, prior to these events Iverson [23] was skeptical of rainfall triggers as he found from a rigid wedge analysis that, in order for ground water head gradients to be large enough to destabilize the Hilina Slump, implausibly large clay layers ( $\sim 200$  m thick) or very low hydraulic diffusivity ( $\sim 10^{-11}$  m<sup>2</sup>/s) needed to be present. Others, however, have found that magmatic injection-induced mechanical and thermal pressurization of fluids may help to explain Canary and Cape Verdes Islands flank instabilities [120].

Excess shear stresses exerted on a basal decollement because of rift zone magma injection have also been shown to be of sufficient magnitude to potentially cause slip [22,23]. Although others have suggested from geodetic observations of discrete dike events at Kilauea's rift zones that the dikes are injected passively, as a response to decollement slip, rather than as a trigger for it [121,122].

### Future Directions: Slow Earthquakes and Submarine Landslide Monitoring

It is clear that expanded seafloor geodetic monitoring of submarine landslides would be extremely important in bettering our understanding of both the slow earthquake process and the hazards associated with submarine landslides. For instance, at the Hilina Slump, it is not even known how much of submarine portion of the landslide actually displaces the seafloor during a slow earthquake. Given the logistical and financial challenges associated with seafloor geodesy, however, it is reasonable to ask if it is worth it to society to monitor submarine landslides. In the Hawaiian Islands, for instance, one compilation suggests that there have been at least 6 tsunami-generating landslide events in the past 300 000 years [123]. For comparison, the average 50 000 year recurrence interval for such events is 1 to 2 orders of magnitude larger than typical earthquake recurrence intervals in Southern California where substantial resources are focused on earthquake-cycle related monitoring (<http://www.wgcep.org/>). As discussed above, however, landslide-generated tsunami can be quite damaging both locally and regionally. One study simulated  $\sim 30$  m wave heights reaching the California coast within 6 h of a massive collapse of the Hilina Slump at Kilauea [124]. Clearly minimizing the impact of such an event would be of societal benefit.

Continuous submarine landslide deformation monitoring, although expensive, is not out of the realm of possibility, as costs for seafloor geodetic instrumentation are decreasing. In the coming years, in order to make continuous monitoring efforts more feasible, research will likely focus on two major technical challenges: (1) instrument power delivery and (2) data transmission. While cabling a network via seafloor pathways is certainly one way of satisfying both requirements, cabling may not be the best long-term solution for a number of reasons. First, the costs of large cable lengths and laying them on the seafloor is often measured in the millions of dollars. Second, especially for submarine landslides, the seafloor across which the cable need be lain can be very rough and cable failure can be a quite common occurrence. Third, if cable runs are long then annual maintenance efforts and cost can also be quite

high. Accordingly, techniques such as acoustic data transmission [125] and local power generation via a buoy, for instance, will be critical.

### Acknowledgments

We thank the Hawaiian Volcano Observatory of the US Geological Survey, in particular Asta Miklius, Maurice Sako, and Kevan Kamibayashi, for their collaboration and efforts in operating the CGPS network at Kilauea. We thank Paul Segall of Stanford University for sharing GPS data from this network. We thank Matt Gould and Nick Witzell of WHOI for engineering support. We thank Capt. Rick Myer and the crew of the R/V Kilo Moana for their help in deploying and retrieving the extensometers. We thank Mark Schenewerk for help with implementing PAGES. We thank Katie Phillips and Dave Chadwell for helpful discussion and sharing a manuscript in preparation. We thank Cecily Wolfe and Gerard Fryer for stimulating discussion. In particular we thank J. Moore, T. Lowry, and D. Chadwell for thorough and insightful reviews which improved this contribution. This research was supported by the Geophysics and Geochemistry programs of the US National Science Foundation, grants from School of Ocean and Earth Science and Technology at the University of Hawaii, and a grant from the Deep Ocean Exploration Institute of the Woods Hole Oceanographic Institution.

### Bibliography

1. Dragert H, Wang K, James TS (2001) A silent slip event on the deeper Cascadia subduction interface. *Science* 292:1525–1528
2. Miller MM, Melbourne TI, Johnson DJ, Sumner WQ (2002) Periodic slow earthquakes from the Cascadia subduction zone. *Science* 295:2423
3. Ozakawa S, Murakami M, Tada T (2001) Time-dependent inversion study of the slow thrust event in the Nankai trough subduction zone, southwestern Japan. *J Geophys Res* 106:782–802
4. Brooks BA, Foster JH, Bevis MF, Frazer LN, Wolfe CJ, Behn M (2006) Periodic slow earthquakes on the flank of Kilauea volcano, Hawaii. *J Earth and Planet Sci Lett* 246:207–216
5. Cervelli P, Segall P, Johnson K, Lisowski M, Miklius A (2002) Sudden aseismic fault slip on the south flank of Kilauea volcano. *Nature* 415:1014–1018
6. Segall P, Desmarais EK, Shelly D, Miklius A, Cervelli P (2006) Earthquakes Triggered by Silent Slip Events on Kilauea Volcano, Hawaii. *Nature* 442:71–74
7. Spiess FN et al (1998) Precise GPS/Acoustic positioning of seafloor reference points for tectonic studies. *Phys Earth Planet Inter* 108:101–112
8. Moore JG et al (1989) Prodigious Submarine Landslides on the Hawaiian Ridge. *J Geophys Res* 94:17465–17484
9. Krastel S et al (2001) Submarine landslides around the Canary Islands. *J Geophys Res* 106:3977–3997
10. Carracedo JC, Day SJ, Guillou H, Perez Torrado FJ (1999) Giant Quaternary landslides in the evolution of La Palma and El Hierro, Canary Islands. *J Volcan Geotherm Res* 94:169–190
11. Day SJ, Heleno da Silva SIN, Fonseca JFBD (1999) A past giant lateral collapse and present-day flank instability of Fogo, Cape Verde Islands. *J Volcan Geotherm Res* 94:191–218
12. Tucholke B (1992) Massive submarine rockslide in the rift-valley wall of the Mid-Atlantic Ridge. *Geology* 20:129–132
13. Driscoll N, Weissel JK, Goff JA (2000) Potential for large-scale submarine slope failure and tsunami generation along the US Mid-Atlantic coast. *Geology* 28:407–410
14. Moore JG (1964) Giant submarine landslides on the Hawaiian Ridge. *US Geol Survey Prof Paper D* 501:95–98
15. Moore JG, Fiske RS (1969) Volcanic substructure inferred from dredge samples and ocean-bottom photographs, Hawaii. *Geol Soc Am Bull* 80:1191–1202
16. Marti J, Hurlimann M, Ablay G, Gudmundsson A (1997) Vertical and lateral collapses on Tenerife (Canary Islands) and other volcanic ocean islands. *Geology* 25:879–882
17. Ward SN, Day S (2001) Cumbre Vieja Volcano – Potential collapse and tsunamis at La Palma, Canary Islands. *Geophys Res Lett* 28:3397–3400
18. Clague DA, Moore JG (2002) The proximal part of the giant submarine Waialau landslide, Molokai, Hawaii. *J Volcan Geotherm Res* 113:259–287
19. Coombs ML, Clague DA, Moore GF, Cousens BL (2004) Growth and collapse of Waianae Volcano, Hawaii, as revealed by exploration of its submarine flanks. *Geochem Geophys Geosyst* 5: doi:10.1029/2004GC000717
20. Morgan JK, Clague DA, Borchers DC, Davis AS, Milliken KL (2007) Mauna Loa's submarine western flank: Landsliding, deep volcanic spreading, and hydrothermal alteration. *Geochem Geophys Geosyst* 8:Q05002; doi:10.1029/2006GC001420
21. Morgan JK, Moore GF, Clague DA (2003) Slope failure and volcanic spreading along the submarine south flank of Kilauea volcano, Hawaii. *J Geophys Res* 108:2415, doi:10.1029/2003JB002411
22. Dieterich J (1988) Growth and Persistence of Hawaiian Volcanic Rift Zones. *J Geophys Res* 93:4258–4270
23. Iverson RM (1995) Can magma-injection and groundwater forces cause massive landslides on Hawaiian volcanoes? *J Volcan Geotherm Res* 66:295–308
24. Cannon EC, Bürgmann R, Owen SE (2001) Shallow normal faulting and block rotation associated with the 1975 Kalaupana earthquake, Kilauea volcano, Hawaii. *Bull Seismol Soc Am* 91:1553–1562
25. Moore JG, Normark WR, Holcomb RT (1994) Giant Hawaiian landslides. *Ann Rev Earth Planet Sci* 22:119–144
26. Bardet J-P, Synolakis CE, Davies HL, Imamura F, Okal EA (2003) Landslide Tsunamis: Recent Findings and Research Directions. *Pure Appl Geophys* 160:1793–1809
27. Hasegawa HS, Kanamori H (1987) Source Mechanism of the Magnitude 7.2 Grand Banks Earthquake of November 18, 1929: Double-couple or Submarine Landslide? *Bull Seismol Soc Am* 77:1984–2004
28. Yeh H et al (1993) The Flores Island Tsunami. *Eos, Transactions, Am Geophys Union* 74:371–373
29. Imamura F, Synolakis CE, Titov V, Lee S (1995) Field Survey of

- the 1994 Mindoro Island, Phillipines Tsunami. *Pure Appl Geophys* 144:875–890
30. Geist EL (2000) Origin of the 17 July 1998 Papua New Guinea Tsunami: Earthquake or Landslide? *Seismol Res Lett* 71:344–351
  31. Yalciner AC et al (1999) Field Survey of the 1999 Izmit Tsunami and Modeling Effort of New Tsunami Generation Mechanism. *Eos, Transactions, Am Geophys Union* F751(abstract):80
  32. Caminade JP et al (2001) Vanuatu Earthquake and Tsunami Caused Much Damage, Few Casualties. *Eos Trans Am Geophys Union* 81:641,646–647
  33. Nettles M, Ekstrom G (2004) Long-Period Source Characteristics of the 1975 Kalapana, Hawaii, Earthquake. *Bull Seismol Soc Am* 94:422–429
  34. Tilling RI et al (1976) Earthquake and Related Catastrophic Events, Island of Hawaii, November 29, 1975: A preliminary report. *US Geol Surv Circ* 740:33
  35. Day SJ, Watts P, Grilli ST, Kirby JT (2004) Mechanical models of the 1975 Kalapana, Hawaii earthquake and tsunami. *Mar Geol* 215:59–92, doi:10.1016/j.margeo.2004.11.008
  36. Okal EA, Synolakis CE (2003) A Theoretical Comparison of Tsunamis from Dislocations and Landslides. *Pure Appl Geophys* 160:2177–2188
  37. Ward S (2001) Landslide tsunamis. *J Geophys Res* 106:11201–11215
  38. Murty TS (2003) Tsunami wave height dependence on landslide volume. *Pure Appl Geophys* 160(10–11):2147–2153
  39. Fine IV, Rabinovich AB, Bornhold BD, Thomson RE, Kulikov EA (2005) The Grand Banks landslide-generated tsunami of November 18, 1929, preliminary analysis and numerical modeling. *Mar Geol* 215:45–57
  40. Phillips KA, Chadwell CD, Hildebrand JA (2008) Vertical deformation measurements on the submerged south flank of Kilauea volcano, Hawai'i reveal seafloor motion associated with volcanic collapse. *J Geophys Res* 113:B05106; doi:10.1029/2007JB005124
  41. Owen S et al (2000) Rapid deformation of Kilauea Volcano: Global Positioning System measurements between 1990 and 1996. *J Geophys Res B: Solid Earth* 105:18983–18998
  42. Delaney PT et al (1998) Volcanic spreading at Kilauea, 1976–1996. *J Geophys Res* 103:18003–18023
  43. Kanamori H, Stewart GS (1979) A slow earthquake. *Phys Earth Planet Inter* 18:167–175
  44. Sacks IS, Suyehiro S, Linde AT, Snoko JA (1978) Slow earthquakes and stress redistribution. *Nature* 275:599–602
  45. Ihmlle PF, Jordan TH (1994) Teleseismic search for slow precursors to large earthquakes. *Science* 266:1547–1551
  46. Kedar S, Watada S, Tanimoto T (1994) The 1989 Macquarie Ridge earthquake: Seismic moment estimation from long-period free oscillations. *J Geophys Res* 99:17893–17908
  47. McGuire JJ, Ihmlle PF, Jordan TH (1996) Time-domain observations of a slow precursor to the 1994 Romanche transform earthquake. *Science* 274:82–85
  48. Kanamori H, Cipar JJ (1974) Focal process of the great Chilean earthquake May 22:1960. *Phys Earth Planet Inter* 9:128–136
  49. Burgmann R et al (2001) Rapid aseismic moment release following the 5 December, 1997 Kronotsky, Kamchatka, earthquake. *Geophys Res Lett* 28:1331–1334
  50. Heki K, Miyazaki S-I, Tsuji H (1997) Silent fault slip following an interplate thrust earthquake at the Japan Trench. *Nature* 386:595–598
  51. Segall P, Burgmann R, Matthews M (2000) Time-dependent triggered afterslip following the 1989 Loma Prieta earthquake. *J Geophys Res B: Solid Earth* 105:5615–5634
  52. Linde AT, Gladwin MT, Johnston MJS, Gwyther RL, Bilham RG (1996) A slow earthquake sequence on the San Andreas Fault. *Nature (London)* 383:65–68
  53. Gwyther RL, Gladwin MT, Mee M, Hart RHG (1996) Anomalous shear strain at Parkfield during 1993–94. *Geophys Res Letters* 23:2425–2428
  54. Gao SS, Silver PG, Linde AT (2000) Analysis of deformation data at Parkfield, California: Detection of a long-term strain transient. *J Geophys Res* 105:2955–2967
  55. Lienkaemper JL, Galehouse JS, Simpson RW (1997) Creep Response of the Hayward Fault to Stress Changes Caused by the Loma Prieta Earthquake. *Science* 276:2014–2016
  56. Kanamori H, Kikuchi M (1993) The 1992 Nicaragua earthquake: a slow tsunami earthquake associate with subducted sediments. *Nature* 361:714–716
  57. Lowry AR, Larson KM, Kostoglodov V, Bilham R (2001) Transient fault slip in Guerrero, southern Mexico. *Geophys Res Lett* 28:3753–3756
  58. Lowry AR (2006) Resonant slow fault slip in subduction zones forced by climatic load stress. *Nature* 442(7104):802–805
  59. Ito Y, Obara K, Shiomi K, Sekine S, Hirose H (2006) Slow earthquakes coincident with episodic tremors and slow slip events. *Science* 26:503–506
  60. Rogers G, Dragert H (2003) Episodic Tremor and slip on the Cascadia Subduction Zone: The chatter of silent slip. *Science* 300:1942–1943
  61. Ide S, Beroza GC, Shelly DR, Uchide T (2007) A scaling law for slow earthquakes. *Nature* 447:76–79; doi:10.1038/nature05780
  62. Aki K, Fehler M, Das S (1977) Source mechanism of volcanic tremors: fluid driven crack models and their application to the 1963 Kilauea eruption. *J Volcan Geotherm Res* 141:259–287
  63. Rubenstein JL et al (2007) Non-volcanic tremor driven by large transient shear stresses. *Nature* 448:579–582
  64. Shelly DR, Beroza GC, Ide S (2007) Non-volcanic tremor and low-frequency earthquake swarms. *Nature* 446:305–307; doi:10.1038/nature05666
  65. Peacock S, Wang K (1999) Seismic Consequences of Warm Versus Cool Subduction Metamorphism: Examples from Southwest and Northeast Japan. *Science* 286:937–939
  66. Julian B (2002) Seismological detection of slab metamorphism. *Science* 296:1625–1626
  67. Kato N (2004) Interaction of slip on asperities: Numerical simulation of seismic cycles on a two-dimensional planar fault with nonuniform frictional property. *J Geophys Res* 109:B12306; doi:10.1029/2004JB003001
  68. Liu Y, Rice JR (2005) Aseismic slip transients emerge spontaneously in three-dimensional rate and state modeling of subduction earthquake sequences. *J Geophys Res* 110:B08307; doi:10.1029/2004JB003424
  69. Miyazaki S, McGuire JJ, Segall P (2003) A transient subduction zone slip episode in southwest Japan observed by the nationwide GPS array. *J Geophys Res* 108(B2):2087; doi:10.1029/2001JB000456
  70. Kostoglodov V et al (2003) A large silent earthquake in the Guerrero seismic gap, Mexico. *Geophys Res Lett* 30:1807; doi:10.1029/2003GL017219



71. Freymueller J et al (2001) The great alaska 'earthquake' of 1998–2001. *EOS Trans Am Geophys Un* 82(47)
72. Hirose H, Hirahara K, Kimata F, Fujii F, Miyazaki S (1999) A slow thrust slip event following the two 1996 Hyuganada earthquakes beneath the Bungo Channel, southwest Japan. *Geophys Res Lett* 26:3237–3240
73. Baum RL, Reid ME (1995) Geology, hydrology, and mechanics of a slow-moving, clay-rich landslide, Honolulu, Hawaii. In: Haneberg WC, Anderson SA (eds) *Clay and Shale Slope Instability*. Geol. Soc. of Am, Boulder, Colorado
74. Malet JP, Maquaire O, Calais E (2002) The use of Global Positioning System techniques for the continuous monitoring of landslides: application to the Super-Sauze earthflow (Alpes-de-Haute-Provence, France). *Geomorphology* 43:33–54
75. Hilley GE et al (2004) Dynamics of slow-moving landslides from permanent scatterer analysis. *Science* 304(5679):1952–1955
76. Chadwell CD et al (1999) No spreading across the southern Juan de Fuca Ridge axial cleft during 1994–1996. *Geophys Res Lett* 26:2525–2528
77. Chadwick WW, Embley RW, Milburn HR, Meining C, Stapp M (1999) Evidence for deformation associated with the 1998 eruption of Axial Volcano, Juan de Fuca Ridge, from acoustic extensometer measurements. *Geophys Res Lett* 26:3441–3444
78. Gagnon K, Chadwell CD, Norabuena E (2005) Measuring the onset of locking in the Peru-Chile trench with GPS and acoustic measurements. *Nature* 434:205–208
79. Sweeney AD, Chadwell CD, Hildebrand JA, Spiess FN (2005) Centimeter-level positioning of seafloor acoustic transponders from a deeply-towed interrogator. *Mar Geol* 28:39–70, doi:10.1080/01490410590884502
80. Cervelli P, Murray MH, Segall P, Aoki Y, Kato T (2001) Estimating source parameters from deformation data, with an application to the March 1997 earthquake swarm off the Izu Peninsula, Japan *J Geophys Res* 106:11217–11237
81. Dixon TH (1991) An Introduction to the Global Positioning System and Some Geological Applications. *Rev Geophys* 29:249–276
82. Hager BH, King RW, Murray MH (1991) Measurement of Crustal Deformation Using the Global Positioning System. *Ann Rev Earth Planet Sci* 19:351–382
83. Segall P, Davis JL (1997) GPS applications for geodynamics and earthquake studies. *Ann Rev Earth Planet Sci* 25:301–336
84. Eckl MC, Snay RA, Soler T, Cline MW, Mader GL (2001) Accuracy of GPS-derived relative positions as a function of inter-station distance and observing-session duration. *J Geodesy* 75:633–640
85. Leick A (2004) *GPS satellite surveying*. Wiley, Hoboken
86. Warnant R, Kutiev I, Marinov P, Bavier M, Lejeune S (2007) Ionospheric and geomagnetic conditions during periods of degraded GPS position accuracy: 1. Monitoring variability in TEC which degrades the accuracy of Real-Time Kinematic GPS applications. *Adv Space Res* 39:875–880
87. Bock Y, Nikolaidis RM, de Jonge PJ, Bevis M (2000) Instantaneous geodetic positioning at medium distances with the Global Positioning System. *J Geophys Res* 105:28233–28253
88. Langbein, J (2004) Noise in two-color electronic distance meter measurements revisited. *J Geophys Res* 109:B04406; doi:10.1029/2003JB002819
89. Brooks BA, Foster JF, Miklius A, Schenewerk M (2005) Extended GPS Network and Near Real-Time processing, Mauna Loa Volcano, Hawai'i. *Eos Trans AGU*, 86:Fall Meet Suppl
90. Schenewerk M, Dillinger W, Hilla S (2000) <http://www.ngs.noaa.gov/GRD/GPS/DOC/toc.html>
91. Mader G, MacKay JR (1996) Geoscience Laboratory, Office of Ocean and Earth Sciences NOS, NOAA. Silver Spring, Maryland
92. McCarthy D, Petit G (eds) (2003) *IERS Technical Note 32*
93. Cartwright DE, Edden AC (1973) Corrected tables of tidal harmonics. *Geophys J R Astron Soc* 33:253–264
94. Cartwright DE, Taylor RJ (1971) New computation in the tide-generating potential. *Geophys J R Astron Soc* 23:45–74
95. Foster J, Bevis M, Chen Y-L, Businger S, Zhang Y (2003) The Ka'u Storm (Nov 2000): Imaging precipitable water using GPS. *J Geophys Res* 108:4585
96. Segall P, Matthews M (1997) Time dependent inversion of geodetic data. *J Geophys Res* 102:22391–22409
97. Flores A, Ruffini G, Rius A (2000) 4D tropospheric tomography using GPS slant wet delays. *Ann Geophys* 18:223–234
98. Spiess FN (1985) Analysis of a possible sea floor strain measurement system. *Mar Geodesy* 9:385–398
99. Zumberge M (1997) Precise Optical Path Length Measurement Through an Optical Fiber: Application to Seafloor Strain Monitoring. *Ocean Eng* 24:532–542
100. Zumberge M et al (2006) In: Nadim F, Pottler R, Einstein H, Klapperich H, Kramer S (eds) 2006 ECI Conference on Geohazards. Lillehammer, Norway
101. Okada Y (1985) Surface deformation due to shear and tensile faults in a half-space. *Bull Seismol Soc Am* 75:1135–1154
102. Steketee JA (1958) Some Geophysical Applications of the Elasticity Theory of Dislocations. *Can J Phys* 36:1168–1198
103. Menke W (1984) *Geophysical Data Analysis: Discrete Inverse Theory*. Academic Press Inc, San Diego
104. Brooks BA, Frazer LN (2005) Importance reweighting reduces dependence on temperature in Gibbs samplers: an application to the coseismic geodetic inverse problem. *Geophys J Int* 161:12–21
105. Mosegaard K, Sambridge M (2002) Monte Carlo analysis of inverse problems. *Inverse Probl* 18:29–54
106. Kirkpatrick S, Gelatt CDJ, Vecchi MP (1983) Optimization by simulated annealing. *Science* 220:671–680
107. Hudnut KW et al (1996) Co-Seismic Displacements of the 1994 Northridge, California, Earthquake. *Bull Seismol Soc Am* 86:19–36
108. Dieterich J, Cayol V, Okubo PG (2000) The use of earthquake rate changes as a stress meter at Kilauea volcano. *Nature* 408:457–460
109. Stein RS (1999) The role of stress transfer in earthquake occurrence. *Nature* 402:605–609
110. Hansen S, Thurber CH, Mandernach MJ, Haslinger F, Doran C (2004) Seismic velocity and attenuation structure of the East Rift Zone and South Flank of Kilauea Volcano, Hawaii. *Bull Seismol Soc Am* 94:1430–1440
111. Got J-L, Okubo PG (2003) New insights into Kilauea's volcano dynamics brought by large-scale relative relocation of microearthquakes. *J Geophys Res* 108(B7):2337; doi:10.1029/2002JB002060
112. Wolfe CJ, Brooks BA, Foster JH, Okubo PG (2007) Microearthquake streaks and seismicity triggered by slow earthquakes on the mobile south flank of Kilauea Volcano, Hawai'i. *Geophys Res Lett* (In press)

113. McGuire JJ, Segall P (2003) Imaging of aseismic fault slip transients recorded by dense geodetic networks. *Geophys J Int* 155:778–788
114. Scholz CJ (1998) Earthquakes and friction laws. *Nature* 391:37–42
115. Perfettini H, Schmittbuhl J, Rice JR, Cocco M (2001) Frictional response induced by time-dependent fluctuations of the normal loading. *J Geophys Res* 106:435–438
116. Liu Y, Rice JR (2007) Spontaneous and triggered aseismic deformation transients in a subduction fault model. *J Geophys Res* 112:B09404; doi:10.1029/2007JB004930
117. Rice JR, Gu JC (1983) Earthquake aftereffects and triggered seismic phenomena. *Pure Appl Geophys* 121:187–219
118. Perfettini H, Schmittbuhl J (2001) Periodic loading on a creeping fault: Implications for tides. *Geophys Res Lett* 28:435–438
119. Shibasaki B, Shimamoto T (2007) Modelling of short-interval silent slip events in deeper subduction interfaces considering the frictional properties at the unstable-stable transition regime. *Geophys J Int* 171(1):191–205
120. Elsworth D, Day SJ (1999) Flank collapse triggered by intrusion: the Canarian and Cape Verde Archipelagoes. *J Volcan Geotherm Res* 94:323–340
121. Cervelli P et al (2002) The 12 September 1999 upper east rift zone dike intrusion at Kilauea Volcano, Hawaii. *J Geophys Res B: Solid Earth* 107:3-1–3-13
122. Owen S et al (2000) January 30, 1997 eruptive event on Kilauea Volcano, Hawaii, as monitored by continuous GPS. *Geophys Res Lett* 27:2757–2760
123. McMurtry GM, Watts P, Fryer GJ, Smith JR, Imamura F (2004) Giant landslides, mega-tsunamis, and paleo-sea level in the Hawaiian Islands. *Mar Geol* 203:219–233
124. Ward S (2002) Slip-sliding away. *Nature* 415:973–974
125. Frye D et al (2006) An Acoustically Linked Moored–Buoy Ocean Observatory. *EOS Trans AGU* 87:213–218

# System Dynamics Models of Environment, Energy and Climate Change

ANDREW FORD  
School of Earth and Environmental Sciences,  
Washington State University, Pullman, Washington,  
USA

## Article Outline

Glossary  
Definition of the Subject  
Introduction  
The Model of Mono Lake  
The Model of the Salmon in the Tucannon River  
Models of Climate Change  
System Dynamics Models of the Carbon Cycle  
Lessons from the Regulated Power Industry  
in the 1970s  
Simulating the Power Industry Response  
to a Carbon Market  
Conditions for Effective Interdisciplinary Modeling  
Future Directions  
Bibliography

## Glossary

**CO<sub>2</sub>** Carbon dioxide is the predominant greenhouse gas. Anthropogenic CO<sub>2</sub> emissions are created largely by the combustion of fossil fuels.

**CGCM** Coupled general circulation model, a climate model which combines the atmospheric and oceanic systems.

**GCM** General circulation model, a term commonly used to describe climate models maintained at large research centers.

**GHG** GHG is a greenhouse gas such as CO<sub>2</sub> and methane. These gases contribute to global warming by capturing some of the outgoing infrared radiation before it leaves the atmosphere.

**GT** Gigaton, a common measure of carbon storage in the global carbon cycle. A GT is a billion metric tons.

**IPCC** The Intergovernmental Panel on Climate Change was formed in 1988 by the World Meteorological Organization and the United Nations Environmental Program. It reports research on climate change. Their assessments are closely watched because of the requirement for unanimous approval by all participating delegates.

## Definition of the Subject

System dynamics is a methodology for studying and managing complex systems which change over time. The method uses computer modeling to focus our attention on the information feedback loops that give rise to the dynamic behavior. Computer simulation is particularly useful when it helps us understand the impact of time delays and nonlinearities in the system. A variety of modeling methods can aid the manager of complex systems. Coyle (p. 2 in [3]) puts the system dynamics approach in perspective when he describes it as that “branch of control theory which deals with socio-economic systems, and that branch of management science which deals with problems of controllability.” The emphasis on controllability can be traced to the early work of Jay Forrester [9] and his background in control engineering [10]. Coyle highlighted controllability again in the following, highly pragmatic definition:

*System dynamics is a method of analyzing problems in which time is an important factor, and which involve the study of how a system can be defended against, or made to benefit from, the shocks which fall upon it from the out-side world.*

The emphasis on controllability is important as it directs our attention to understanding and managing the system, not to the goal of forecasting the future state of the system. Making point predictions is the objective of some modeling methods, but system dynamics models are used to improve our understanding of the general patterns of dynamic behavior. System dynamics has been widely used in business, public policy and energy and environmental policy making. This article describes applications to energy and environmental systems.

## Introduction

System dynamics has been used extensively in the study of environmental and energy systems. This article describes some of these applications, paying particular attention to the problem of global climate change. The applications were selected to illustrate the power of the method in promoting an interdisciplinary understanding of complex problems.

The applications to environmental and energy systems are similar to applications to other systems described in this encyclopedia. They usually begin with the recognition of a dynamic pattern that represents a problem. System dynamics is based on the premise that we can improve our understanding of the dynamic behavior by the construction and testing of computer simulation models. The

models are especially helpful when they illuminate the key feedbacks that give rise to the problematic behavior.

System dynamics is explained in the core article in this volume, in the early texts by Forrester [9], Coyle [3] and Richardson [18] and in more recent texts on strategy by Warren [22] and by Morecroft [17]. The most comprehensive explanation is provided in the text on business dynamics by Sterman [19]. Applications to environmental systems are explained in the text by Ford [7]. The most widely read application to the environment is undoubtedly *The Limits to Growth* [16]. Collections of environmental applications appear in special issues of the *System Dynamics Review* [11,20].

The models are normally implemented with visual software such as Stella (<http://www.iseesystems.com>), Vensim (<http://www.vensim.com/>) or Powersim (<http://www.powersim.com/>). These programs use stock and flow icons to help one see where the accumulations of the system take place. They also help one to see the information feedback in the simulated system. The programs use numerical methods to show the dynamic behavior of the simulated system. The examples selected for this article make use of the Stella and Vensim software.

This article begins with textbook examples of environmental resources in the western US. The management of water levels at Mono Lake in Northern California is the first example. It shows a hydrological model to simulate the decline in lake levels due to water exported out of the basin. The second example involves the declining salmon population in the Tucannon River in Eastern Washington. These examples demonstrate the clarity of the approach, and they illustrate the potential for interdisciplinary modeling.

The article then turns to the topic of climate change and global warming. The focus is on the global carbon cycle and the growing concentration of carbon dioxide (CO<sub>2</sub>) in the atmosphere. A wide variety of models have been used to improve our understanding of the climate system and the importance of anthropogenic CO<sub>2</sub> emissions. Examples of system dynamics models are presented to show how they can improve our understanding and provide a platform for interdisciplinary analysis.

System dynamics has also been widely applied in the study of energy problems, especially problems in the electric power industry. The final section describes two applications to electric power. The first involved the financial problems of regulated electric utilities in the US during the 1970s. It demonstrates the usefulness of the method in promoting an interdisciplinary understanding of the utilities' financial problems. The second study dealt with the CO<sub>2</sub> emissions in the large electricity system in the West-

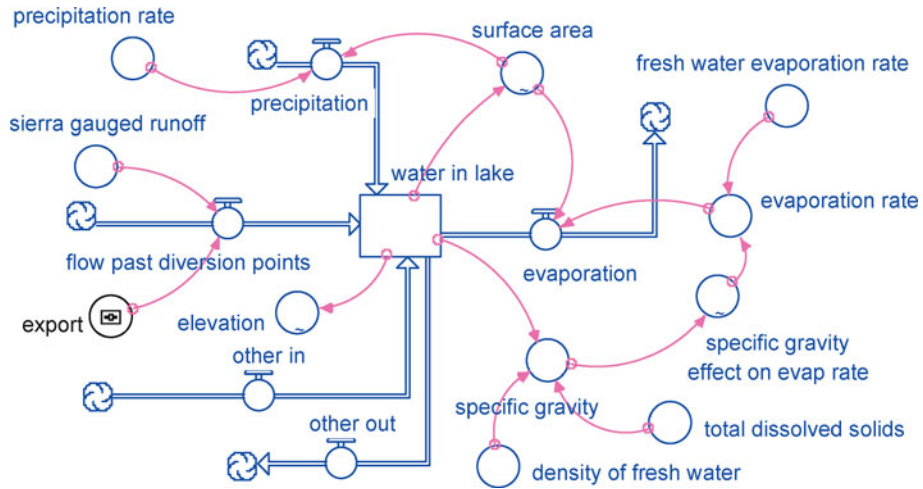
ern USA and Canada. It demonstrated how the power industry could lead the way in reducing CO<sub>2</sub> emissions in the decades following the implementation of a market in carbon allowances.

### The Model of Mono Lake

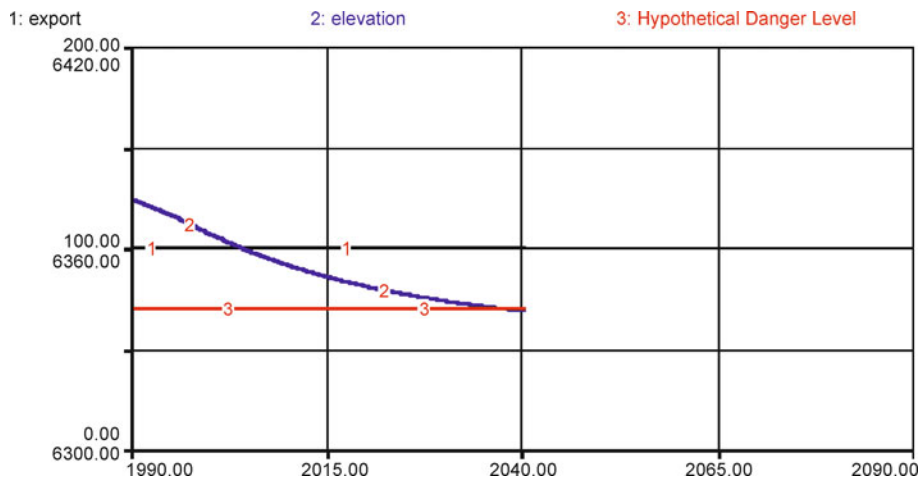
Mono Lake is an ancient inland sea on the east side of the Sierra Nevada Mountains in California. Microscopic algae thrive in its saline waters, and the algae support huge populations of brine flies and brine shrimp which can, under the right conditions, provide a virtually limitless food supply for migratory and nesting birds. Starting in 1941, stream flows toward Mono Lake were diverted into the aqueduct for export to Los Angeles. The large export deprived the lake of the historical flows, and the volume shrunk over the next four decades. By 1980, the lake's volume was cut approximately in half, and its salinity nearly doubled. Higher salinity levels posed risks to the ecosystem, and environmental scientists feared for the future of the lake ecosystem. Various groups filed suit in the 1970s to limit exports, and the California Supreme Court ruled in 1983 that public trust doctrine mandated a reconsideration of the management of the waters of the Mono Basin. That reconsideration led to a long-term plan to limit exports until the lake's elevation would return to safer levels.

Figure 1 shows a system dynamics model to simulate water flows and storage in the Mono Basin. The goal was to understand the pattern of decline over four decades and to study the responsiveness of the lake to a change in export policy. The model is implemented with the Stella software, and Fig. 1 shows how the model appears when using the software. A single stock variable is used to represent the storage in the basin. The main flow into the lake is the flow from gauged streams that bring runoff from the Sierra to the lake. The aqueduct system diverts a portion of this flow south to Los Angeles, and the flow allowed past the diversion points is the main flow into the lake. The main outflow is the evaporation. It depends on the surface area of the lake and the evaporation rate. The surface area depends in a nonlinear way on the volume of water in the lake. Figure 1 shows that this model follows the standard, system dynamics practice of using familiar names to convey the meaning of the variables in the model. (These particular names match the terms used by water managers and hydrological models of the basin.)

Figure 2 shows the simulated decline in the lake if exports were allowed to continue at high levels for 50 years. The lake would decline from 6374 to around 6342 feet above sea level, a value which is designated as a hypothetical danger level for this simulation. The long, gradual



System Dynamics Models of Environment, Energy and Climate Change, Figure 1  
 Stella diagram of the model of Mono Lake



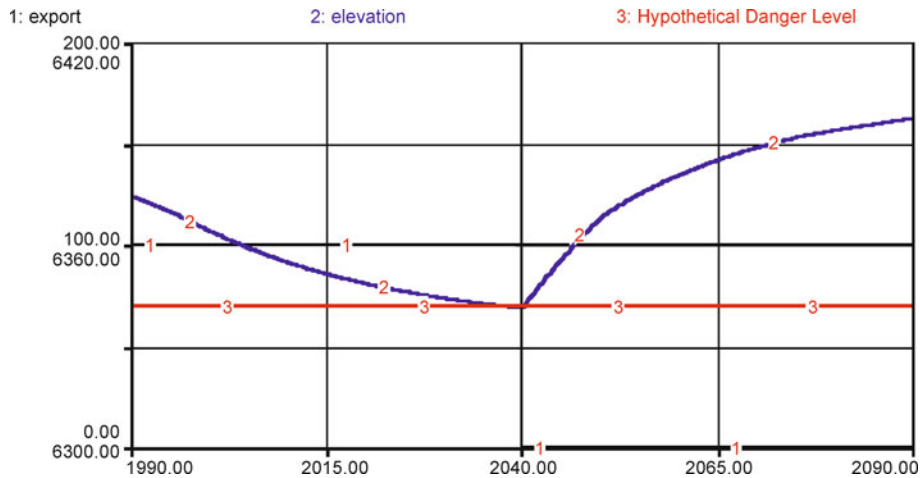
System Dynamics Models of Environment, Energy and Climate Change, Figure 2  
 Simulated decline in Mono Lake elevation if historical export were allowed to continue until the year 2040

decline is a match of projections by the other hydrological models used in the management plan for the basin. The lake will continue to fall until the area has been reduced sufficiently to create an evaporation which will lead to a balance of the flows in and out of the basin.

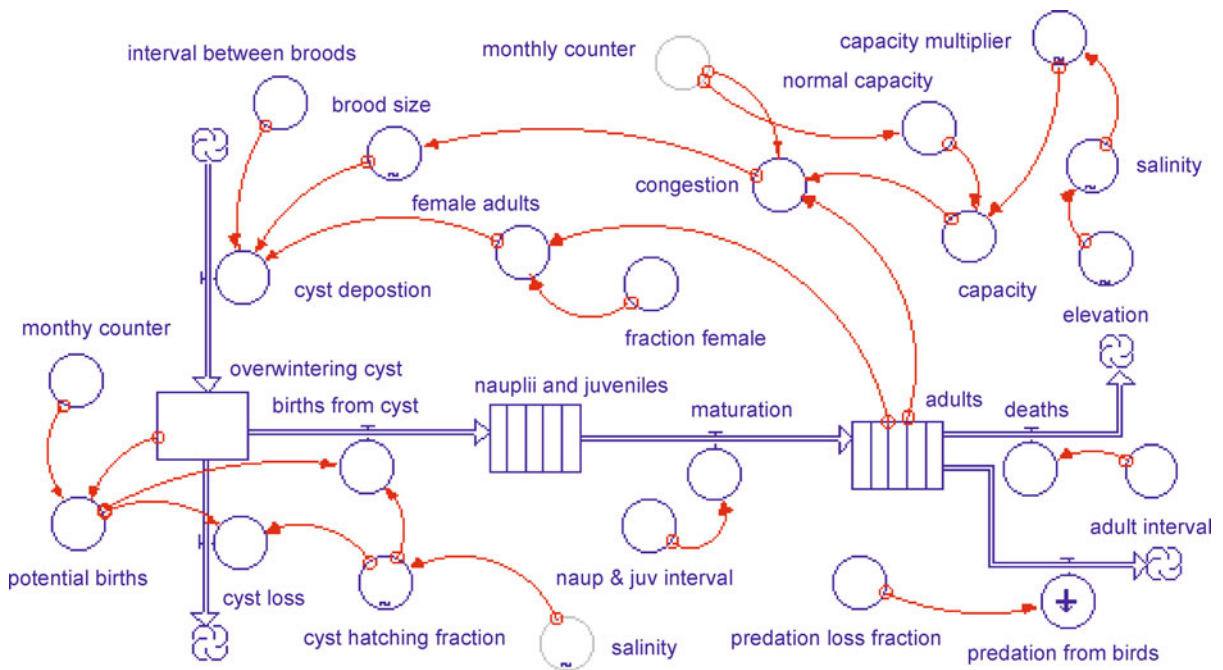
Figure 3 shows the simulated responsiveness of the lake to a change in export. The export is cut to zero midway through the simulation, and the elevation increases rapidly in the ensuing decade. The simulation reveals an immediate and rapid response, indicating that there is little downward momentum associated with the hydrology of the basin. This responsiveness is highly relevant to the management plan. When the lake falls to a dan-

gerous level, the export could be reduced, and the lake would climb to higher elevations within a few years after the change in policy. This rapid response supports the “wait and see” argument by those who advocated waiting for full signs of a dangerous salinity before changing export policy.<sup>1</sup> But there is far more than hydrology at work in this system. The waters of Mono Lake support a complex ecosystem which may or may not recover as quickly as the lake elevation. To explore the larger system requires

<sup>1</sup>“Wait and see” may be supported by an analysis of the hydrology of the basin, but it does not necessarily make sense when considering the long delays in the political and managerial process to change water export.



System Dynamics Models of Environment, Energy and Climate Change, Figure 3  
 Simulated recovery of Mono Lake elevation if export is set to zero for the second half of the simulation



System Dynamics Models of Environment, Energy and Climate Change, Figure 4  
 Stella model of the brine shrimp population of Mono Lake

an interdisciplinary model, one that looks at both hydrology and population biology.

Figure 4 shows a model of the population of brine shrimp that live in Mono Lake. The life cycle begins when the adult females deposit cysts in the summer. A stock is assigned to the overwintering cysts. The nauplii and juvenile phases are combined into a second stock, and the maturation leads to a new population of adults in the fol-

lowing summer. The model operates in months and is simulated over a long time interval to show the population response to long-term changes in elevation and in salinity. The model shows the population's response to changes in lake elevation, so one can learn about the delays in the population's response to the changes in lake elevation. Since the shrimp life cycle is 12 months, one would expect the population to rebound rapidly after the increase in ele-

vation and the reduction in salinity. The model confirms that the shrimp population would increase rapidly in the years following the elimination of water export from the basin.

The Mono Lake models are textbook models [7]. They demonstrate the clarity that the system dynamics approach brings to the modeling of environmental systems. The stock and flow icons help one see the structure of the system, and the long variable names help one appreciate the individual relationships. The simulation results help one understand the downward momentum in the system. In this particular case, there is no significant downward momentum associated with either the hydrological dynamics or the population dynamics.

The model in Fig. 1 allows for a system dynamics portrayal of the type of calculations commonly performed by hydrologists. Compared to the previous methods in hydrology, system dynamics adds clarity and ease of experimentation. The population model in Fig. 4 is a system dynamics version of the type of modeling commonly performed by population biologists. System dynamics adds clarity and ease of experimentation in this discipline as well.

The main theme of this article is that system dynamics offers the opportunity for interdisciplinary modeling and exploration. The Mono Lake case illustrates this opportunity with the combination of the hydrological and biological models that allows one to simulate management policies that control export based on the size of the brine shrimp population. The new model is no longer strictly hydrology nor strictly population biology; it is an interdisciplinary combination of both. And by using stock and flow symbols that are easily recognized by experts from many fields of study, the system dynamics enables quick transfer of knowledge. The ability to combine perspectives from different disciplines is one of the most useful aspects of the system dynamics approach to environmental and energy systems. This point is illustrated further with each of the remaining examples in the article.

### The Model of the Salmon in the Tucannon River

The next example involves the decline in salmon populations in the Snake and Columbia River system of the Pacific Northwest. By the end of the 1990s, the salmon had disappeared from 40% of their historical breeding ranges despite a public and private investment of more than \$1 billion. The annual salmon and steelhead runs had dwindled to less than a quarter of the runs from one hundred years ago. Figure 5 shows a system dynamics model one of the salmon runs, the population of Spring Chinook

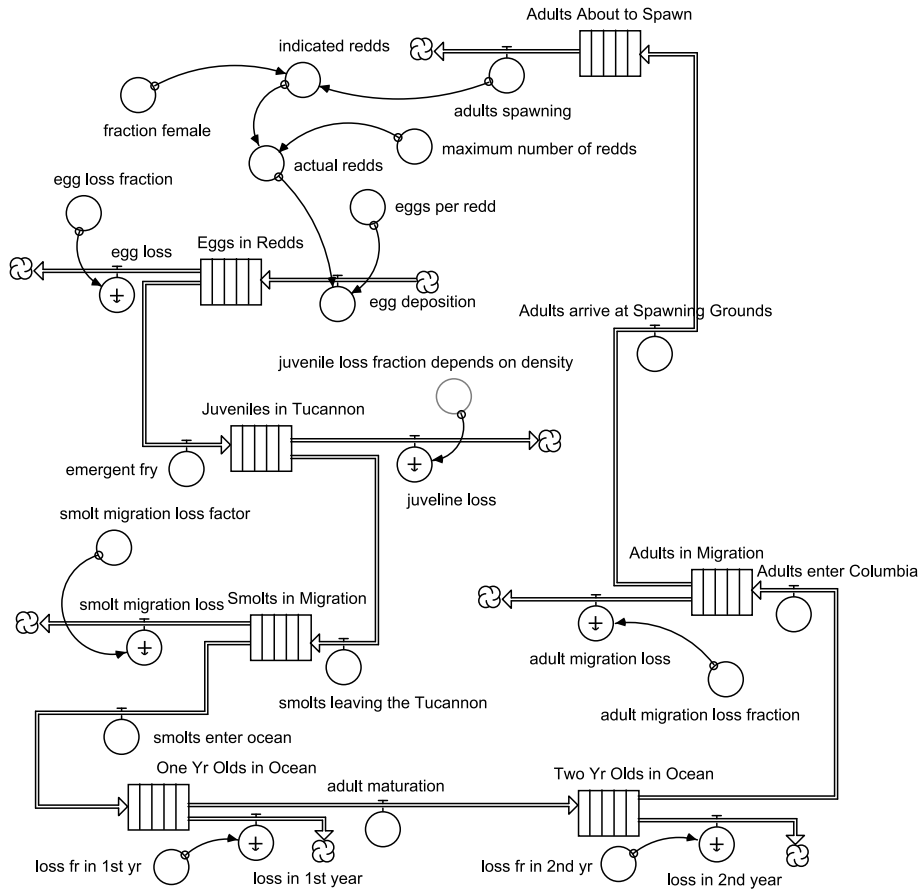
System Dynamics Models of Environment, Energy and Climate Change, Table 1  
Inputs to simulate the salmon population under pre-development conditions

Months in each phase		Population parameters	
Adults ready to spawn	1	fraction female	50%
eggs in redds	6	eggs per redd	3,900
juveniles in Tucannon	12	egg loss fraction	50%
smolts in migration	1	smolt migration loss factor	90%
one yr olds in ocean	12	loss fr for first yr	35%
two yr olds in ocean	12	loss fr for second yr	10%
adults in migration	4	adult migration loss fraction	25%

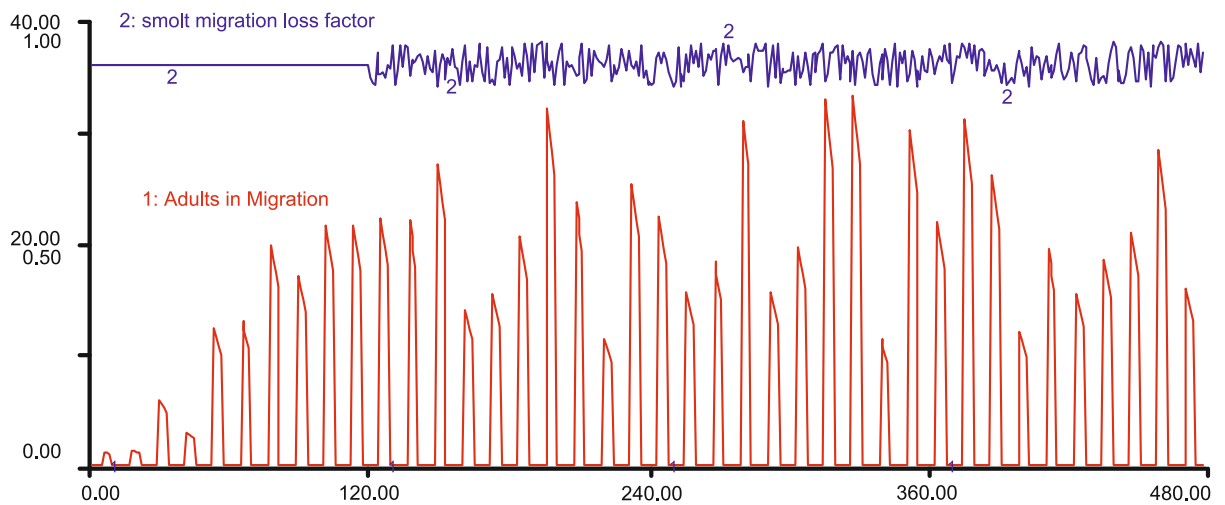
that spawn in the Tucannon River. The river rises in the Blue Mountains of Oregon and flows 50 miles toward the Snake River in Eastern Washington. It is estimated that the river originally supported runs of 20 thousand adults. But the number of returning adults has declined substantially due to many changes in the past sixty years. These changes include agricultural development in the Tucannon watershed, hydro-electric development on the Snake and Columbia, and harvesting in the ocean.

Each of the stocks in Fig. 5 correspond to a different phase in the salmon life cycle (see Table 1), with a total life-cycle of 48 months. The parameters represent predevelopment conditions, the conditions prior to agricultural development in the Tucannon watershed and hydro-electric development on the Snake and Columbia. Each of these parameters is fixed regardless of the size of the salmon populations. One of the most important variables is the "juvenile loss fraction depends on density." It can be as low as 50% when there are only a few emergent fry each spring. With higher densities, however, juvenile survival becomes more difficult due to crowding in the cool and safe portions of the river.

Figure 6 shows the model results over a 480 month period with the population parameters in Table 1. The simulation begins with a small number to see if the population will grow to the 20 thousand adults that were thought to have returned to the river in earlier times. The time graph shows a rapid rise to around 20 thousand adults within the first 120 months of the simulation. The remainder of the simulation tests the population response to variability in environmental conditions, as represented by random variations in the smolt migration loss fraction. (This loss tends to be high in years with low runoff and low in years with high runoff.) Figure 6 confirms that the model simulates the major swings in returning adults due to environmental variability. The runs can vary from a low of ten thousand to a high of thirty thousand.

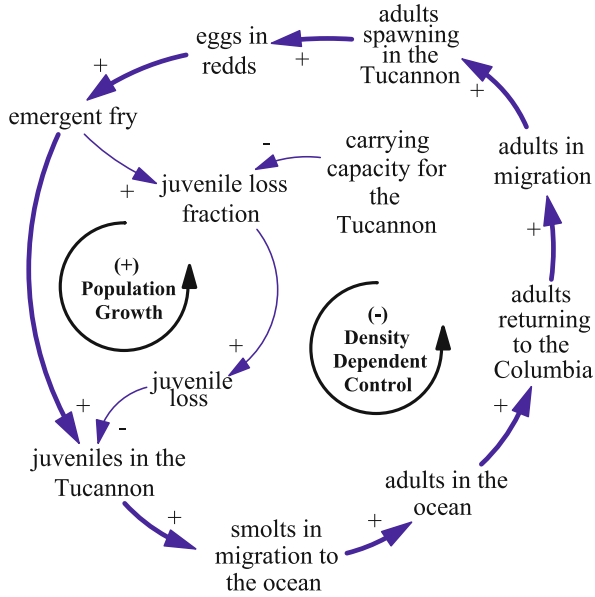


System Dynamics Models of Environment, Energy and Climate Change, Figure 5  
 Stella diagram of the model of the salmon life cycle



System Dynamics Models of Environment, Energy and Climate Change, Figure 6  
 Test of the salmon model with random variations in the smolt migration losses





System Dynamics Models of Environment, Energy and Climate Change, Figure 7  
Key feedback loops in the salmon model

System dynamics models are especially useful when they help us to understand the key feedbacks in the system. Positive feedback loops are essential to our understanding of rapid, exponential growth; negative feedbacks are essential to our understanding of the controllability of the system. Causal loop diagrams are often used to depict the feedback loops at work in the simulated system. Figure 7 shows an example by emphasizing the most important feedback loops in the salmon model.

Most readers will immediately recognize the importance of the outer loop which is highlighted by bold arrows in the diagram. Starting near the top, imagine that there are more spawning adults and more eggs in redds. We would then expect to see more emergent fry, more juveniles, more smolts in migration, more salmon in the ocean, more adults entering the Columbia, and a subsequent increase in the number of spawning adults. This is the positive feedback loop that gives the salmon population the opportunity to grow rapidly under favorable conditions.

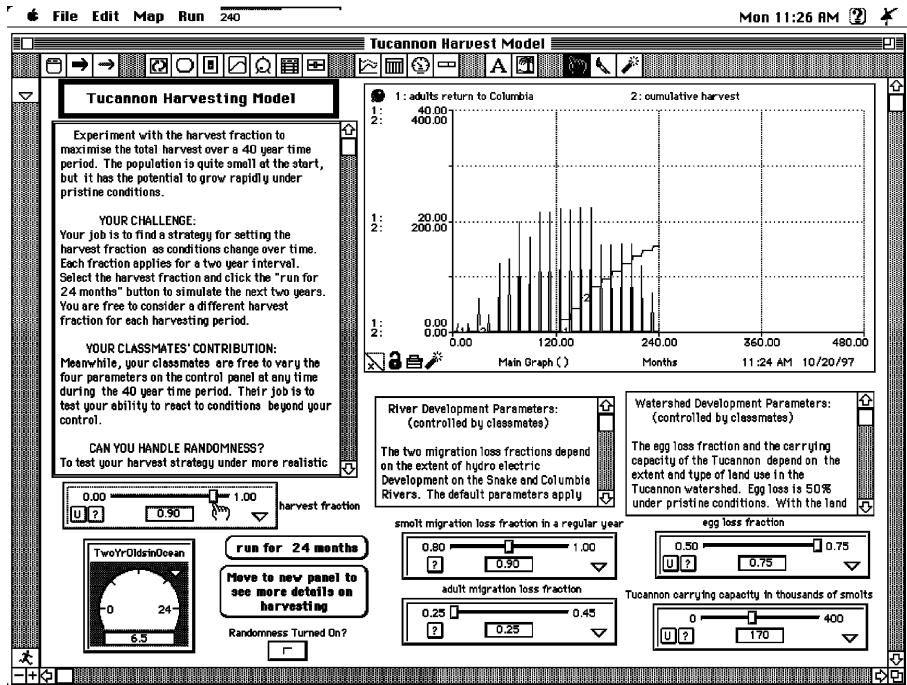
An equally important feedback works its way around the inner loop in the diagram. If we begin at the top with more spawners, we would expect to see more eggs, more fry and a greater juvenile loss fraction as the fry compete for space in the river. With a higher loss fraction, we expect to see fewer juveniles survive to be smolts, fewer smolts in migration, and fewer adults in the ocean. This means

we would see fewer returning adults and less egg deposition. This “density dependent feedback” becomes increasingly strong with larger populations, and it turns out to be crucial to the eventual size of the population. Simulating density dependent feedback is also essential to our understanding of the recovery potential of the salmon population. Suppose, for example, that the salmon experience high losses during the adult migration. This will mean that fewer adults reach the spawning grounds. There will be less egg deposition and fewer emergent fry in the following spring. The new cohort of juveniles will then experience more favorable conditions, and a larger fraction will survive the juvenile stage and migrate to the ocean. The density dependent feedback is crucial to the population’s ability to withstand shocks from external conditions.<sup>2</sup>

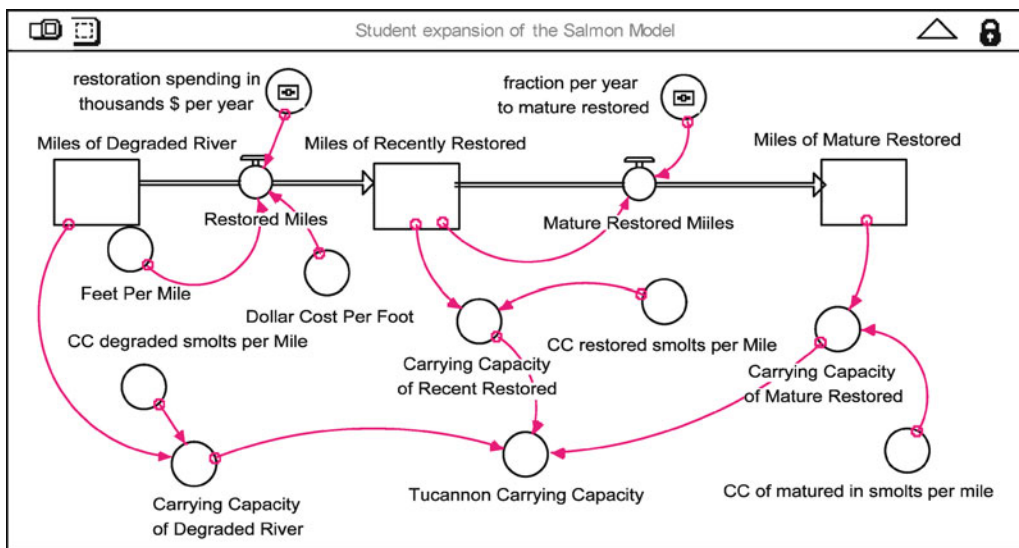
Figure 8 shows a version of the model to encourage student experimentation with harvesting policies. The information fields instruct the students to work in groups of three with one student playing the role of “the harvest manager”. The harvest manager’s goal is to achieve a large, sustainable harvest through control of the harvest fraction. The other students are given control of the parameters that describe conditions on the Snake and Columbia and in the Tucannon watershed. These students are encouraged to make major and unpredictable changes to test the instincts of the harvest manager.

Models designed for highly interactive simulations of this kind are sometimes called “management flight simulators” because they serve the same function as actual flight simulators. With a pilot simulator, the trainee takes the controls of an electro-mechanical model and tests his instincts for managing the simulated airplane under difficult conditions. The Tucannon harvesting model provides a similar opportunity for environmental students. They can learn the challenge of managing open access fisheries that are vulnerable to over harvesting and the tragedy of the commons [12]. In this particular exercise, students learn that they can achieve a sustainable harvest under a wide variety of difficult and unpredictable conditions. The key to sustainability is harvest manager’s freedom to change the harvest fraction in response to recent trends in number of returning adults. This is an important finding for fishery management because it reveals that the population dynamics are not the main obstacle to sustainability. Rather, unsustainable harvesting is more likely to occur

<sup>2</sup>The shocks could take the form of changes in ocean mortalities, changes in harvesting and changes in the migration mortalities. These shocks are external to the boundary of this model, so one is reminded of Coyle’s definition of system dynamics. That is, the model helps us understand how the salmon population could withstand the shocks which fall upon it from the out-side world.



System Dynamics Models of Environment, Energy and Climate Change, Figure 8  
 Salmon harvesting model to encourage student experimentation



System Dynamics Models of Environment, Energy and Climate Change, Figure 9  
 Student addition to simulate river restoration

when the managers find it difficult to change the harvest fraction in response to recent trends. This is the fundamental challenge of an open-access fishery.

The salmon model is a system dynamics version of the type of modeling commonly performed by popula-

tion biologists. System dynamics adds clarity and ease of experimentation compared to these models. It also provides a launching point for model expansions that can go beyond population biology. Figure 9 shows an example. This is a student expansion to change the carrying capac-

ity from a user input to a variable that responds to the user's river restoration strategy. The student was trained in geomorphology and was an expert on restoring degraded rivers in the west. The Tucannon began the simulation with 25 miles of river in degraded condition and the remaining 25 miles in a mature, fully restored river with a much higher carrying capacity. The new model permits one to experiment with the timing of river restoration spending and to learn the impact on the management of the salmon fishery.

The student's model provides another example of interdisciplinary modeling that aids our understanding of environmental systems. In this particular case, the modeling of river restoration is normally the domain of the geomorphologist. The model of the salmon population is the domain of the population biologist. Their work is often conducted separately, and their models are seldom connected. This is unfortunate as the experts working in their separate domains miss out on the insights that arise when two perspectives are combined within a single model. In the student's case, surprising insights emerged when the combined model was used to study the economic value of the harvesting that could be sustained in the decades following the restoration of the river. To the student's surprise, the new harvesting could "pay back" the entire cost of the river restoration in less than a decade.

### Models of Climate Change

Scientists use a variety of models to keep track of the greenhouse gasses and their impact on the climate. Some of the models combine simulations of the atmosphere, soils, biomass and ocean response to anthropogenic emissions. The more developed models include CO<sub>2</sub>, methane, nitrous oxides and other greenhouse gas (GHG) emissions and their changing concentrations in the atmosphere. Claussen [2] classifies climate models as simple, intermediate and comprehensive. The simple models are sometimes called "box models" since they represent the storage in the system by highly aggregated stocks. The parameters are usually selected to match the results from more complicated models. The simple models can be simulated faster on the computer, and the results are easier to interpret. This makes them valuable for sensitivity studies and in scenario analysis [13].

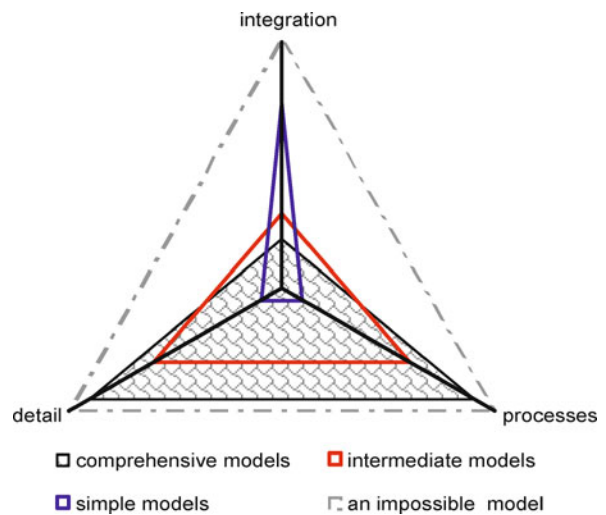
The comprehensive models are maintained by large research centers, such as the Hadley Center in the UK. The term "comprehensive" refers to the goal of capturing all the important processes and simulating them in a highly detailed manner. The models are sometimes called GCMs (general circulation models). They can be used to describe

circulation in the atmosphere or the ocean. Some simulate both the ocean and atmospheric circulation in a simultaneous, interacting fashion. They are said to be coupled general circulation models (CGCMs) and are considered to be the "most comprehensive" of the models available [2]. They are particularly useful when a high spatial resolution is required. However, a disadvantage of the CGCMs is that only a limited number of multi-decadal experiments can be performed even when using the most powerful computers.

Intermediate models help scientists bridge the gap between the simple and the comprehensive models. Claussen [2] describes eleven models of intermediate complexity. These models aim to "preserve the geographic integrity of the Earth system" while still providing the opportunity for multiple simulations to "explore the parameter space with some completeness. Thus, they are more suitable for assessing uncertainty". Figure 10 characterizes the different categories of models based on their relative emphasis on:

- number of processes (right axis)
- detailed treatment of the each process (left axis), and the
- extent of integration among the different processes (top axis).

Regardless of the methodology, climate modeling teams must make some judgments on where to concentrate their attention. No model can achieve maximum performance along all three dimensions. (Figure 10 uses the dashed lines



System Dynamics Models of Environment, Energy and Climate Change, Figure 10

Classification of climate models

to draw our attention to the impossible task of doing every thing within a single model.)

The comprehensive models strive to simulate as many processes as possible with a high degree of detail. This approach provides greater realism, but the models often fail to simulate the key feedback loops the link that atmospheric system with the terrestrial and oceanic systems. (An example is the feedback between CO<sub>2</sub> emissions, temperatures and the decomposition of soil carbon. If higher temperatures lead to accelerated decomposition, the soils could change from a net sink to a net source of carbon [15].) The simple models sacrifice detail and the number of processes in order to focus on the feedback effects between the processes. Using Claussen's terminology, one would say that such models aim for a high degree of "integration". However, the increased integration is achieved by limiting the number of processes and the degree of detail in representing each of the processes.

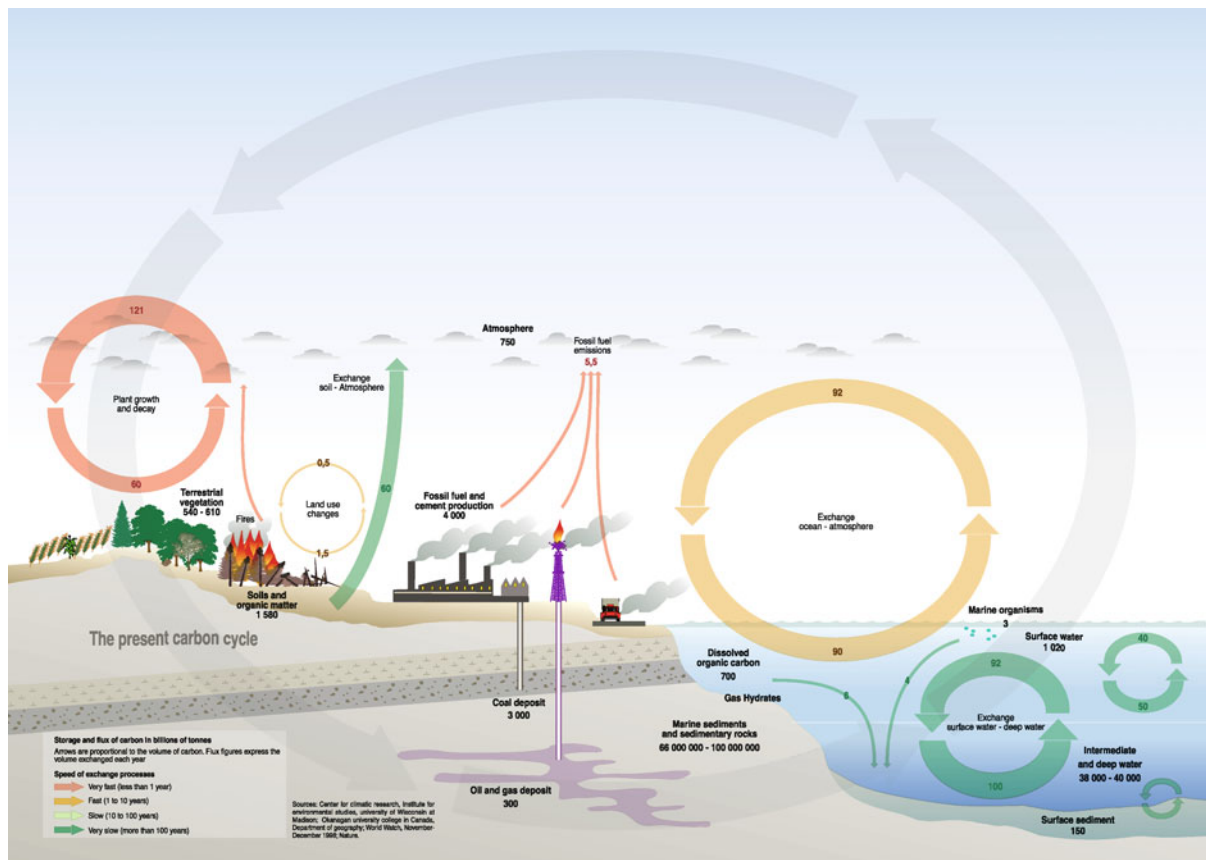
System dynamics has been used in a few applications to climate change. These applications fit in the category of

simple models whose goal is to provide a highly integrated representation of the system. Two examples are described here; both deal with the complexities of the global carbon cycle.

### System Dynamics Models of the Carbon Cycle

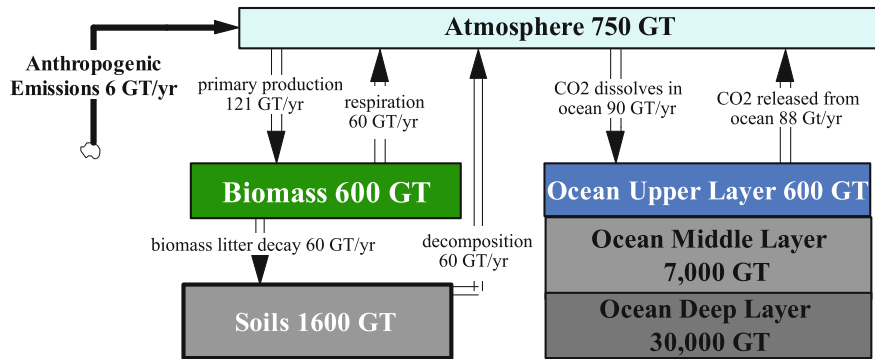
Figures 11 and 12 depict the global carbon cycle. Figure 11 shows the carbon flows in a visual manner. Figure 12 uses the Vensim stock and flow icons to summarize carbon storage and flux in the current system. The storage is measured in GT, gigatons of carbon, (where carbon is the C in CO<sub>2</sub>). The flows are in GT/year of carbon with values rounded off for clarity.

The left side of Fig. 12 shows the flows to the terrestrial system. The primary production removes 121 GT/yr from the atmosphere. This outflow exceeds the return flows by 1 GT/year. This imbalance suggests that around 1 GT of carbon is added to the stocks of biomass and soil each year. So the carbon stored in the terrestrial system would



System Dynamics Models of Environment, Energy and Climate Change, Figure 11

The global carbon cycle. (Source: United Nations Environmental Program (UNEP) <http://www.unep.org/>)

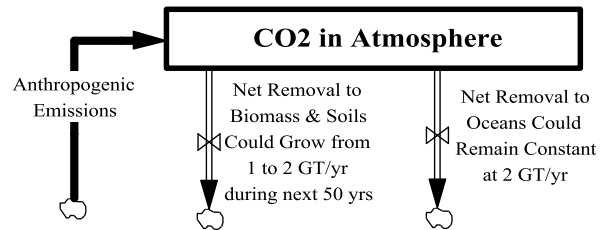


System Dynamics Models of Environment, Energy and Climate Change, Figure 12  
Diagram of the stocks and flows in the carbon cycle

grow over time (perhaps due to extensive reforestation of previously cleared land.) The right side of Fig. 12 shows the flows from the atmosphere to the ocean. The CO<sub>2</sub> dissolved in the ocean each year exceeds the annual release back to the atmosphere by 2 GT. The total, net-flow out of the atmosphere is 3 GT/year which means that natural processes are acting to negate approximately half of the current anthropogenic load.

As the use of fossil fuels grows over time, the anthropogenic load will increase. But scientists do not think that natural processes can continue to negate 50% of an ever increasing anthropogenic load. On the terrestrial side of the system, there are limits on the net flow associated with reforestation of previously cleared land. And there are limits to the carbon sequestration in plants and soils due to nitrogen constraints. On the ocean side of the system, the current absorption of 2 GT/year is already sufficiently high to disrupt the chemistry of the ocean's upper layer. Higher CO<sub>2</sub> can reduce the concentration of carbonate, the ocean's main buffering agent, thus affecting the ocean's ability to absorb CO<sub>2</sub> over long time periods.

Almost of the intermediate and comprehensive climate models may be used to estimate CO<sub>2</sub> accumulation in the atmosphere in the future. For this article, it is useful to draw on the mean estimate published in *Climatic Change* by Webster [23]. He used the climate model developed at the Massachusetts Institute of Technology, one of the eleven models of "intermediate complexity" in the review by Claussen [2]. The model began the simulation in the year 2000 with an atmospheric CO<sub>2</sub> concentration of 350 parts per million (ppm). (This concentration corresponds to around 750 GT of carbon in the atmosphere.) The mean projection assumed that anthropogenic emissions would grow to around 19 GT/year by 2100. The mean projection of atmospheric CO<sub>2</sub> was around 700 ppm



System Dynamics Models of Environment, Energy and Climate Change, Figure 13  
Simple model to understand accumulation of CO<sub>2</sub> in the atmosphere

by 2100. The amount of CO<sub>2</sub> in the atmosphere would be twice as high at the end of the century.

Figure 13 shows the simplest possible model to explain the doubling of atmospheric CO<sub>2</sub>. The stock accumulates the effect of three flows, each of which is specified by the user. Anthropogenic emissions are set to match Webster's assumption. They grow to 19 GT/year by the end of the century. Net removal to oceans is assumed to remain constant at 2 GT/year for the reasons given previously. Net removal to biomass and soils is then subject to experimentation to allow this simple model to match Webster's results. A close match is provided if the net removal increases from 1 to 2 GT/year during the first half of the century and then remains at 2 GT/year for the next fifty years. With these assumptions, the CO<sub>2</sub> in the atmosphere would double from 750 to 1500 GT during the century. This means that the atmospheric concentration would double from 350 to 700 ppm, the same result published by Webster [23].

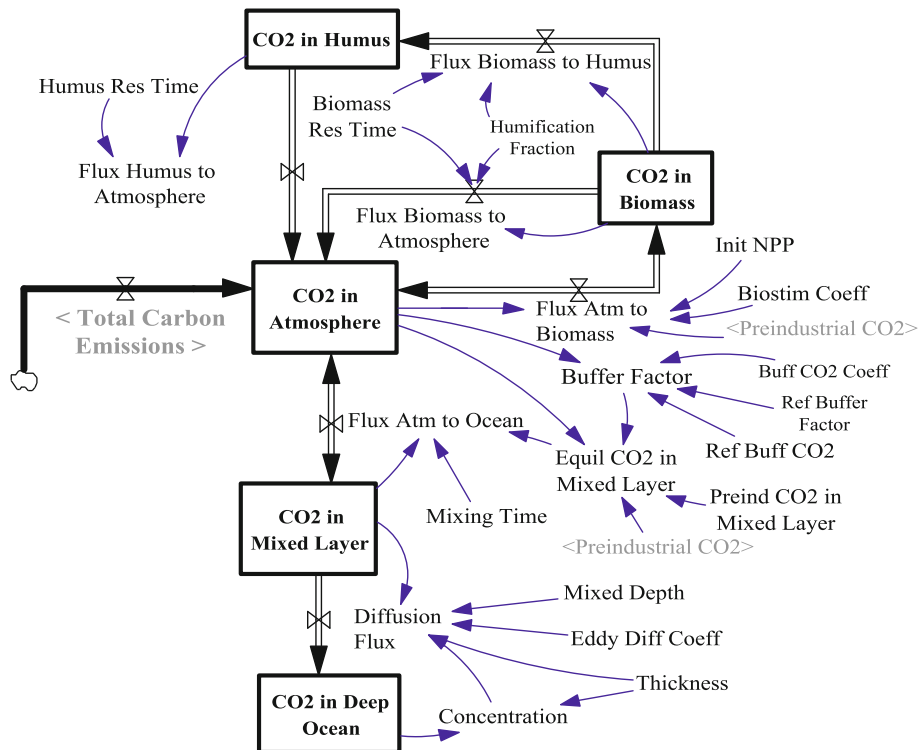
The model in Fig. 13 is no more than an accumulator. This is the simplest of possible models to add insight on the dynamics of CO<sub>2</sub> accumulation in the atmosphere. It includes a single stock and only three flows, with

all of the flows specified by the user. There are no feedback relationships which are normally at the core of system dynamics models. This extreme simplification is intended to make the point that simple models may provide perspective on the dynamics of a system. In this case, a simple accumulator can teach one about the sluggish response of atmospheric CO<sub>2</sub> in the wake of reductions in the anthropogenic emissions. As an example, suppose carbon policies were to succeed in cutting global emissions dramatically in the year 2050. By this year, emissions would have reached 10 GT/yr, so the supposed policy would reduce emissions to 5 GT/yr. What might then happen to CO<sub>2</sub> concentrations in the atmosphere for the remainder of the century? Experiments with highly educated adults [21] suggest that some subjects would answer this question with “pattern matching” reasoning. For example, if emissions are cut in half, it might make sense that CO<sub>2</sub> concentrations would be cut in half as well. But pattern matching leads one astray since the accumulation of CO<sub>2</sub> in the atmosphere responds to the total effect of the flows in Fig. 13. Were anthropogenic emissions to be reduced to 5 GT/year and net removals were to remain at 4 GT/year, the CO<sub>2</sub> concentration would continue to grow, and at-

mospheric CO<sub>2</sub> would reach 470 ppm by the end of the century.

The model in Fig. 13 is an extreme example to make a point about the usefulness of simple models. The next example is by Fiddaman [6]. It was selected as illustrative of the type of model that would emerge after a system dynamics study. Figure 14 shows the view of the carbon cycle, one of 30 views in the model. The model simulates the climate system within a larger system that includes growth in human population, growth in the economy, and changes in the production of energy. The model was organized conceptually as nine interacting sectors with a high degree of coupling between the energy, economic and the climate sectors.

Fiddaman focused on policy making, particularly the best way to put a price on carbon. In the current debate, this question comes down to a choice between a carbon tax and a carbon market. His simulations add support to those who argue that the carbon tax is the preferred method of putting a price on carbon. The simulations also provide another example of the usefulness of system dynamics models that cross disciplinary boundaries. By representing the economy, the energy system and the climate system



System Dynamics Models of Environment, Energy and Climate Change, Figure 14  
Representation of the carbon cycle in the model by Fiddaman [6]

within a single, tightly coupled model, he provides another example of the power of system dynamics to promote interdisciplinary exploration of complex problems.

System dynamics has also been applied to a wide variety of energy problems [1,7]. Indeed, a key word frequency count in 2004 revealed nearly 400 energy entries in the System dynamics bibliography [11]. Many of these applications deal with the electric power industry, and I have selected two electric studies to illustrate the usefulness of the approach. The first involves the regulatory and financial challenges of the investor owned electric utilities in the United States.

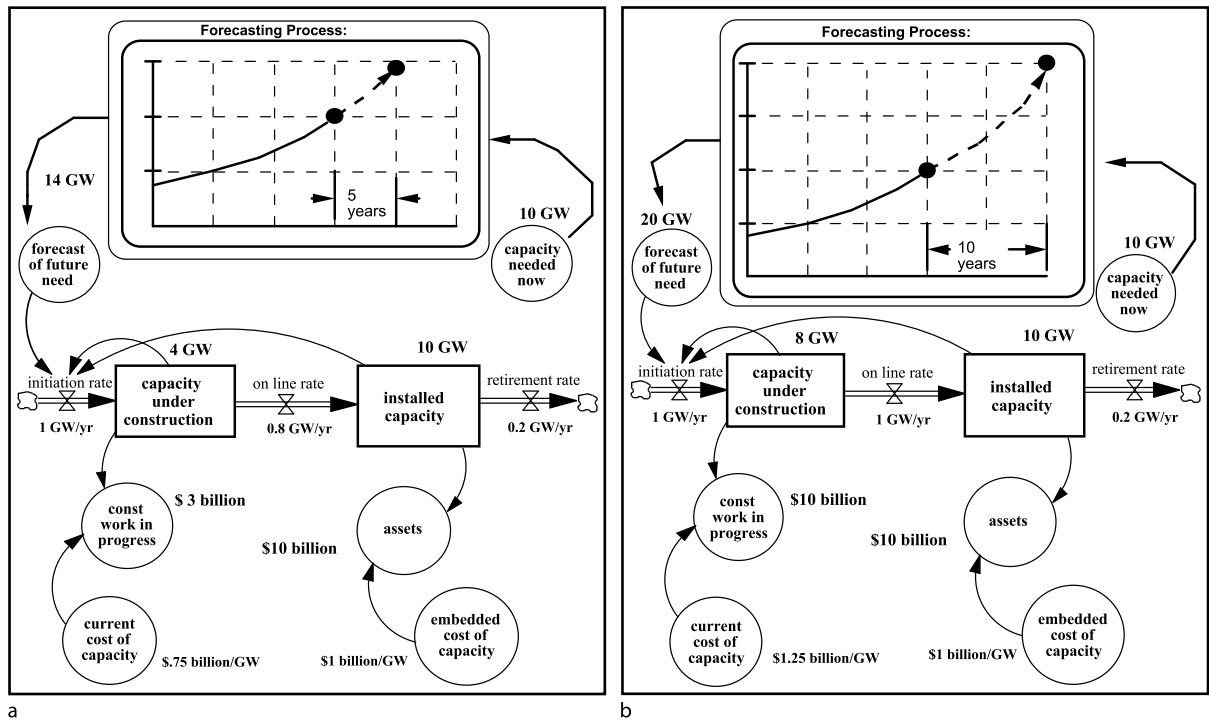
### Lessons from the Regulated Power Industry in the 1970s

The 1970s was a difficult decade for the regulated power companies in the United States. The price of oil and gas was increasing rapidly, and the power companies were frequently calling on their regulators to increase retail rates to cover the growing cost of fuel. The demand for electricity had been growing rapidly during previous decades, often at 7%/year. At this rate, the demand doubled every decade, and the power companies faced the challenge of

doubling the amount of generating capacity to ensure that demand would be satisfied. The power companies dealt with this challenge in previous decades by building ever larger power plants (whose unit construction costs declined due to economies of scale). But the economies of scale were exhausted by the 1970s, and the power companies found themselves with less internal funds and poor financial indicators. Utilities worried that the construction of new power plants would not keep pace with demand, and the newspapers warned of curtailments and blackouts.

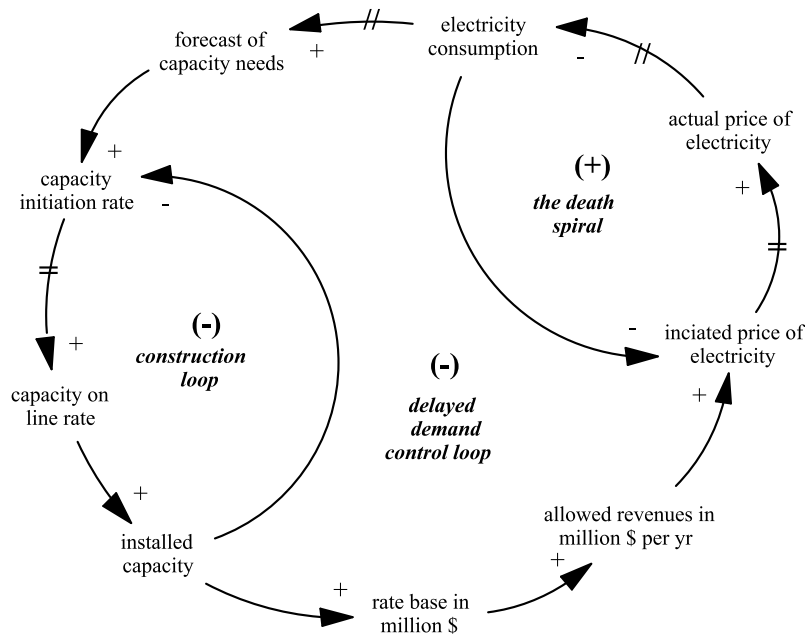
Figure 15 puts the financial problems in perspective by showing the forecasting, planning and construction processes. The side by side charts allows one to compare the difficult conditions of the 1970s with conditions in previous decades. Figure 15a shows the situation in the 1950s and 1960s. Construction lead times were around 5 years, so forecasts would extend 5 years into the future. Given the costs at the time, the power company would need to finance \$3 billion in construction. This was a substantial, but manageable task for a company with \$10 billion in assets.

Figure 15b shows the dramatic change in the 1970s. Construction lead times had grown to around 10 years, and construction costs had increased as well. The power



System Dynamics Models of Environment, Energy and Climate Change, Figure 15

a The electric utility's financial challenge during the 1950s and 1960s. b The electric utility's financial challenge during the 1970s



System Dynamics Models of Environment, Energy and Climate Change, Figure 16  
Key feedbacks and delays faced by power companies in the 1970s

company faced the challenge of financing \$10 billion in construction with an asset base of \$10 billion. The utility executives turned to the regulators for help. They asked for higher electricity rates in order to increase annual revenues and improve their ability to attract external financing. The regulators responded with substantial rate increases, but they began to wonder whether further rate increases would pose a problem with consumer demand. If consumers were to lower electricity consumption, the utility would have less sales and less revenues. The executives might then be forced to request another round of rate increases. Regulators wondered if they were setting loose a “death spiral” of ever increasing rates, declining sales and inadequate financing.

Figure 16 puts the problem in perspective by showing the consumer response to higher electricity rates along side of the other key feedback loops in the system. Higher electricity rates do pose the problem which came to be called “the death spiral”. But the death spiral does not act in isolation. Figure 16 reminds us that higher rates lead to lower consumption and to a subsequent reduction in the demand forecast and in construction. After delays for the new power plants to come on line, the power companies experiences a reduction in its “rate base” and the “allowed revenues”. When the causal relationships are traced around the outer loop, one sees a negative feedback loop that could act to stabilize the situation. The problem, how-

ever, is that the delays around the outer loop are substantially longer than the delay for the death spiral.

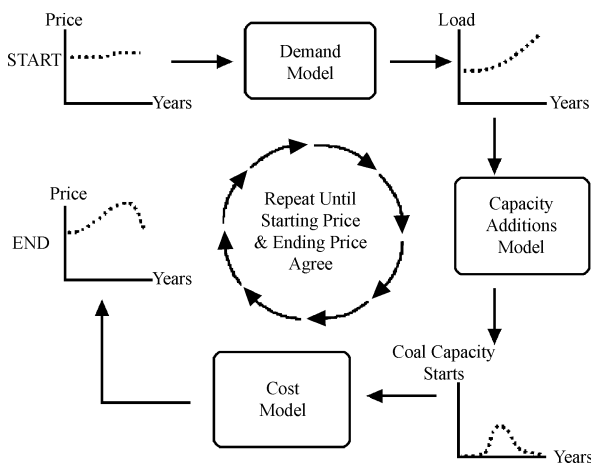
The utility companies financial challenge was the subject of several system dynamics studies in the 1970s and 1980s [7]. The studies revealed that the downward spiral could pose difficult problems, especially if consumers reacted quickly while utilities were stuck with long-lead time, capital intensive power plants under construction. The studies showed that utility executives needed to do more than rely on regulators to grant rate increases; they needed to take steps on their own to soften the impact of the death spiral. The best strategy was to shift the investments to technologies with shorter lead times. (As an example, a power company in coal region would do better to switch from large to smaller coal plants because of the small plants’ shorter lead time.) The studies also revealed that the company’s financial situation would improve markedly with slower growth in demand. By the late 1970s and early 1980s, many power companies began to provide direct financial incentives to their customers to slow the growth in demand. System dynamics studies showed that the company-sponsored efficiency programs would be beneficial to the both the customers (lower electric bills) and to the power companies (improved financial performance).

An essential feature of the utility modeling was the inclusion of power operations along side of consumer be-



havior, company forecasting, power plant construction, regulatory decision making and company financing. This interdisciplinary approach is common within the system dynamics community because practitioners believe that insights will emerge from simulating the key feedback loops. (This belief leads one to follow the cause and effect connections around the key loops regardless of the disciplinary boundaries that are crossed along the way.) This approach contrasts strongly with the customary modeling framework of large power companies who were not familiar with system dynamics. Their approach was to assign models to different departments (i. e., operations, accounting and forecasting) and string the models together to provide a view of the entire corporation over the long-term planning interval.

Figure 17 shows what can happen when models within separate departments are strung together. A large corporation might use 30 models, but this diagram makes the point by describing three models. The analysis would begin with an assumption on future electricity prices over the 20-year interval. These are needed to prepare a forecast of the growth in electricity load. The forecast is then given to the planning department which may run a variety of models to select the number power plants to construct in the future. The construction results are then handed to the accounting and rate making departments to prepare a forecast of electricity prices. When the company finally completes the many calculations, the prices that emerge may not agree with the prices that were assumed at the start. The company must then choose whether to ignore the contradiction or to repeat the entire process with a new es-



System Dynamics Models of Environment, Energy and Climate Change, Figure 17

The iterative approach often used by large power companies in the 1970s

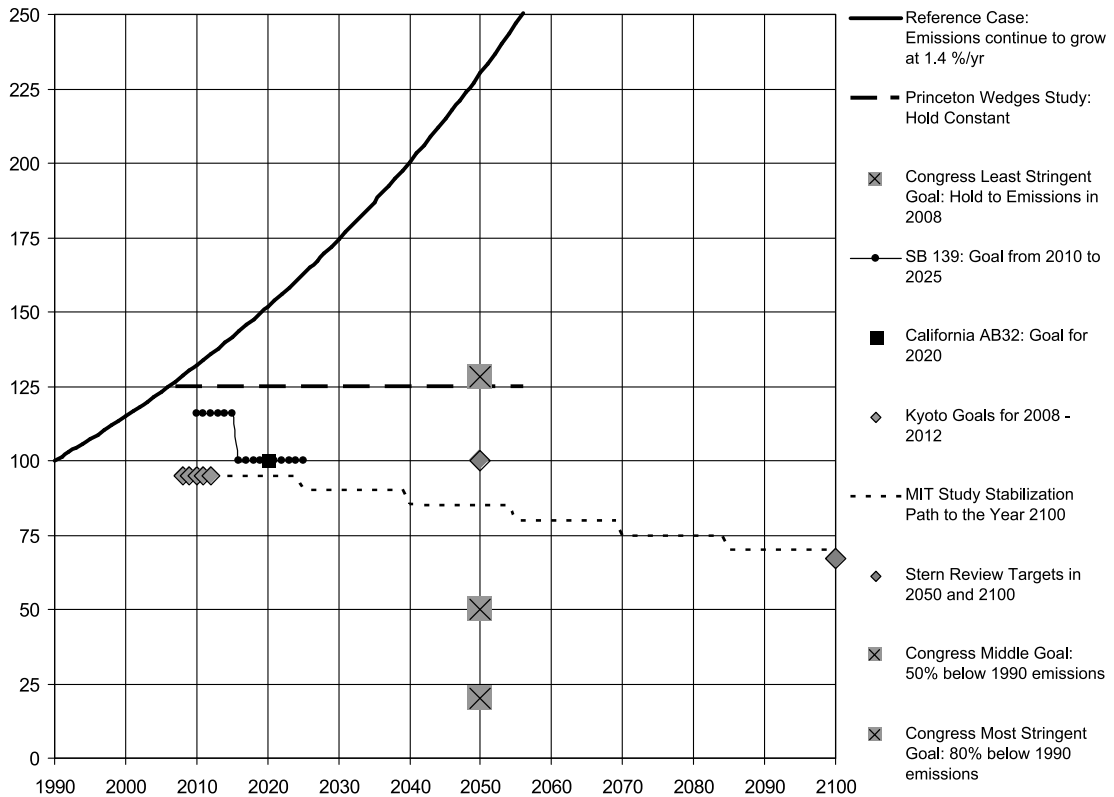
timate of the prices at the top of the diagram. This was not an easy choice. Ignoring the price discrepancy was problematic because it was equivalent to ignoring the “death spiral,” one of the foremost problems of the 1970s. Repeating the analysis was also problematic. The new round of calculations would be time consuming, and there was no guarantee that consistent results would be obtained at the end of the next iteration.

The power companies’ dilemma from the 1970s is described here to make an important point about the usefulness of system dynamics. System dynamics modeling is ideally suited for the analysis of dynamic problems that require a feedback perspective. The method allows one to “close the loop”, as long as one is willing to cross the necessarily disciplinary boundaries. In contrast, other modeling methods are likely to be extremely time consuming or fall short in simulating the key feedbacks that tie the system together.

### Simulating the Power Industry Response to a Carbon Market

The world is getting warmer, both in the atmosphere and in the oceans. The clearest and most emphatic description of global warming was issued by the intergovernmental panel on climate change (IPCC) in February of 2007. Their summary for policymakers (p. 4 in [14]) reported that the “Warming of the climate system is unequivocal, as is now evident from observations of increases in global average air and ocean temperatures, widespread melting of snow and ice and rising global mean sea level”. The IPCC concluded that “most of the observed increase is very likely due to the observed increase in anthropogenic greenhouse gas concentrations”. As a consequence of the IPCC and other warnings, policymakers around the world are calling for massive reductions in CO<sub>2</sub> and other greenhouse gas (GHG) emissions to reduce the risks of global warming.

Figure 18 summarizes some of the targets for emission reductions that have been adopted or proposed around the world. In many cases, the targets are specified relative to a country’s emissions in the year 1990. So, for ease of comparison, the chart uses 100 to denote emissions in the year 1990. Emissions have been growing at around 1.4%/year. The upward curve shows the future emissions if this trend continues: emissions would reach 200 by 2040 and 400 by 2090. The chart shows the great differences in the stringency of the targets. Some call for holding emissions constant; others call for dramatic reductions over time. Some targets apply to the next two decades; many extend to the year 2050; and some extend to the year 2100. However,



System Dynamics Models of Environment, Energy and Climate Change, Figure 18  
Comparison of goals for emissions (100 on the vertical axis represent emissions in the year 1990)

when compared to the upward trend, all targets require major reductions relative to business as usual.

The targets from the Kyoto treaty are probably the best known of the goals in Fig. 18. The treaty became effective in February of 2005 and called for the Annex I countries to reduce emissions, on average, by 5% below 1990 emissions by the year 2008 and to maintain this limit through 2012. The extension of the Kyoto protocol beyond 2012 is the subject of ongoing discussions. The solid line from 2010 to 2050 represents the “stabilization path” used in the climate modeling by Webster [23]. The limit on emissions was imposed in modeling calculations designed to stabilize atmospheric CO<sub>2</sub> at 550 ppmv or lower. The scenario assumed that the Kyoto emissions caps are adopted by all countries by 2010. The policy assumed that the caps would be extended and then further lowered by 5% every 15 years. By the end of the century, the emissions would be 35% below the value in 1990.

This article concentrates on Senate Bill 139, The Climate Stewardship Act of 2003. Figure 19 shows the S139 targets over the interval from 2010 to 2025. The bill called for an initial cap on emissions from 2010 to 2016. The



System Dynamics Models of Environment, Energy and Climate Change, Figure 19  
Map of the western electricity system

cap would be reduced to a more challenging level in 2016, when the goal was to limit emissions to no more than the emissions from 1990. S139 was introduced by Senators McCain and Lieberman in January of 2003. It did

not pass, but it was the subject of several studies including a highly detailed study by the Energy Information Administration [5]. The EIA used a wide variety of models to search for the carbon market prices that would induce industries to lower emissions to come into compliance with the cap. The carbon prices were estimated at \$22 per metric ton of CO<sub>2</sub> when the market was to open in 2010. They were projected to grow to \$60 by the year 2025.

The EIA study showed that the electric power sector would lead the way in reducing emissions. By the year 2025, power sector emissions would be reduced 75% below the reference case. This reduction was far beyond the reductions to be achieved by other sectors of the economy. This dramatic response was possible given the large use of coal in power generation and the power industry's wide range of choices for cleaner generation.

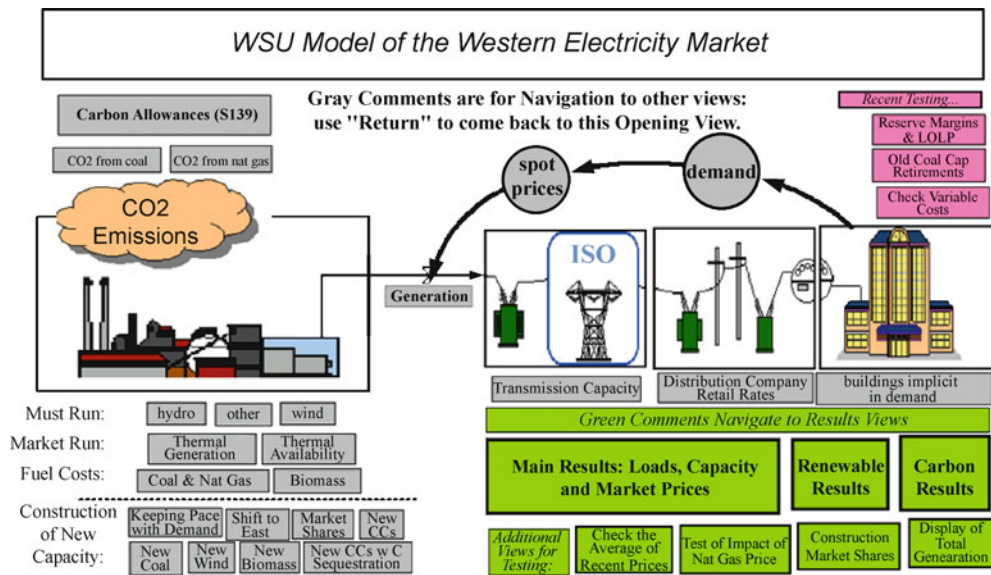
A system dynamics study of S139 was conducted at Washington State University (WSU) to learn if S139 could lead to similar reductions in the west. Electricity generation in the western system is provided in a large, interconnected power system shown in Fig. 19. This region has considerably more hydro resources, and it makes less use of coal-fired generation than the nation as a whole. The goal was to learn if dramatic reductions in CO<sub>2</sub> emissions could be possible in the west and to learn if they could be achieved with generating technologies that are commercially available today.

The opening view of the WSU model is shown in Fig. 20. The model deals with generation, transmission and

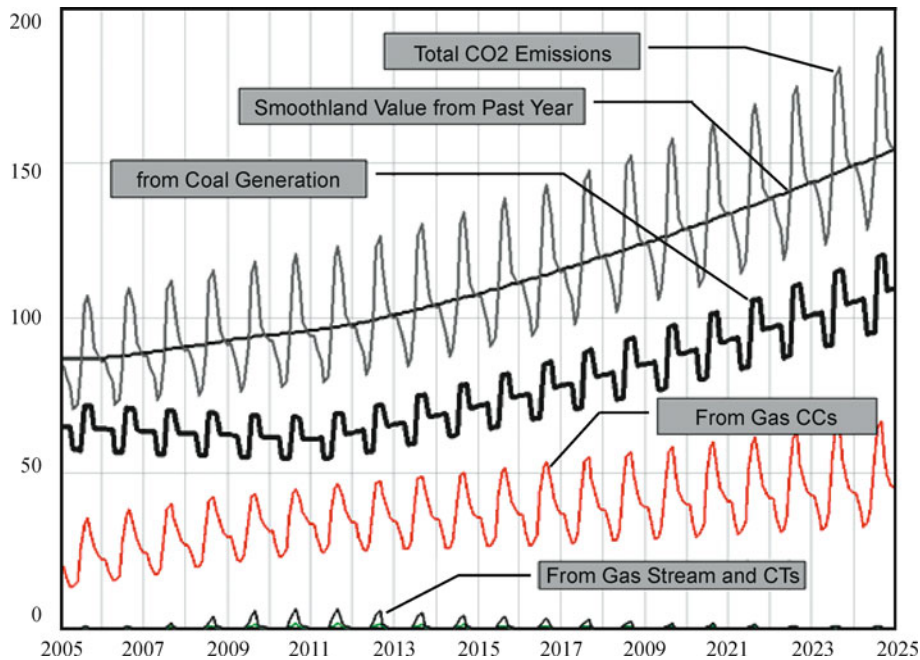
distribution to end use customers, with price feedback on the demand for electricity. The model is much larger than the textbook models described earlier in this article. Fifty views are required to show the all the diagrams and the simulation results. The opening view serves as a central hub to connect with all the other views.

The opening view uses Vensim's comment icons to draw attention to the CO<sub>2</sub> emissions in the model. The emissions arise mainly from coal-fired power plants, as shown in Fig. 21. A smaller, but still significant fraction of the emissions is caused by burning natural gas in combined cycle power plants. Total emissions vary with the seasons of the year, with the peak normally appearing in the summer when almost all of the fossil-fueled plants are needed to satisfy peak demand. The base case shows annual emissions growing by over 75% by the year 2025.

A major challenge for the system dynamics model is representing power flows across a transmission grid. Finding the flows on each transmission line and the prices in each area is difficult with the standard tools of system dynamics. It simply doesn't make sense to represent the power flows with a combination of stocks, flows and feedback processes to explain the flows. It makes more sense to calculate the flows and prices using traditional power systems methods, as explained by Dimitrovski [4]. The power flows were estimated using an algebraic approach which power engineers label as a reduced version of a direct-current optimal power flow calculation. The solution to the algebraic constraints were developed with the Matlab soft-



System Dynamics Models of Environment, Energy and Climate Change, Figure 20  
Opening view of the model of the western electricity system



System Dynamics Models of Environment, Energy and Climate Change, Figure 21  
Annual emissions in a base case simulation (annual emissions are in million metric tons of carbon)

ware and then transferred to user-defined functions to operate within the Vensim software. The Vensim simulations were set to run over twenty years with time in months. (A typical simulation required 240 months with changes during a typical day handled by carrying along separate calculations for each of 24 h in a typical day.) These are extensive calculations compared to many system dynamics models, so there was concern that we would lose the rapid simulation speed that helps to promote interactive exploration and model testing. The important methodological accomplishment of this project was the inclusion of network and hourly results within a long-term model without losing the rapid simulation response that encourages users to experiment with the model.

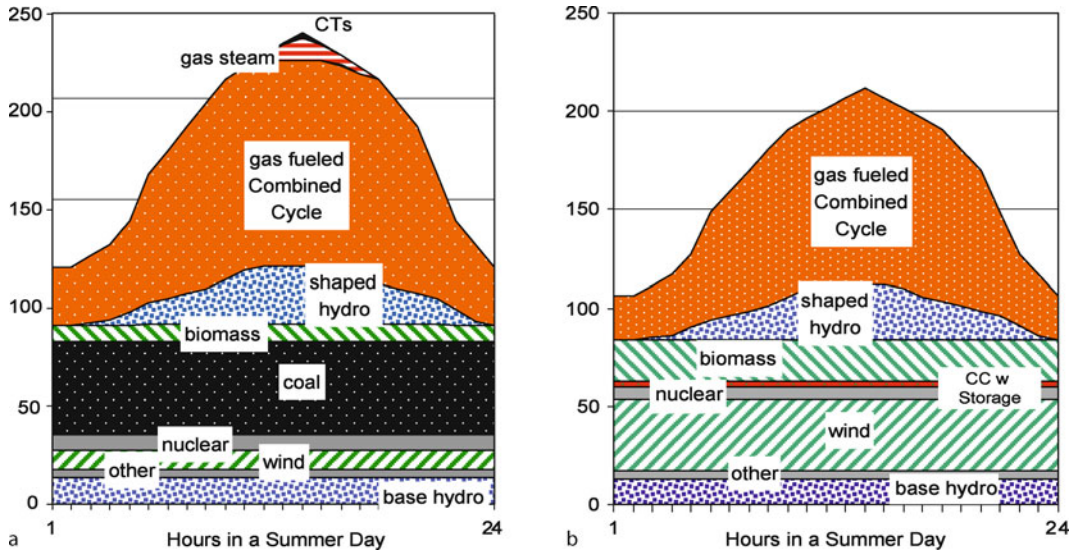
One of the model experiments called for a new simulation with carbon prices set to follow the \$20 to \$60 trajectory projected by the EIA for S139. These prices were specified as a user input, and the model responded with a change in both short-term operations and long-term investments. The important result was a 75% reduction in CO<sub>2</sub> emissions by the end of the simulation. This dramatic reduction corresponds almost exactly to the EIA estimate of CO<sub>2</sub> reduction for the power industry in the entire US.

Figure 22 helps one understand how CO<sub>2</sub> emissions could be reduced by such a large amount. These diagrams

show the operation of generating units across the Western US and Canada for a typical day in the summer of the final year of the simulation. Figure 22a shows the reference case; Figure 22b shows the case with S139. The side by side comparison helps one visualize the change in system operation. A comparison of the peak loads shows that the demand for electricity would be reduced. The reduction is 9%, which is due entirely to the consumers' reaction to higher retail electric prices over time.

Figure 22b shows large contributions from wind and biomass generation. Wind generation is carbon free, and biomass generation is judged to be carbon neutral, so these generating units make an important contribution by the end of the simulation. Both of these generating technologies are competitive with today's fuel prices and tax credits. The model includes combined cycle gas generation equipped with carbon capture and storage, a technology that is not commercially available today. The model assumes that advances in carbon sequestration over the next two decades would allow this technology to capture a small share of investment near the end of the simulation. By the year 2025, the combined cycle plants with sequestration equipment would provide 2% of the generation.

The most important observation from Fig. 22 is the complete elimination of coal-fired generation in the S139 case. Coal-fired units are shown to operate in a base load



System Dynamics Models of Environment, Energy and Climate Change, Figure 22

**a** Projected generation for a peak summer day in 2024 in the reference case. **b** Projected generation for a peak summer day in 2024 in the S139 case

mode in the reference simulation. They provide around 28% of the annual generation, but they account for around two-thirds of the CO<sub>2</sub> emissions in the western system. The carbon prices from S139 make investment in new coal-fired capacity unprofitable at the very start of the simulated market in 2010. As the carbon prices increase, utilities to cut back on coal-fired generation and compensate with increased generation from gas-fired CC capacity. In the simulations reported here, this fuel switching would push the coal units into the difficult position of operating fewer and fewer hours in a day. Eventually this short duration operation is no longer feasible, and coal generation is eliminated completely by the end of the simulation.

The WSU study of the western electric system was selected as the concluding example because of its novel treatment of network flows inside a system dynamics model [4]. The model is also interesting for its treatment of daily price changes within a long-term model. (Such changes are important to the simulation of revenues in the wholesale market.) From a policy perspective, the study confirms previous modeling of the pivotal role of the electric power industry in responding to carbon markets. The study indicated that the western electricity system could achieve dramatic reductions in CO<sub>2</sub> emissions within 15 years after the opening of a carbon market, and it could do so with technologies that are commercially available today [8].

### Conditions for Effective Interdisciplinary Modeling

All of the applications demonstrate the usefulness of system dynamics in promoting interdisciplinary modeling. The article concludes with comments on the level of effort and the conditions needed for effective, interdisciplinary modeling.

The examples in this article differ substantially in the level of effort required, from several weeks for the classroom examples to several years for the energy studies. The textbook examples involved student expansions of models of Mono Lake and the Tucannon salmon. The expansions were completed by undergraduate students in projects lasting two or three weeks. The key was the students' previous education (classes from many different departments) and their receptiveness to an interdisciplinary approach.

Fiddaman's model of the climate and energy system [6] was a more ambitious exercise, requiring several years of effort as part of his doctoral research. Bringing multi-year interdisciplinary modeling projects to a successful conclusion requires one to invest the time to master several disciplines and to maintain a belief that there are potential insights at the end of the effort.

The electric power industry examples were also ambitious projects that required several years of effort. The modeling of the western electricity system was a four-year project with support from the National Science Founda-

tion. The long research period was crucial for it allowed the researchers from power systems engineering, system dynamics and environmental science to take the time to learn from one another. The modeling of the electric company problems in the 1970s was also spread over several years of effort. The success of this modeling was aided by utility planners, managers and modelers who were looking for a systems view of their agency and its problems. They saw system dynamics as a way to tie existing ideas together within an integrated portrayal of their system. Their existing ideas were implemented in models maintained by separate functional areas (i. e., forecasting, accounting, operations). The existing models often provided a foundation for the system dynamics models (i. e., in the same way that the comprehensive climate models in Fig. 11 provide support for the development of the more integrated models). The key to effective, interdisciplinary modeling within such large organizations is support from a client with a strong interest in learning and with managerial responsibility for the larger system.

### Future Directions

This article concludes with future directions for system dynamics applications to climate change. People often talk of mitigation and adaptation. Mitigation refers to the challenge of lowering greenhouse gas emissions to avoid dangerous anthropogenic interference with the climate system. Adaptation refers to the challenge of living in a changing world.

**Mitigation:** The challenge of lowering CO<sub>2</sub> and other GHG emissions is the fundamental challenge of the coming century. The next two decades will probably see various forms of carbon markets, and system dynamics can aid in learning about market design. It is important that we learn how to make these markets work well. And if they don't work well, it's important to speed the transition to a carbon tax policy with better prospects for success. System dynamics can aid in learning about markets, especially if it is coupled with simulating gaming to allow market participants and regulators to "experience" and better understand market dynamics.

**Adaptation:** The world will continue to warm, and sea levels will continue to rise. These trends will dominate the first half of this century even with major reductions in CO<sub>2</sub> emissions. These and other climate changes will bring a wide variety of problems for management of water resource, public health planning, control of invasive species, preservation of endangered species, control of wildfire, and coastal zone management, just to name a few. Our understanding of the adaptation challenges can be improved

through system dynamics modeling. The prospects for insight are best if the models provide an interdisciplinary perspective on adapting to a changing world.

## Bibliography

### Primary Literature

1. Bunn D, Larsen E (1997) Systems modelling for energy policy. Wiley, Chichester
2. Claussen M et al (2002) Earth system models of intermediate complexity: closing the gap in the spectrum of climate system models. *Climate Dyn* 18:579–586
3. Coyle G (1977) Management system dynamics. Wiley, Chichester
4. Dimitrovski A, Ford A, Tomsovic K (2007) An interdisciplinary approach to long term modeling for power system expansion. *Int J Crit Infrastruct* 3(1–2):235–264
5. EIA (2003) United States Department of Energy, Energy Information Administration, Analysis of S139, the Climate Stewardship Act of 2003
6. Fiddaman T (2002) Exploring policy options with a behavioral climate-economy model. *Syst Dyn Rev* 18(2):243–264
7. Ford A (1999) Modeling the environment. Island Press, Washington
8. Ford A (2008) Simulation scenarios for rapid reduction in carbon dioxide emissions in the western electricity system. *Energy Policy* 36:443–455
9. Forrester J (1961) Industrial dynamics. Pegasus Communications
10. Forrester J (2000) From the ranch to system dynamics: An autobiography, in management laureates. JAI Press
11. Ford A, Cavana R (eds) (2004) Special Issue of the *Syst Dyn Rev*
12. Hardin G (1968) The tragedy of the commons. *Science* 162:1243–1248
13. IPCC (1997) An introduction to simple climate models used in the IPCC second assessment report. ISBN 92-9169-101-1
14. IPCC (2007) Climate change 2007: The physical science basis, summary for policymakers. [www.ipcc.ch/](http://www.ipcc.ch/)
15. Kump L (2002) Reducing uncertainty about carbon dioxide as a climate driver. *Nature* 419:188–190
16. Meadows DH, Meadows DL, Randers J, Behrens W (1972) The limits to growth. Universe Books
17. Morecroft J (2007) Strategic modelling and business dynamics. Wiley, Chichester
18. Richardson J, Pugh A (1981) Introduction to system dynamics modeling with dynamo. Pegasus Communications
19. Sterman J (2000) Business dynamics. McGraw-Hill, Irwin
20. Sterman J (ed) (2002) Special Issue of the *Syst Dyn Rev*
21. Sterman J, Sweeney L (2007) Understanding public complacency about climate change. *Clim Chang* 80(3–4):213–238
22. Warren K (2002) Competitive strategy dynamics. Wiley, Chichester
23. Webster M et al (2003) Uncertainty analysis of climate change and policy response. *Climat Chang* 61:295–320

### Books and Review

- Houghton J (2004) Global warming: The complete briefing, 3rd edn. Cambridge University Press, Cambridge

## Tomography, Seismic

JOSE PUJOL

Dept. of Earth Sciences, The University of Memphis,  
Memphis, USA

### Article Outline

Definition of the Subject

Introduction

Fundamentals of X-ray Computerized Tomography

Arrival-Time Seismic Tomography

Solution of Ill-Posed Linear Problems

Examples

Future Directions

Bibliography

### Definition of the Subject

Seismic tomography refers to a number of techniques designed to investigate the interior of the earth using arrival times and/or waveforms from natural and artificial sources. The most common product of a tomographic study is a velocity model, although other parameters, such as attenuation, are also studied. The importance of seismic tomography stems from two facts. One, it generally has higher resolution than that provided by other geophysical methods. Two, it provides information that (a) can help solve fundamental problems concerning the internal structure of the earth at a global scale, and (b) has been used in tectonic and seismic hazards studies at a local scale. Seismic tomography has also been applied to data collected in boreholes, but because of the high expenses associated with drilling, borehole tomography is relatively little used.

### Introduction

In the most general terms, seismic tomography problems are inverse problems, and before the word “tomography” entered the seismological literature the term inversion was used. This change occurred as a consequence of the revolution in medical imaging caused by computerized (or computed) X-ray tomography (CT) (also known as computer assisted or axial, tomography, CAT), introduced in the 1970s. The importance of CT on seismic tomography was that it provided efficient numerical techniques for the solution of systems of equations with extremely large number of unknowns ( $10^5$  or more), which allowed a great expansion of the seismological inverse problems that could be tackled. In addition, according to [1], chang-

ing the name inversion to tomography enhanced the “believability” of the inverse method.

Seismic tomography covers a wide range of scales, from the very small (e.g., borehole tomography, involving distances of a few kilometers at most) to the very large (i.e., whole earth tomography). Initially, seismic tomography involved body-wave arrival times, but later it was extended to include waveform and surface wave information. Arrival-time tomography is simpler and for this reason it is the most popular. Waveform tomography requires software for the computation of synthetic seismograms for comparison with the observed ones, which makes the whole inversion process more difficult both theoretically and practically. Surface wave tomography involves the determination of phase velocities, which is more complicated than picking arrival times. The three approaches, however, have one thing in common, namely, they all end up requiring the solution of a linear system of equations. This task might sound simple, but in practice it is not because the system generally does not have a unique solution, which means that solving it requires making decisions (either implicitly or explicitly) regarding the nature of the solution. In addition, given the complexity of wave propagation in heterogeneous media (such as the earth), the inverse problems solved by seismologists involve highly simplified models of the earth. These simplifications enter into the linear system to be solved, which adds another layer of uncertainty to the solution. This is in sharp contrast with X-ray tomography, which is comparatively unaffected by this kind of uncertainties. For this reason it might be argued that the name tomography does not do justice to the kind of problems that seismologists solve, and this fact should be born in mind by readers from other disciplines. Other differences between X-ray and seismic tomography, of more practical nature, are given below.

The goal of most of the seismic tomography work is to derive 3-D velocity models of portions of the earth. Currently, most of the research concentrates on two scales, global and local. At the global scale, the tomographic models generally have higher resolution than that provided by other geophysical methods, and for this reason it has the potential to provide constraints on the fate of the subducted slabs, on models of mantle convection, on petrological and geochemical models, on studies of the geomagnetic and gravity fields, on mineral physics, and on the core-mantle boundary, among others (see, e.g., [41,96,129,161]). Local tomography usually involves the simultaneous determination of a 3-D velocity model and the relocation of the seismic events used to determine the model. Traditionally, local velocity models have been used in structural and tectonic interpretations, but more recently

they have become important in seismic hazard studies, as the ground motion caused by earthquakes are amplified by low-velocity materials, which increases the hazard in the case of large events (see, e. g., [160]). As discussed below, the standard practice of locating earthquakes with layered velocity models may lead to significant location errors when the lateral velocity variations are significant. Therefore, the simultaneous velocity-inversion and event relocation has the potential to produce improved event locations, which is also important in the context of seismic hazards studies.

## Fundamentals of X-ray Computerized Tomography

### Historical Overview

Computerized tomography began in the early 1970s with the introduction of the X-ray scanner developed by G. Hounsfield [65,66], who was able to combine data generated using X-ray techniques with computerized data acquisition and processing techniques. This resulted in the first commercially viable CT machine, with clinical applications presented in [6]. The history of the development of CT is extremely interesting and will be summarized here, but to put it into a broader perspective we will review briefly the X-rays acquisition and processing techniques in use before CT was introduced.

Röntgen's discovery of X-rays in 1895 revolutionized the practice of medicine by allowing a view of the interior of the human body. However, an X-ray image is the 2-D projection of a 3-D body, which means that its interpretation is subject to ambiguity. This fact was noticed soon after the radiographic technique was introduced, and a number of researchers began to develop techniques to produce 3-D views (see, e. g., [155]). The goal was to generate X-ray images of thin slices of a patient. This was achieved by moving the film and the X-ray source in such a way that the objects in a particular plane were emphasized while others were blurred by the motion. This process is discussed in, e. g., [138]. Work on this goal began in 1914 in several European countries (Germany, France, Italy, Holland), and was motivated, at least in part, by World War I. Different approaches were tried and different names were assigned to some of them, but only one, tomography, remained. This word comes from the Greek word *tomos*, which means cut or slice, and was introduced by the Berliner physician G. Grossmann, whose tomography was commercially available in 1935 [155,156].

Conventional tomography represented a considerable improvement over the original radiographic techniques, but the advent of computers opened new research avenues based on the digital processing of the X-ray images, or

shadowgrams. A shadowgram is a photographic plate developed after it has been illuminated by X-rays that passed through an object and gives a measure of the absorptivity of the rays by the object, which in many materials is roughly proportional to its density [12]. The availability of computers allowed shadowgrams, as well as other images, to be scanned and digitized for further processing using Fourier transform techniques as well as numerical solution of matrix equations. This work will be referred to below, but it is important to realize that they were in place when CT was introduced, and were major contributors to the improvement of the quality of the early CT results.

One of the earliest papers on medical computerized tomography is by A. Cormack [26], who received the Nobel Prize in Physiology or Medicine in 1979 together with G. Hounsfield for their contributions to the development of the technique, which were carried out independently. Cormack was a South African physicist working at the University of Cape Town. In 1955 the physicist at the Cape Town hospital resigned and Cormack replaced him for six months in 1956. During this time he became interested in the determination of the absorption of X or gamma rays passing through an inhomogeneous medium. His motivation was medical, in the context of radiotherapy, which required a good knowledge of the absorption coefficient of the bones and tissues of a patient. Cormack's approach was to consider a 2-D problem, as a 3-D one can be solved in terms of a succession of 2-D layers. The 2-D problem was well known and can be formulated as follows. A beam of monoenergetic rays of intensity  $I_o$  traverses a finite 2-D domain  $\mathcal{D}$  along a straight line  $L$ , and the intensity of the ray emerging from  $\mathcal{D}$  is  $I$ , given by

$$I = I_o \exp \int_L f(l) dl \quad (1)$$

where  $f$  is the absorption coefficient, which is a function of position within  $\mathcal{D}$ , and  $dl$  denotes a length element along  $L$ . Dividing both sides of Eq. (1) by  $I_o$  and taking the natural logarithm gives

$$g_L \equiv \ln \frac{I}{I_o} = \int_L f(l) dl . \quad (2)$$

The quantity  $g_L$  is known, and the question posed by Cormack was whether  $f$  could be computed using  $g_L$  determined for a number of lines. Cormack solved the problem in terms of a series expansion and demonstrated the feasibility of the method with two simple samples made of aluminum and either wood or Lucite [26,27]. His work, however, went essentially unnoticed until the publication of a paper [28] where the connection between the CT



problem and the Radon transform was discussed. That paper in turn, was stimulated by Hounsfield's work, which became known in 1971 [29]. Cormack moved to Tufts University (United States) in 1957.

The actual implementation of the CT technique as a viable commercial enterprise is due to Hounsfield [65], who was an engineer working at the Central Research Laboratories of Electrical and Musical Industries, the English company that eventually became the well known musical records company EMI. As recounted by Hounsfield [67], after finding that a project he had been working on would not be commercially viable, he was given the opportunity to think about other areas of research that he thought might be more fruitful. This rare opportunity was probably the result of the considerable amount of money that the group *The Beatles* had brought to EMI [107]. One of the projects Hounsfield had been working on was connected with automatic pattern recognition, which in 1967 led to the idea of what eventually became the technique of computerized tomography. To materialize this idea Hounsfield built the CT machine and developed the numerical technique to process the data (see Subsect. "Iterative Solutions"). According to some, the second task may have been the more fundamental of the two [107].

We close this summary with two notes of historical interest. First, B. Korenblyum, S. Tetel'baum, and A. Tyutin worked on tomography at the Kiev Polytechnic Institute and published their results in obscure Russian journals in 1957 and 1958. These authors formulated the tomography problem using line integrals, solved it exactly in terms of the inverse Radon transform, and discussed a recon-

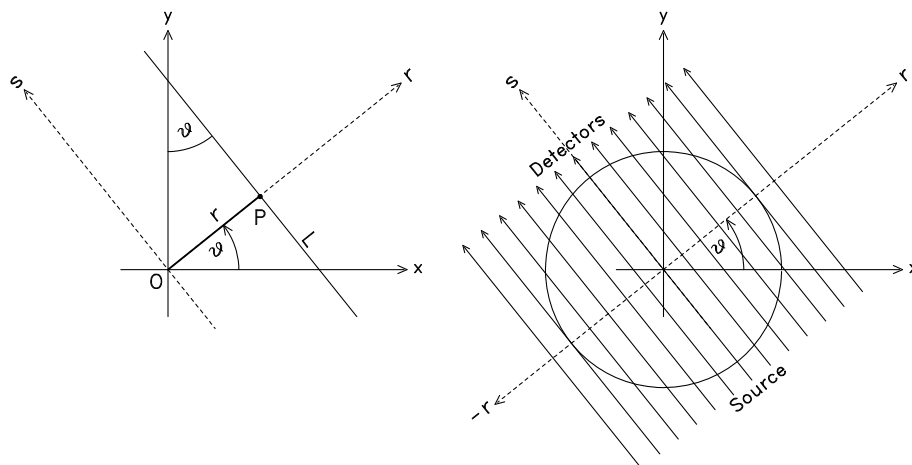
struction algorithm. However, additional references on their work could not be found [11]. Second, W. Oldendorf, a neurologist at the University of California, Los Angeles, developed a scanning instrument, although he did not solve the problem mathematically [106]. Oldendorf patented his instrument in 1963 but was not able to attract companies interested in manufacturing a commercial version of it. Because of this contribution, Oldendorf was considered for the Nobel Prize together with Cormack and Hounsfield, but eventually was excluded. Possible reasons for the decision of the Nobel committee and the politics involved can be found in [19].

### Solving the CT Problem

The computerized tomography problem is to solve Eq. (2) when  $g_L$  has been determined for a large number of lines having different positions and directions. To make this problem more precise, the definition of projection will be introduced. First consider the following line integral of a function  $f(x, y)$

$$p(r, \vartheta) = \int_L f(x, y) dl \quad (3)$$

where the line  $L$  is defined in terms of its angle  $\vartheta$  with the  $y$  axis and its distance  $r$  to the origin (Fig. 1). Alternatively, it can be said that  $L$  passes through point  $P$  with polar coordinates  $(r \cos \vartheta, r \sin \vartheta)$  and is perpendicular to  $\overline{OP}$ . The collection of line integrals along parallel lines is known as a *parallel projection* of  $f(x, y)$ . Because this is the only type of projection considered here, the qualifier



Tomography, Seismic, Figure 1

**Left:** Geometry for the definition of projection.  $L$  is the projection line and  $r$  is the distance from the line to the origin. The axis  $s$  is parallel to  $L$ . **Right:** Geometry for parallel X-ray tomography. The circle bounds an object to be projected. Each line is equivalent to the line  $L$  on the left. There is a source of X-rays on one end of the lines and detectors on the opposite end

parallel will be dropped. In terms of Eq. (3), the projection of  $f$  is obtained by letting  $r$  be a variable while keeping the value of  $\vartheta$  fixed (i. e.,  $\vartheta$  must be considered a parameter). Note that some authors call a particular line integral a projection along that line (see, e. g., [132]).

Equation (3) was solved using two different approaches that broadly speaking can be referred to as analytical and iterative. The analytical approach followed different paths, but in each case the result was a closed-form solution, which had to be solved numerically. This approach is relevant to seismology for at least two reasons. First, it provides the theoretical underpinnings of the concept of backprojection, which has entered the seismic tomography literature. Second, it is directly related to the concepts of slant-stack and Radon transform popular in the reflection seismology literature. The iterative approach was to write the CT problem in matrix form and to solve it iteratively. This approach was widely applied to the solution of seismological inverse problems. Because of their close relation to seismology, the two approaches will be discussed here.

**Analytical Solutions** The most popular solutions are based on the use of the Fourier transform, but before considering them we will briefly mention the work of Radon, the Bohemian mathematician that investigated Eq. (3) in a paper published in 1917. It was written in German, and an English translation (by R. Lohner) can be found in [35].

Radon considered the function

$$\begin{aligned} p(r, \theta) &= p(-r, \theta + \pi) \\ &= \int_{-\infty}^{\infty} f(r \cos \vartheta - s \sin \vartheta, r \sin \vartheta + s \cos \vartheta) ds \end{aligned} \quad (4)$$

(see Eq. (11) below; his symbols were somewhat different), and the mean value of  $p(r, \theta)$  for the lines tangent to a circle with center at a point  $Q = (x, y)$  and radius  $a$

$$\bar{p}_Q(a) = \frac{1}{2\pi} \int_0^{2\pi} p(x \cos \vartheta + y \sin \vartheta + a, \theta) d\theta \quad (5)$$

and proved that

$$f(Q) = -\frac{1}{\pi} \int_0^{\infty} \frac{d\bar{p}_Q(a)}{a} = -\frac{1}{\pi} \int_0^{\infty} \frac{d\bar{p}_Q}{da} \frac{da}{a}, \quad (6)$$

where the last equality is given in [28]. A proof of Radon's result can be found in [62]. Equations (4) (or equivalent expressions) and (6) are known as the Radon and inverse Radon transform, respectively. Although Eq. (6) looks simple, its practical implementation is not [132] and probably for this reason it did not receive as much at-

tention (after Cormack's 1973 paper, [28]) as the other methods developed to solve the problem. As a matter of historical interest we note that Radon's problem arises in a number of scientific fields and had been solved independently more than once in the early part of the twentieth century [30]. Yet, these results were not widely known and had to be derived again.

The analytical approach developed along two different lines and produced results formally different from that of Radon. One was pioneered by Cormack [26,27], but, as noted earlier, it went essentially unnoticed. A second line originated in radio astronomy [17,18], electron microscopy (e. g., [34,36]) and X-ray radiography (e. g., [12,124]), and was based on the use of the Fourier transform. This is the approach that will be taken here and its description will be based mainly on [20,76,88] and [132].

The early development of the CT technique was based on the use of a number of parallel projections determined for different values of  $\vartheta$  in Eq. (3). This is the case that will be analyzed here. To find the equation that represents  $L$  we will use that fact that its slope is equal to  $-1/\tan \vartheta$  and its intercept with the  $y$  axis is  $r/\sin \vartheta$ . Thus

$$y = -\frac{\cos \vartheta}{\sin \vartheta} x + \frac{r}{\sin \vartheta} \quad (7)$$

so that

$$x \cos \vartheta + y \sin \vartheta = r. \quad (8)$$

To tie this definition to the concept of *slant stack* in seismology we note that it is defined using Eq. (3) with  $f$  replaced by a function  $u(x, t)$ , where  $u$ ,  $x$ , and  $t$  represent seismic wave amplitude, distance, and time, respectively,  $y$  is replaced by  $t$ , and  $t$  is written in terms of the intercept and slope of the line  $L$  (generally indicated with  $\tau$  and  $p$ , so that  $t = px + \tau$ ) (see, e. g., [40,127]).

Now we will introduce a new coordinate system  $(r, s)$  obtained by a rotation of angle  $\vartheta$  of the  $(x, y)$  system (Fig. 1). In the  $(r, s)$  system the points on the line  $L$  have constant  $r$  and variable  $s$ . The two systems are related by the following transformations of coordinates

$$\begin{pmatrix} r \\ s \end{pmatrix} = \begin{pmatrix} \cos \vartheta & \sin \vartheta \\ -\sin \vartheta & \cos \vartheta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \cos \vartheta + y \sin \vartheta \\ -x \sin \vartheta + y \cos \vartheta \end{pmatrix} \quad (9)$$

and

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos \vartheta & -\sin \vartheta \\ \sin \vartheta & \cos \vartheta \end{pmatrix} \begin{pmatrix} r \\ s \end{pmatrix} = \begin{pmatrix} r \cos \vartheta - s \sin \vartheta \\ r \sin \vartheta + s \cos \vartheta \end{pmatrix}. \quad (10)$$

Using Eq. (10), Eq. (3) can be rewritten as

$$\begin{aligned}
 p(r, \theta) &= \int_{-\infty}^{\infty} f(r \cos \vartheta - s \sin \vartheta, r \sin \vartheta + s \cos \vartheta) ds \\
 &\equiv \int_{-\infty}^{\infty} \hat{f}_{\vartheta}(r, s) ds
 \end{aligned}
 \tag{11}$$

where  $\hat{f}_{\vartheta}$  represents the function  $f$  when it is written in terms of  $r$  and  $s$ , and the subscript  $\vartheta$  is used to emphasize that it enters in the computations as a parameter. In Eq. (11)  $r$  is allowed to vary continuously, although in practical applications (such as CT)  $r$  is a discrete variable, as sketched in Fig. 1. Also note that  $r$  is allowed to be negative, which means that  $0 \leq \vartheta < \pi$ , and that although  $|r|$  is allowed to extend to infinity, for functions defined over a finite domain in the  $(x, y)$  plane the projections for lines outside of the domain will be zero.

To proceed further we will work in the wavenumber domain. Let the one-dimensional Fourier transform (or 1-D F.T.) of  $p(r, \theta)$  with respect to  $r$  be

$$P(k, \theta) = \int_{-\infty}^{\infty} p(r, \theta) e^{-i2\pi kr} dr
 \tag{12}$$

where  $k$  indicates wavenumber (equivalent to the frequency in the time domain). Introducing Eq. (11) into this expression and then going back to the  $(x, y)$  coordinate system gives

$$\begin{aligned}
 P(k, \theta) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \hat{f}_{\vartheta}(r, s) e^{-i2\pi kr} dr ds \\
 &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) e^{-i2\pi k(x \cos \vartheta + y \sin \vartheta)} dx dy \\
 &= F_{\vartheta}(k_x, k_y)
 \end{aligned}
 \tag{13}$$

where  $r$  has been written using Eq. (9), the Jacobian of the coordinates transformation is equal to one,  $F_{\vartheta}$  is the two-dimensional Fourier transform of  $f$  and

$$k_x = k \cos \vartheta; \quad k_y = k \sin \vartheta.
 \tag{14}$$

Note that

$$k^2 = k_x^2 + k_y^2; \quad k_y/k_x = \tan \theta.
 \tag{15}$$

In summary,

$$P(k, \theta) = F_{\vartheta}(k_x, k_y).
 \tag{16}$$

In words, the 1-D F.T. of the projection  $p(r, \theta)$  of  $f(x, y)$  is equal to the 2-D F.T. of  $f(x, y)$ . This relation is known as the Fourier (central) slice theorem. This name comes from the fact that the 2-D F.T. is known in a slice (i. e.,

a line) through the origin in the  $(k_x, k_y)$  space, as Eq. (15) shows. Moreover, the angle of this line with the  $k_x$  axis is  $\vartheta$ , which is equal to the angle between the  $x$  and  $r$  axes. Equation (16) suggests one approach to the determination of  $f(x, y)$ . Find the 1-D F.T. of the projections for discrete values of  $\theta$  between 0 and  $\pi$  and then use the 2-D inverse F.T. to recover  $f(x, y)$  numerically. However, because the values of the 2-D F.T will be defined on a polar coordinates grid, it must be interpolated to a Cartesian grid in the  $(k_x, k_y)$  space.

A different approach is as follows. Let  $F(k_x, k_y)$  be the 2-D F.T. of  $f(x, y)$ . Then,  $f(x, y)$  is given by the inverse F.T.

$$f(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F(k_x, k_y) e^{i2\pi(k_x x + k_y y)} dk_x dk_y.
 \tag{17}$$

In this expression  $k_x$  and  $k_y$  are defined on a Cartesian coordinate system, which will be converted to polar coordinates using Eq. (14). This gives

$$f(x, y) = \int_0^{\pi} \int_{-\infty}^{\infty} P(k, \vartheta) e^{i2\pi k(x \cos \vartheta + y \sin \vartheta)} |k| d\vartheta dk
 \tag{18}$$

where Eq. (16) was used and  $|k| d\vartheta dk$  is the Jacobian for the transformation of coordinates. Because the Jacobian relates the elements of area in the two coordinate systems, the absolute value of  $k$  is needed because the area is positive. Two comments are in order here. First,  $\vartheta$  is no longer a parameter; now it is used as a variable. Second, the polar coordinates have been defined using  $0 \leq \theta < \pi$  and  $-\infty < k < \infty$ . This convention is equivalent to the standard one (i. e.,  $0 \leq \theta < 2\pi, 0 \leq k < \infty$ .)

Now Eq. (18) will be rewritten as follows

$$f(x, y) = \int_0^{\pi} p^*(x \cos \vartheta + y \sin \vartheta, \vartheta) d\vartheta
 \tag{19}$$

where

$$p^*(r, \vartheta) = \int_{-\infty}^{\infty} P(k, \vartheta) |k| e^{i2\pi kr} dk.
 \tag{20}$$

The integral in Eq. (20) involves the product of two 1-D F.T.s,  $P(k, \theta)$  and  $|k|$ , so, in principle, one way to solve it would be to use the following property. Given two functions  $g(r)$  and  $h(r)$  with F.T.s  $G(k)$  and  $H(k)$ , then

$$\int_{-\infty}^{\infty} G(k) H(k) e^{i2\pi kr} dk = \int_{-\infty}^{\infty} g(\rho) h(r - \rho) d\rho
 \tag{21}$$

where the integral on the right-hand side is the convolution of  $g$  and  $h$  (see, e. g., [94]). A problem with this formulation is that  $|k|$  does not have an inverse F.T., as can

be seen from the fact that it does not go to zero as  $k$  goes to infinity, which is one of the properties of the F.T. [149]. There are two ways to get around this problem. One is to solve Eq. (21) in terms of distributions. A basic yet rigorous introduction to distributions, which can be considered an extension of the concept of function, can be found in [120]. Let us rewrite Eq. (20) as follows

$$p^*(r, \theta) = \frac{1}{2\pi^2} \int_{-\infty}^{\infty} P(k, \theta) (i2\pi k) (-i\pi \operatorname{sgn} k) e^{i2\pi kr} dk \tag{22}$$

where

$$\operatorname{sgn} k = \begin{cases} 1 & k > 0 \\ -1 & k < 0. \end{cases} \tag{23}$$

Then, using Eq. (21) we can write

$$p^*(r, \theta) = \frac{1}{2\pi^2} \mathcal{F}^{-1} \{ i2\pi k P(k, \theta) \} * \mathcal{F}^{-1} \{ -i\pi \operatorname{sgn} k \} \tag{24}$$

where  $\mathcal{F}^{-1}$  represents the inverse F.T. of the function in braces and the  $*$  stands for convolution. The first inverse is just  $\partial p(r, \theta) / \partial r$  and the second inverse is equal to  $1/r$  [120]. Combining these two results gives

$$p^*(r, \theta) = -\frac{1}{2\pi} \left( -\frac{1}{\pi r} * \frac{\partial p(r, \theta)}{\partial r} \right) \equiv -\frac{1}{2\pi} \mathcal{H} \left\{ \frac{\partial p(r, \theta)}{\partial r} \right\} \tag{25}$$

where  $\mathcal{H}$  stands for the Hilbert transform of the function in braces [120]. The inverse expression for the slant stack can be derived using expressions similar to those presented above [25].

The importance of Eq. (25) is mostly theoretical, and from a practical point of view a different method of solution was found to be more useful. Before proceeding, however, note that Eq. (20) without  $|k|$  on the right-hand side corresponds to the 1-D inverse F.T. of  $P(k, \theta)$ , equal to  $p(r, \theta)$  (see Eq. (12)). Therefore,  $p^*$  is a *filtered projection*, with  $|k|$  the filter response. Clearly, the effect of the filter is to amplify the components of  $p$  corresponding to the higher wavenumbers. The approach used here is also based on the use of Eq. (21), with  $g(r) = p(r, \theta)$  (as before) and  $h(r)$  a function whose F.T. is an approximation to  $|k|$  in the sense that

$$H(k) \approx \begin{cases} |k|; & |k| < K \\ 0; & \text{elsewhere} \end{cases} \tag{26}$$

where  $K$  can be taken as the spatial Nyquist frequency  $k_N$  (equal to  $1/2a$ , where  $a$  is the spacing between projection

lines). To avoid aliasing,  $P(k, \theta)$  must be zero for  $|k| > k_N$ . Clearly, there is some freedom in the selection of  $h(r)$ . For example, in [18] and [124]  $H(k)$  is defined by an expression similar to Eq. (26) with an equal sign in place of the less than sign. This work was followed by the introduction of an improved function [133]. For our purposes, however, what is important is not the particular choice of  $h$  but the fact that we can write the following approximation for  $f(x, y)$

$$f(x, y) \approx \tilde{f}(x, y) \equiv \int_0^\pi \int_{-\infty}^{\infty} p(\rho, \theta) h(x \cos \vartheta + y \sin \vartheta - \rho) d\rho d\theta. \tag{27}$$

Finally, because the projections are determined for discrete and equispaced values of  $r$  and  $\theta$  we can write

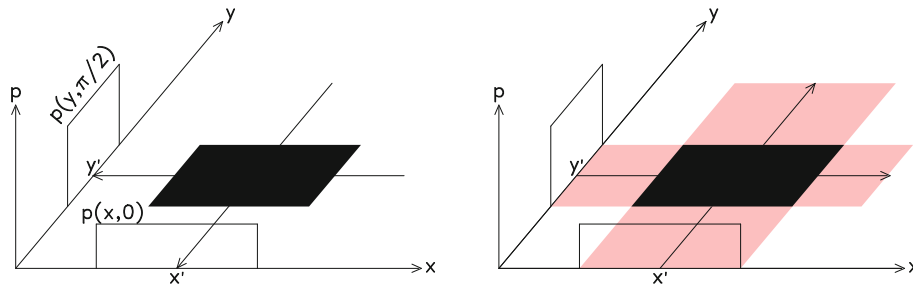
$$\tilde{f}(x, y) \approx \frac{\pi a}{N} \sum_{i=1}^N \sum_{j=-J}^J p(ja, \theta_i) h(x \cos \vartheta_i + y \sin \vartheta_i - ja) \tag{28}$$

where  $\theta_i = (i-1)\pi/N$ ,  $N$  is fixed, and  $J$  is a finite limit that takes into account the finite size of the object (after [132]).

To end this discussion of the transform methods of solution we will go back to Eq. (19). The integration there is an example of the operation known as *backprojection*, to be indicated with  $\mathcal{B}$  and defined by

$$\mathcal{B}\{q\}(x, y) = \int_0^\pi q(x \cos \vartheta + y \sin \vartheta, \vartheta) d\theta \tag{29}$$

where  $q$  is an arbitrary function of  $r$  and  $\vartheta$ . An example of this operation is provided in Fig. 2a, which shows two projections of a rectangular object corresponding to  $\vartheta$  equal to 0 and  $\pi/2$ . The two projections, identified by  $p(x, 0)$  and  $p(y, \pi/2)$ , are nonzero for limited ranges of the  $y$  and  $x$  axes. To compute the backprojection let us consider a particular value  $y'$  of  $y$  and assign the value  $p(y', \pi/2)$  to every point in the line defined by the points  $(x, y')$ . The object may be part of a larger body, and for this reason the line will extend beyond the  $x$  values that bound the object. Repeating this operation for all the values  $y'$  and then proceeding in a similar way for the  $p(x, 0)$  projection produces an image such as that shown in Fig. 2b. The two backprojections overlap in the area occupied by the rectangular object, but also contribute values to other areas of the  $(x, y)$  plane, which is clearly not correct. The problem is not related to the use of only two projections; even when a larger number of projections is used, the contributions outside of the object do not disappear. The reason for that is that the determination of  $f(x, y)$  requires  $p^*$ , not  $p$ , as shown by Eq. (19) (see, e. g., [23]).



Tomography, Seismic, Figure 2

Illustration of the concept of back projection. *Left*: projection of a rectangular box using two sets of parallel lines (two representative lines are shown). *Right*: backprojection of the two projections shown on the left. Note the change in the direction of the arrows. For each line, the value of its projection is assigned to all the points along the line and are added to the values that come from all the other projections. See text for details. After [20]

**Iterative Solutions** As was the case with the analytical solutions, the iterative solutions were developed before the advent of CT. The earliest one was the *algebraic reconstruction technique* (ART, [51]) and was developed for the solution of electron microscopy problems as an alternative to analytical solutions based on the Fourier transform. These problems are represented by an equation similar to Eq. (2). The basic idea behind ART was the discretization of a plane object in terms of a square grid of points. The goal was to find the optical density  $\rho_{ij}$  at each point  $(i, j)$  of the grid. A ray of a projection at an angle  $\theta$  was defined as a band of width  $w$  across the plane at the same angle, and corresponds to each of our lines  $L$ . One possible choice for  $w$  is to make it equal to the grid spacing. The basis of ART is to start with an average value of the density (computed from the observations) and to use it to compute the projections for all the rays for all the angles. Then, two updating methods were used, one additive and the other multiplicative. In the original version of the additive method, for each ray the difference between observed and computed projections was used to update the density values in the cells along the ray. If an updated value of density became negative it was set to zero. This was done for all the rays, one ray at a time. This completed one iteration. Then a new one was started with the updated densities used as starting values. Originally, the difference between the observed and computed projections was divided equally among all the cells in the ray, but in later applications each cell received a weight representing its contribution to a given ray. The weight was used when the cell densities were updated (see, e. g., [50]). In the multiplicative version of ART the density value of a cell was updated by multiplying it by the ratio of computed to observed projections. Again, a weighting scheme was introduced later. A technique similar to additive ART was developed inde-

pendently by Hounsfield and applied to his X-ray scanner (see, e. g., [20,133]).

A second technique, simultaneous iterative reconstruction technique (SIRT) was introduced in [48], which was highly critical of the performance of ART. SIRT also has additive and multiplicative versions, and the main difference with the ART counterparts is that at each iteration the density of each cell is updated using all the projections passing through that cell. Subsequent work showed that although ART is computationally more efficient than SIRT, it has the problem that is more affected by errors in the data (see, e. g., [20,39,63]). SIRT, on the other hand, has some undesirable properties, referred to in Subsect. “Regularization Approach”.

### Arrival-Time Seismic Tomography

As noted earlier, this type of tomography is much simpler than surface wave and waveform tomography, and lends itself to a fairly self-contained discussion, to be presented here. The other two types of tomography will not be discussed here because they require a knowledge of seismic theory and data processing beyond the scope of this article.

This section is organized as follows. First, the differences between seismic and X-ray tomography will be discussed. The early applications of CT techniques to seismic problems ignored these differences, and although the results thus obtained opened the way to a new research approach in seismology, they may have been affected by several types of errors, the sources of which will become clear here. Second, tomography using local data will be addressed. In this case both the location of the events as well as the velocity model must be determined. In addition, a distinction between velocity and slowness tomography must be made. As noted earlier, before seismic to-

mography there was velocity inversion. Then, after the introduction of CT techniques the inversion parameter became slowness, not velocity. Therefore, the equations to be solved in the two cases are different, and because some of the velocity inversion programs are still in use, the two parameterizations, and their relationship, will be presented here. In addition, regardless of the parameterization used, it is necessary to decouple (or separate) the location part from the tomographic part of the problem, which requires introducing new analytical tools. Another topic related to local tomography is the computation of travel times, which will be briefly considered here. Finally, the case of tomography using teleseismic arrival times is analyzed.

### Comparison with X-ray Tomography

Seismic tomography differs from X-ray tomography in two fundamental ways because, first, the former is nonlinear, and second, the locations of the seismic sources are generally unknown (i. e., when earthquakes are used). Another major difference is that in X-ray tomography there is control over position and number of sources and receivers, which is not the case for seismic tomography. The first two differences will be considered in more detail.

The expression for travel time,  $t$ , along a ray is given by

$$t = \int_R \frac{1}{v} ds = \int_R u ds \quad (30)$$

where  $R$  denotes the raypath between two fixed points within an elastic medium with velocity  $v(\mathbf{x})$  that depends on position (indicated by the vector  $\mathbf{x}$ ),  $ds$  is a line element along the raypath, and  $u$  is the slowness, equal to the inverse of velocity

$$u(\mathbf{x}) = \frac{1}{v(\mathbf{x})}. \quad (31)$$

An advantage of writing  $t$  in terms of  $u$  instead of  $v$  is that in terms of  $u$ ,  $t$  has an expression similar to Eq. (2), which constitutes the basis of X-ray tomography. This similarity, however, is more apparent than real. To see that we will introduce the concept of linear operation. Given an operation  $\mathcal{O}$ , it is said to be linear if

$$\mathcal{O}(f + h) = \mathcal{O}(f) + \mathcal{O}(h) \quad (32)$$

and

$$\mathcal{O}(\alpha f) = \alpha \mathcal{O}(f) \quad (33)$$

where  $f$  and  $h$  are functions and  $\alpha$  is a scalar. A simple example of linear operation is the integration over a given

interval

$$\int_a^b [f(x) + h(x)] dx = \int_a^b f(x) dx + \int_a^b h(x) dx. \quad (34)$$

In particular, the operation defined by Eq. (2) is linear because  $L$  is always a straight line that does not depend on  $f$ . Therefore, if we let  $f = f_1 + f_2$  we can write

$$\begin{aligned} g_L(f) &= \int_L [f_1(s) + f_2(s)] ds \\ &= \int_L f_1(s) ds + \int_L f_2(s) ds \\ &= g_L(f_1) + g_L(f_2). \end{aligned} \quad (35)$$

Now let us consider the travel time problem. Let  $R(u_1)$  and  $R(u_2)$  denote the raypaths between two fixed points within an elastic medium assuming that it had slownesses  $u_1$  and  $u_2 \neq u_1$  at two different times. This situation is generally not possible in the earth but can be simulated on a computer. The corresponding travel times are given by

$$t_1 = \int_{R(u_1)} u_1(s) ds, \quad t_2 = \int_{R(u_2)} u_2(s) ds. \quad (36)$$

Note that the two raypaths will be different because  $u_1$  and  $u_2$  have been assumed to be different. The trivial case of  $u_1 = ku_2$ , where  $k$  is a constant, is ignored. Therefore, it will not be possible, in general, to write  $t_1 + t_2$  as an integral over a common path involving  $u_1 + u_2$ . Moreover, if the medium has slowness  $u_1 + u_2$  the raypath  $R(u_1 + u_2)$  between the two points will be different from  $R(u_1)$  and  $R(u_2)$  and, in general,

$$t_1 + t_2 \neq \int_{R(u_1+u_2)} [u_1(s) + u_2(s)] ds \quad (37)$$

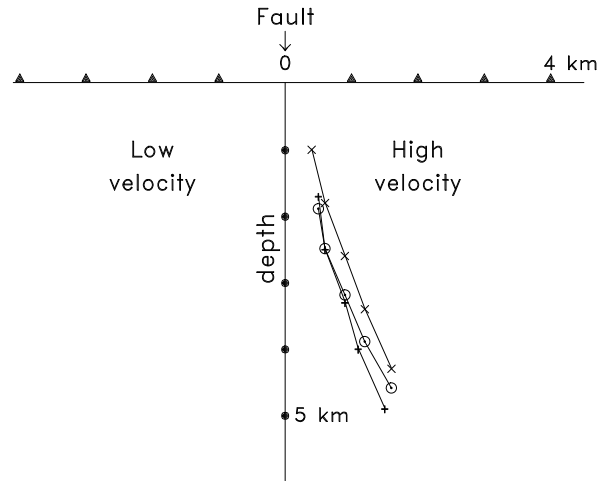
which in turn means that the travel time problem cannot be expressed in terms of a linear operation. Therefore, the seismic tomography problem is nonlinear, and the CT solution techniques cannot be applied directly except when the 3-D velocity variations in the medium are small, a condition that severely limits its application to situations of seismic interest.

Another major difference between seismic tomography and CT is that the locations of the sources (mostly earthquakes) are generally not known. Therefore, the determination of a velocity model requires the simultaneous determination of the source locations. This is particularly important when the events are locally recorded. To see that we will consider a very simple example. Let us assume that a vertical fault separates rocks with low- and high-velocities, denoted  $v_1$  and  $v_2$ , respectively, and that

earthquakes occur along a vertical line within the fault. Let  $v_1 = 3$  km/s and  $v_2 = 4$  km/s. The origin of the coordinate system will be the fault location at the surface and the event depths will be 1, 2, 3, 4, and 5 km. The earthquakes will be assumed to be recorded by a local network of eight stations on the surface. The theoretical arrival times computed for this geometry were used with a single-event location program and two constant velocity models with velocities equal to 3 and 3.5 km/s and with a 1-D model with five layers and velocities between 2.5 and 4 km/s. The locations obtained with these models (Fig. 3) are affected by significant errors, which cannot be inferred by the small root-mean-square errors, which range between 0.02 and 0.09 s, with the larger value corresponding to the deeper event. These results show that the effect of the lateral juxtaposition of high- and low-velocity rocks when the events are located with 1-D velocity models is a mislocation of the events in a direction away from the low-velocity zone. A well-documented example of this situation is provided by the aftershocks of the 1994  $M = 6.7$  Northridge, California, earthquake, which occurred within the sedimentary rocks of the San Fernando basin. These events and the velocity model determined using them are considered in Subsect. “[P Wave Velocity Model for the San Fernando, California, Area](#)”.

This example has a counterpart. What happens when mislocated events are kept at their erroneous locations and are used to determine a 2-D velocity model around the fault? Clearly, such a velocity model will be affected by the location errors and its reliability will be questionable. Examples showing the significant bias that these errors may have in the computed velocities are given in [54,145], and [146] for geometries similar to that in Fig. 3. In [146],  $v_1 = 5$  km/s and  $v_2 = 6$  km/s and there was only one earthquake on the fault. In this case the velocity determined by inversion was 5.44 km/s to the left of the fault and 5.56 km/s to the right of it. This example is clearly an oversimplification, but the result is useful because it shows that ignoring the mislocation introduced by lateral velocity variations will result in a velocity model showing variations smaller than the actual ones. For the case of teleseismic tomography, event mislocation may not be a serious problem except when earthquakes from subducting slabs are used, in which case the effect on the computed velocities may be important (see, e. g., [32,153]).

Another difference between seismic tomography and X-ray tomography is that in the later the position of the sources and the receivers can be chosen so that the computed images have the desired resolution. In contrast, in seismic tomography there is no control over the position of the earthquakes, while the number and position of sta-



Tomography, Seismic, Figure 3

Simple example (2-D) of the effect of lateral velocity variations on earthquake location when those variations are ignored. A fault divides the medium into two quarter spaces with velocities equal to 3 km/s (left) and 4 km/s (right). The triangles and dots represent stations and hypocenters, respectively. Synthetic arrival times computed for these events, stations, and velocity model are used with a single-event location program and three 1-D velocity models. Two models have constant velocities, equal to 3 km/s and 3.5 km/s, and the other has five layers with velocities between 2.5 and 4 km/s. The locations determined using these models are indicated by pluses, crosses and circles

tions is dictated by financial considerations as well as practical constraints imposed by the nature of the terrain. For example, mountains frequently constitute a severe obstacle to station deployments. Even more serious is the presence of oceans, which cover two-thirds of the earth's surface and essentially prevent the routine deployment of permanent seismic networks similar to those on land. As a consequence, seismic tomographic models, particularly at a global scale, are affected to some extent by low-resolution problems.

Finally, we note that soon after X-ray tomography began, several competing methods of solution were proposed (see Subsect. “[Solving the CT Problem](#)”), and to compare their performances synthetic data generated for a model of a cross section of the skull were used [133]. With this approach it was possible to detect artifacts in some of the solutions and to investigate the effect of noise in the data. In contrast, seismic tomography methods were rarely subject to a similar validation analysis in spite of the approximations and simplifying assumptions made. These differences stem, in part, on the complicated nature of wave propagation in the earth, which is difficult to replicate in a computer on account of theoretical difficulties and computational cost. On the other hand, the propaga-

tion of X-rays is relatively simple and easy to simulate on a computer. In addition, medical tomography, in general, is well funded, while seismic tomography is only poorly funded [84].

### Local Velocity Tomography

The simultaneous determination of the locations of a group of earthquakes and a 3-D velocity model can be considered an extension of the standard method of earthquake location. Usually, the velocity structure is modeled in terms of constant-velocity blocks or by velocity values on a grid. The basic ideas were introduced by Aki and Lee [2] and Crosson [33], although the latter only considered layered velocity models. Because this inverse problem is nonlinear in both the earthquake location and the velocity determination, it is solved by linearizing it about initial estimates of the locations and origin times of the events and model velocities. With this approach the original problem is replaced by other in which the unknowns are adjustments to the initial estimates. Then these adjustments are added to the initial estimates, the corresponding values are used as new initial estimates, and the process is repeated iteratively until some stopping criterion is met. This process will be described mathematically in the following.

Let us use subscripts  $i$  and  $j$  to identify the stations and earthquakes used, respectively, and let  $M$  and  $N_j$  be the number of events and the number of stations with arrival times for the  $j$ th event, respectively. The blocks or grid points used to parameterize the velocity structure constitute a 3-D array, but for computational purposes each block or point will be identified with a single subscript, say  $k$ . This requires establishing an ordering scheme that assigns a single index to a triplet of indices. For example, we may use the following scheme:  $(1, 1, 1) \rightarrow 1$ ,  $(2, 1, 1) \rightarrow 2$ ,  $(n, 1, 1) \rightarrow n$ ,  $(n, 2, 1) \rightarrow n + 1$ , and so on. Let  $K$  represent the total number of velocity parameters. The linearized problem can be written as

$$w_{ij}r_{ij} = w_{ij} \left( dT_j + \frac{\partial t}{\partial x} dx_j + \frac{\partial t}{\partial y} dy_j + \frac{\partial t}{\partial z} dz_j + \sum_k \frac{\partial t}{\partial v_k} dv_k \right) \quad (38)$$

$$i = 1, \dots, N_j \quad j = 1, \dots, M$$

where

$$r_{ij} = \mathcal{T}_{ij}^{\text{obs}} - \mathcal{T}_{ij}^{\text{comp}} \equiv \mathcal{T}_{ij}^{\text{obs}} - (T_j + t_{ij}) \quad (39)$$

is the arrival time residual for the  $j$ th earthquake and the  $i$ th station. The subscript  $k$  labels the blocks or grid

points associated with the raypath from the  $j$ th hypocenter to the  $i$ th station. Therefore,  $k$  is a function of  $i$  and  $j$ , but this dependence is left implicit to simplify the notation. Also note that the values of  $k$  associated with a given ray do not have any particular ordering. The meaning of the other variables is as follows:

$dT_j, (dx_j, dy_j, dz_j), dv_k$ : adjustments to origin time, hypocentral coordinates, and velocities, respectively,  
 $\mathcal{T}_{ij}^{\text{obs}}$ : observed arrival time,  
 $\mathcal{T}_{ij}^{\text{comp}}$ : computed arrival time,  
 $T_j$ : origin time,  
 $t_{ij}$ : computed travel time,  
 $t, v$ : travel time and velocity,  
 $w_{ij}$ : quality weight.

In Eqs. (38) and (39) the only unknowns are the adjustments. All the other quantities are computed using initial estimates of the velocities ( $v_k$ ) and the origin time ( $T_j$ ) and hypocentral coordinates ( $x_j, y_j, z_j$ ) of each earthquake. The expressions for the derivatives can be found in [85].

### Local Slowness Tomography

As noted earlier, the travel time problem cannot be expressed in terms of a linear operation on slowness, and for this reason the problem is linearized assuming that the difference between the initial and actual velocity models is small. Let us go back to Eq. (36) with  $u_2 = u_1 + du$ ,  $du \ll u_1$ . Then the difference  $dt$  in travel times is given by

$$dt = t_2 - t_1 = \int_{R(u_1+du)} [u_1(s) + du(s)] ds - \int_{R(u_1)} u_1(s) ds \approx \int_{R(u_1)} du(s) ds. \quad (40)$$

Here we have invoked *Fermat's principle*, which allows replacing  $R(u_1 + du)$  with  $R(u_1)$ . Recall that this principle states that raypaths are paths of stationary travel time (see, e. g., [3,120]), which in turn means that the variation of travel time along a raypath is zero. Two points must be noted here, however. First, the principle applies to small variations in raypaths, which in turn requires a small  $du$ , a condition which may not be valid in reality. This question was tested in the context of the propagation of rays through subducting slabs [31] and it was found that arrival times at teleseismic distances calculated using Fermat's principle were the same or later than those determined using exact ray tracing. The difference between exact and approximate times depended on event depth, with the larger errors corresponding to intermediate-depth events. A consequence of these errors was an underestimation of the



velocity anomaly. Second, Fermat's principle applies to curves with the same end points, but in practice the end point of  $R(u_1)$  corresponding to the event location is not the true location because it has been determined with an incorrect velocity model. Therefore,  $R(u_1)$  is not necessarily close to the true ray path. In practice, however, the approximations involved in the use of Fermat's principle in local tomography improve as the iterations proceed, and can essentially be ignored.

For completeness, the relation between small variations  $dv$  and  $du$  in velocity and slowness will be considered. Because travel time is equal to distance divided by velocity, here we are interested in  $1/(v + dv)$ , with  $dv \ll v$ , which can be approximated as follows

$$\frac{1}{v + dv} = \frac{1}{v} \left(1 + \frac{dv}{v}\right)^{-1} \approx \frac{1}{v} - \frac{1}{v^2} dv \equiv u + du. \quad (41)$$

Therefore, it does not make any theoretical difference whether the inverse problem is formulated in terms of slowness or velocity as long as  $du$  and  $dv$  are such that the approximations introduced are valid.

Now let us go back to the joint hypocentral location and determination of a 3-D slowness model. The only difference with the expression for velocity tomography (Eq. (38)) is in the last term, which is replaced by the right-hand side of Eq. (40). This gives

$$w_{ij}r_{ij} = w_{ij} \left( dT_j + \frac{\partial t}{\partial x} dx_j + \frac{\partial t}{\partial y} dy_j + \frac{\partial t}{\partial z} dz_j + \int_{R_{ij}} du ds \right) \quad (42)$$

$$i = 1, \dots, N_j \quad j = 1, \dots, M$$

where  $R_{ij}$  indicates the raypath between the corresponding station-event pair. To proceed further the integral must be discretized, which can be done using a block model. After that Eq. (42) becomes

$$w_{ij}r_{ij} = w_{ij} \left( dT_j + \frac{\partial t}{\partial x} dx_j + \frac{\partial t}{\partial y} dy_j + \frac{\partial t}{\partial z} dz_j + \sum_k l_{ijk} du_k \right) \quad (43)$$

$$i = 1, \dots, N_j \quad j = 1, \dots, M$$

where  $l_{ijk}$  and  $du_k$  are the length of the ray in the  $k$ th block and the corresponding slowness perturbation.

Note that although the last terms in Eqs. (38) and (43) are different, in principle, the velocity and slowness models derived using the two formulations should satisfy

$u(\mathbf{x}) = 1/v(\mathbf{x})$ . Let us consider this question for the block parameterization. To compute the derivative of travel time with respect to velocity we will use the following approximation

$$\frac{\partial t_{ij}}{\partial v_k} \approx \frac{\partial t_{ijk}}{\partial v_k} \quad (44)$$

[85], where the subscripts in  $t$  are used to identify the rays and blocks involved. The meaning of this relation is that the change in travel time along a ray path due to a change in the velocity of a given block is approximately equal to the change in travel time in that block. Note that

$$t_{ijk} = \frac{l_{ijk}}{v_k} \quad (45)$$

which allows us to write

$$\frac{\partial t_{ij}}{\partial v_k} dv_k \approx l_{ijk} \left( -\frac{1}{v_k^2} dv_k \right) \equiv l_{ijk} du_k. \quad (46)$$

Using this result we see that Eqs. (38) and (43) are identical, which means that the two formulations should give the same results as long as the block structure is the same in the two cases and the approximations introduced are valid. Of course, the numerical implementation of the software corresponding to the two approaches should be the same.

### Decoupling of the Earthquake Location and Tomography Problems

Equation (38) can be written in matrix form as:

$$\mathbf{W}_j \mathbf{A}_j dx_j + \mathbf{W}_j \mathbf{B}_j dv = \mathbf{W}_j \mathbf{r}_j; \quad j = 1, M \quad (47)$$

where  $\mathbf{r}_j$  is the vector of residuals  $r_{ij}$ ,  $\mathbf{A}_j$  is an  $N_j \times 4$  matrix of partial derivatives of time with respect to origin time and hypocentral coordinates,  $dx_j$  is the vector of origin time and hypocenter adjustments,  $\mathbf{W}_j$  is an  $N_j \times N_j$  matrix of weights,  $\mathbf{B}_j$  is an  $N_j \times K$  matrix of partial derivatives of travel time with respect to velocities, and  $dv$  is a vector of  $K$  velocity adjustments. The matrix  $\mathbf{B}_j$  has zero entries for the blocks not traversed by any ray. The matrix  $\mathbf{W}_j$  has only one nonzero entry per row and column and is not necessarily diagonal because the ordering of stations for different earthquakes may not be always the same.

For Eq. (43) the matrix form is

$$\mathbf{W}_j \mathbf{A}_j dx_j + \mathbf{W}_j \mathbf{C}_j du = \mathbf{W}_j \mathbf{r}_j; \quad j = 1, M \quad (48)$$

with  $\mathbf{C}_j$  an  $N_j \times K$  matrix whose entries are raypath lengths in individual blocks and  $du$  is a vector of  $K$  slowness adjustments. All the other quantities are as in Eq. (47).

Clearly, Eqs. (47) and (48) are formally equivalent, which means that for the following discussion on how to solve them it is unnecessary to distinguish between velocity and slowness. For this reason, from now on we will refer exclusively to Eq. (47), with the understanding that the results below apply to Eq. (48) as well.

Equation (47) represents  $M$  systems of equations coupled through the common vector  $dv$ . Therefore, these systems could be combined, in principle, into a single system, but because its size could become computationally unmanageable, it is necessary to decouple the earthquake location part from the inversion part. A very efficient approach to do that is based on the singular value decomposition (SVD) of a matrix [110] and is known as the method of parameter separation. The following presentation is based on [119]. The SVD of an arbitrary  $n \times m$  matrix  $G$  is given by

$$G = U\Lambda V^T \tag{49}$$

(see, e. g., [3,46]) where the superscript T indicates matrix transposition,  $U$  and  $V$  are  $n \times n$  and  $m \times m$  matrices with columns given by the eigenvectors of  $GG^T$  and  $G^TG$ , respectively, and  $\Lambda$  is an  $n \times m$  matrix with diagonal elements equal to the singular values of  $G$  and off-diagonal elements equal to zero. The singular values are positive and equal to the square roots of the eigenvalues of  $GG^T$  and  $G^TG$ . The number of nonzero singular values, say  $p$ , cannot exceed the minimum of  $m$  and  $n$ . Matrices  $U$  and  $V$  are orthogonal, i. e.,

$$UU^T = I_n, \quad VV^T = I_m, \tag{50}$$

where  $I_n$  is the  $n \times n$  identity matrix. Let us assume that the singular values are sorted in nonincreasing order (i. e., largest first). Then, matrices  $U$  and  $V$  can be partitioned as follows:

$$U = (U_p \quad U_0), \quad V = (V_p \quad V_0) \tag{51}$$

where the subscripts  $p$  and  $0$  indicate that the columns of the matrices come from the eigenvectors corresponding to the nonzero and zero singular values, respectively. Partitioned matrices are discussed in, e. g., [103]. Matrix  $A$  can be partitioned in a similar way. Then, writing Eq. (49) in terms of these partitioned matrices gives

$$G = (U_p \quad U_0) \begin{pmatrix} A_p & O \\ O & O \end{pmatrix} \begin{pmatrix} V_p^T \\ V_0^T \end{pmatrix} = U_p A_p V_p^T \tag{52}$$

where  $A_p$  is the  $p \times p$  diagonal matrix having as elements the  $p$  nonzero singular values and  $O$  represents zero matrices of appropriate sizes. Using this result and

$$U_0^T U_p = O \tag{53}$$

which is a consequence of the fact that the columns of  $U$  are orthogonal to each other, we get the following important result

$$U_0^T G = U_0^T U_p A_p V_p^T = O. \tag{54}$$

To apply Eq. (54) to our problem we will use

$$G = W_j A_j = U_j A_j V_j^T \\ = (U_{jp} \quad U_{j0}) A_j (V_{jp} \quad V_{j0})^T = U_{jp} A_{jp} V_{jp}^T. \tag{55}$$

Because  $W_j A_j$  has four columns and at least as many rows,  $p \leq 4$ . If  $p < 4$ , this means that one or more of the columns of the matrix is a linear combination of the others and the location of the corresponding event cannot be determined uniquely. Events with  $p < 4$  should not be used in the inversion. Now multiply Eq. (47) by  $U_{j0}^T$  on the left and use the equivalent of Eq. (54) for the  $j$ th event. This gives

$$U_{j0}^T W_j B_j dv = U_{j0}^T W_j r_j; \quad j = 1, M. \tag{56}$$

Because  $A_j$  does not appear in Eq. (56), we have been able to decouple the velocity parameters from the earthquake parameters. Note, however, that this equation depends on  $A_j$  implicitly via  $U_{j0}$ , which means that location errors will translate into errors in  $U_{j0}$  and, thus, in  $dv$ .

To simplify the following discussion Eq. (56) will be rewritten as

$$\underbrace{B'_j}_{(N_j-4) \times K} \underbrace{dv}_{K \times 1} = \underbrace{r'_j}_{(N_j-4) \times 1}; \quad j = 1, M \tag{57}$$

with obvious expressions for the primed quantities. The sizes of the matrix and vectors involved are also indicated (assuming  $p = 4$ ). Equation (57) represents  $M$  systems of equations with the common unknown vector  $dv$  and can be written in compact form in terms of a partitioned matrix and vector

$$B dv = \rho \tag{58}$$

where

$$B = (B_1^T \ B_2^T \ \dots \ B_M^T)^T \tag{59}$$

and

$$\rho = (r_1^T \ r_2^T \ \dots \ r_M^T)^T. \tag{60}$$

Let  $N$  be the largest of the  $N_j$ . Then, matrix  $B$  has  $K$  columns and up to  $M \times (N - 4)$  rows, which means

that it may become very large. For example, there may be thousands of earthquakes, thousands of velocity parameters, and tens of stations, which means that how to solve Eq. (60) becomes an issue. If the size of the matrix and the computer resources allow it, a solution based on the use of the SVD would be convenient (see Sect. “**Solution of Ill-Posed Linear Problems**”). If this approach is not feasible, one can solve Eq. (60) by least squares, which requires solving

$$\mathbf{B}^T \mathbf{B} \, d\mathbf{v} = \mathbf{B}^T \boldsymbol{\rho} \quad (61)$$

which on account of Eqs. (59) and (60) becomes

$$\left( \sum_{j=1}^M \mathbf{B}'_j{}^T \mathbf{B}'_j \right) d\mathbf{v} = \sum_{j=1}^M \mathbf{B}'_j{}^T \mathbf{r}_j. \quad (62)$$

This approach however, may have two problems. One is that the resulting matrix may still be too large for the computer facilities available and the other is the possibility of numerical loss of precision due to the matrix multiplications. This second problem can be alleviated through the use of double precision, although at the expense of longer computer times. For these reasons, for large tomographic problems the ensuing linear systems are solved using iterative matrix solvers (see Subsect. “**Regularization Approach**”). Also note that Eq. (62) is not the result of the “accumulation” of individual equations  $\mathbf{B}'_j{}^T \mathbf{B}'_j d\mathbf{v} = \mathbf{B}'_j{}^T \mathbf{r}_j$ , as it is sometimes stated.

Once Eq. (58) has been solved, the earthquakes must be relocated. To do that Eq. (47) will be used, but to make the analysis more general it will be assumed that both  $P$  and  $S$  wave arrivals are available. Because the  $P$  wave arrivals are independent of the  $S$  wave velocity, and vice versa, we can write two pairs of equations similar to Eqs. (47) and (56), and to solve for  $d\mathbf{v}^P$  and  $d\mathbf{v}^S$ , where the superscripts identify the type of arrivals. After that is done we can write

$$\mathbf{W}_j^P \mathbf{A}_j^P d\mathbf{x}_j = \mathbf{W}_j^P \mathbf{r}_j^P - \mathbf{W}_j^P \mathbf{B}_j^P d\mathbf{v}^P; \quad j = 1, M \quad (63)$$

and

$$\mathbf{W}_j^S \mathbf{A}_j^S d\mathbf{x}_j = \mathbf{W}_j^S \mathbf{r}_j^S - \mathbf{W}_j^S \mathbf{B}_j^S d\mathbf{v}^S; \quad j = 1, M. \quad (64)$$

These two equations are coupled through the common vector  $d\mathbf{x}_j$  and can be written as a single equation as follows

$$\begin{pmatrix} \mathbf{W}_j^P \mathbf{A}_j^P \\ \mathbf{W}_j^S \mathbf{A}_j^S \end{pmatrix} d\mathbf{x}_j = \begin{pmatrix} \mathbf{W}_j^P \mathbf{r}_j^P - \mathbf{W}_j^P \mathbf{B}_j^P d\mathbf{v}^P \\ \mathbf{W}_j^S \mathbf{r}_j^S - \mathbf{W}_j^S \mathbf{B}_j^S d\mathbf{v}^S \end{pmatrix}; \quad j = 1, M. \quad (65)$$

[122]. The matrix on the left-hand side of this equation is small ( $2N \times 4$  at most) and can be solved using any of the standard methods. Once each  $d\mathbf{x}_j$  has been found, it is used to get new initial estimates of origin time and hypocentral estimates (equal to  $T_j + dT_j$ ,  $x_j + dx_j$  and so on). After updating the velocities (i. e., they become  $v_k + dv_k$ ) a new iteration is started. This iterative process is repeated until some stopping criterion is met. An early criterion [144] was based on the use the statistical  $F$  test. With this test it is possible to determine if the decrease in the sum of arrival time residuals squared from a given iteration to the next is statistically significant ([80] and references therein). A simpler, yet effective approach is to stop when the root-mean square of all the travel-time residuals reaches the expected value of the error in the data.

### Computation of Local Travel Times

The iterative determination of a 3-D velocity model and simultaneous determination of earthquake locations requires the availability of software for the computation of travel times in that type of models. The theoretical solution to this problem using ray theory was well known (see, e. g., [22,44,85,112], and references therein) at the time the seismic tomographic method began to be developed, but because exact ray tracing is a computationally time-consuming task, in most of the earlier tomographic codes approximate ray tracing methods were used. One of them was introduced by Thurber [144], and was based on the use of arcs of circles to approximate the ray paths connecting the source and station. Although the method was fast, its accuracy was questionable. For example, in [42] the weights of arrivals with epicentral distances between 20 and 45 km were decreased and arrivals with distances larger than 45 km were not used. Because this approximate method was popular, the readers of the earlier literature should be aware of the limitations of the method, which can result in errors in the computed velocities and in the locations determined with them.

A significant improvement to Thurber’s [144] method was introduced by Um and Thurber [152], who developed a method based on the perturbation of an initial raypath between the source and the station using a ray-theoretical equation. Ray tracing methods that find the raypath between two points by iterative perturbation of an initial estimate are known as *bending methods* (see, e. g., [75] and references therein). The Um and Thurber method is based on an approximate computation of the perturbations, and for this reason it is commonly referred to as a *pseudo-bending method*. However, one of the disadvantages of this method, noted by the authors, is that it is not appropriate

for use in media with velocity discontinuities and where nearly constant velocities are present. This method is included in a popular tomographic package, and these limitations should be taken into account when considering the results obtained using it. Pseudo-bending solutions that do not have the limitations of the Um and Thurber method exist [99,115], but their applications to seismic tomography appear to be limited. The Um and Thurber's method was used in combination with Snell's law to account for velocity discontinuities at the Conrad and Moho boundaries and the upper boundary of subducting plates [163], but these boundaries must be known from other studies.

An early example of tomography with exact ray tracing can be found in [61]. In this case the ray tracing problem was solved by Runge–Kutta integration of six simultaneous differential equations that can be derived from the eikonal equation. This approach corresponds to the *shooting method*, which requires the specification of the two angles at the source that define a ray. These two angles must be changed until the end point is within a specified distance from the station. This approach does not introduce any approximations (beyond those inherent to ray theory), but, as already noted, is more time consuming than the approximate methods. The software based on this approach was later improved [14] by incorporation of the method of parameter separation described in Subsect. “**Decoupling of the Earthquake Location and Tomography Problems**”. This software was applied to the 1989,  $M = 7.1$ , Loma Prieta, California, mainshock-aftershock sequence [116]. Another example of tomography with exact ray tracing [79] is based on the division of the earth into constant-velocity blocks and ray tracing based on the use of Snell's law.

A different kind of approach to the computation of travel times, which does not require ray tracing, was introduced by Vidale [157,158]. In his method, the velocity model consists of a set of values assigned to points in a regular equispaced 2-D or 3-D grid and travel times are computed using a finite-difference approximation to the partial derivatives of travel times with respect to the spatial coordinates and plane or circular wavefront approximations. It must be noted, however, that the eikonal equation is not solved using the classical finite-difference method, which is not a simple task and is seldomly used [21]. The output of Vidale's method is a series of wavefronts of minimum travel times. This method is more time consuming than the approximate methods referred to above and has some problems when large velocity contrasts are present. Moreover, the approximations involved may not be adequate in the vicinity of the source, where the wavefronts are highly curved [21]. Vidale's method has been imple-

mented in software for earthquake location in media with 3-D variations [102].

Vidale's [157,158] approach motivated a method developed by Podvin and Lecomte [113], who used Huygens's principle for the computation of travel times. The application of the principle is equivalent to the propagation of local wavefronts, which is done as follows. Let us assume that the four vertices of one of the sides of a unit cube in the grid have known arrival times. These points can be combined into four groups of three adjacent points. Each group defines a plane wavefront, which is used to compute the arrival time at each of the other four corners of the same cube. The selection of the appropriate local wavefront follows a set of pre-established rules. In addition, each point within the grid is the common vertex of eight unit cubes around it. These cubes form a larger cube with the grid point at its center and each side contributes sixteen local wavefronts. Therefore, for each grid point within the grid, ninety-six wavefronts must be considered. The advantage of this approach is that it accounts for the existence of transmitted, diffracted, and head waves and performs well even in the presence of large velocity contrasts. The computations can be carried out in parallel in a multiprocessor machine or sequentially when only one processor is used. In the latter case the method becomes computationally very time consuming, but its implementation in a tomographic software package [15] has resulted in highly detailed velocity models. Examples are given in Sect. “**Examples**”. Podvin and Lecomte's software is becoming very popular and for this reason it is worth noting that the software has a minor flaw that results in time differences for rays moving equal distances to the left and right of the source [151]. For example, for a 0.5 km grid size and source-receiver offsets up to 30 km the difference is between about 0.03 and 0.04 s, with the larger values for offsets less than 10 km. Decreasing the grid size reduces the error.

The two wavefront methods referred to above do not compute raypaths directly. However, in seismic tomography they are needed to associate arrival time residuals with specific velocity grid points. In Podvin and Lecomte's software, rays are traced from the receiver to the source following a direction opposite to the time gradient, with the endpoint of the ray at most a distance  $h$  from the source, where  $h$  is the grid spacing. Recall that in isotropic media rays are defined as curves whose tangents are everywhere perpendicular to a wavefront (see, e. g., [69,120]).

There is yet another approach to the computation of travel times, namely, the shortest-path method, which is based on concepts borrowed from network theory. In this approach the velocity model is also based on a grid

of points and connections are established between close points. Each connection is given a weight, equal to the travel time between them. The shortest path between any two points is that along the connections for which the sum of the weights is smallest. This path is an approximation to the seismic raypath. This method was introduced in seismology in [101] and was further developed in [98]. Additional references can be found in, e. g., [9,21], and [147]. Applications to seismic tomography and 3-D earthquake location can be found in [7,8] and [100], respectively.

To end this section we note that the methods considered here assume a Cartesian coordinate system, so that the curvature of the earth is ignored, i. e., the earth is assumed to be flat. This assumption, however, has limitations (see, e. g., [125,134]), which should be taken into account when dealing with epicentral distances that exceed about 150–200 km. This range is provided as guide only. For a given epicentral distance, the error in travel time generally increases with hypocentral depth. For example, for a depth of 200 km and an epicentral distance of 2° the error is about 0.25 s [125]. When the errors introduced by ignoring the earth's curvature become important (e. g., larger than the arrival-time picking errors) it is advisable to introduce a cut-off epicentral distance beyond which the arrival times are either not used or weighted down (see, e. g. [78] and Subsect. “*P and S Wave Velocity Models for Taiwan*”).

### Teleseismic Tomography

The data most commonly used in teleseismic tomographic studies are *P* wave arrivals for the investigation of the mantle and *PcP* and *PKP* arrivals for the investigation of the core-mantle boundary (see, e. g. [96]), although other arrivals have also been used. A major source for the data is the International Seismological Center, which has collected millions of arrivals contributed by hundreds of stations around the world. Although it is possible to perform a simultaneous event location and velocity determination following a formulation similar to that described for local tomography (see, e. g. [137]), the most common approach is to keep the event locations fixed. This is the approach to be describe here.

The use of teleseismic data to determine a 3-D velocity model of the crust and upper mantle underneath a seismic array was introduced by Aki et al. [4]. They divided the portion of the earth being investigated into homogeneous horizontal layers subdivided into blocks, computed the theoretical arrival times for earthquakes having published locations, formed arrival time residuals, and solved for slowness perturbations in each of the blocks with re-

spect to a model with constant slowness in each layer. A simplified version of the arguments presented in [4] follows. Let  $u_k$  and  $u_k^o$  be the actual and reference slowness of block  $k$  and let  $m_k$  be the fractional slowness perturbation, equal to

$$m_k = \frac{u_k - u_k^o}{u_k^o} \equiv \frac{du_k}{u_k^o} \quad (66)$$

where  $du_k$  is equivalent to the  $du$  in Eq. (40) and is assumed to be much smaller than  $u_k$ . Solving for  $u_k$  we get

$$u_k = u_k^o(1 + m_k). \quad (67)$$

Given a particular earthquake-station pair, the actual travel time in the  $k$ th block traveled by the corresponding ray is given by

$$t_k = l_k u_k = l_k u_k^o(1 + m_k) \approx t_k^o + t_k^o m_k \quad (68)$$

where  $l_k$  is the actual length of the ray in the  $k$ th block and the approximate sign arises because the actual path and the path in the reference medium are assumed to be close on account of Fermat's principle. Then, the difference in travel time introduced by the slowness perturbation  $m_k$  is given by

$$dt_k = t_k - t_k^o \approx t_k^o m_k. \quad (69)$$

Introducing the right-hand side of Eq. (66) in this expression gives

$$t_k \approx \frac{t_k^o}{u_k^o} du_k = l_k^o du_k \quad (70)$$

where  $l_k^o$  is the raypath length in the  $k$ th block for the reference model. The right-hand side of this expression is equivalent to  $l_{ijk} du_k$  in Eq. (43).

Let us consider the arrival time residual that arises when using teleseismic data and the earthquake locations used are those determined with the reference slowness model. Using the notation introduced earlier we can write

$$r_{ij} = \delta t_j + \sum_k l_{ijk}^o du_k \quad (71)$$

where the subscripts  $i$  and  $j$  denote station and event, and the sum is over all the blocks. A block not traversed by any ray will have the corresponding  $l_{ijk}^o$  equal to zero. The term  $\delta t_j$  includes the contributions to the residuals not accounted for by the second term. For example,  $\delta t_j$  includes errors in event origin time and location, picking errors, possible errors in the reference slowness model outside of the volume investigated, and possible errors in the model

parameterization. This  $\delta t_j$  is unknown, but because it is assumed to be effectively constant for a given earthquake, it can be eliminated by averaging  $r_{ij}$  over all the stations that recorded the  $j$ th event and subtracting the result from Eq. (71). The average residual is given by

$$\overline{r_{ij}} = \frac{1}{N_j} \sum_{i=1}^{N_j} r_{ij} = \delta t_j + \sum_k \overline{l_{ijk}^o} du_k \quad (72)$$

where the overbar indicates average, as defined by the first equality. Next, subtracting Eq. (72) from Eq. (71) gives the relative residual

$$r'_{ij} = r_{ij} - \overline{r_{ij}} = \sum_k \left( l_{ijk}^o - \overline{l_{ijk}^o} \right) du_k. \quad (73)$$

A similar result can be found in [3]. This equation shows that the slowness perturbations  $du_k$  are obtained by solving a linear system, as expected, and that the main difference with Eq. (43) is the presence of source terms in the latter (aside from a weight factor). Once Eq. (73) is solved, the perturbations are added to the reference slownesses and the tomographic problem is solved. It is important to note however, that the resulting model is not well resolved vertically. In fact, the effect of a uniform perturbation over a layer cannot be distinguished from a change in  $\delta t_j$ . This problem is intrinsic to the method and cannot be removed [3]. Examples showing the capabilities and limitations of the method can be found in [45]. A potentially severe limitation is the effect introduced by the presence of large sedimentary basins, which usually have significantly smaller velocities than the surrounding rocks. These basins may introduce travel time anomalies as large as the upper mantle residuals [93]. Clearly, the tomographic results will be incorrect if the arrival times are not corrected for large crustal effects, which requires information derived independently (see, e. g., [93,131]).

### Solution of Ill-Posed Linear Problems

Regardless of the type of arrivals considered, the tomographic problem reduces to the solution of equations of the form

$$\mathbf{Ax} = \mathbf{b} \quad (74)$$

where  $\mathbf{A}$  is a known  $m \times n$  matrix with information about the model,  $\mathbf{x}$  is an  $n$  vector of unknowns, and  $\mathbf{b}$  is a known  $m$  vector. Solving this problem requires consideration of the question of the existence of a solution, but before proceeding it is convenient to introduce the following definition, due to the French mathematician Hadamard

(see, e. g., [47,53], and references therein). The problem represented by Eq. (74) is said to be *well posed* if it has a solution, it is unique, and depends continuously on  $\mathbf{b}$ . The latest condition means that small changes in  $\mathbf{b}$  cause small changes in  $\mathbf{x}$ . When any of these conditions is not satisfied the problem is said to be *ill posed*. For a unique solution of Eq. (74) to exist a necessary (but not sufficient) condition is that  $m \geq n$ . If  $m > n$ , the system is said to be *overdetermined*. It must be noted, however that this condition (to be assumed here) does not imply that a solution exists, or that it is unique (if it exists). For example, if some of the rows of  $\mathbf{A}$  constitute a linearly dependent set (e. g., some of them are linear combinations of other rows) a solution may exist but it will not be unique. On the other hand, a solution may not exist at all. In this case the system is said to be *inconsistent*, a condition that in seismic tomography arises because of errors in both the data and the model. The magnitude of the data errors depend on arrival type. Teleseismic data generally have larger errors than local data, while local S wave arrivals have larger errors than the corresponding P arrivals because they are more difficult to identify. Model errors arise because the parameterization of the velocity (or slowness) field does not reproduce the actual variations appropriately. Although a more faithful representation of the earth would be advantageous, an obvious problem is that the data may not be adequate to resolve the parameters of a more detailed model. Clearly, if the block size or grid spacing used to parameterize the velocity (or slowness) field are too small with respect to the interstation spacing, and the spatial distribution of the seismic sources is not favorable, then a large number of blocks may not be traversed by any ray, and matrix  $\mathbf{A}$  will be very large with a large number of zero entries (i. e., the matrix is said to be *sparse*). For a given set of arrival times, an increase in the number of unknown parameters results in a relative decrease in the number of constraints, and in the resolution of the solution (see below).

A very powerful tool for the solution of the system (74) is provided by the Moore–Penrose generalized inverse  $\mathbf{A}^\dagger$  of  $\mathbf{A}$ . Let  $\mathbf{A}$  have  $p$  nonzero singular values, sorted by decreasing order. Then, using the SVD (see Eq. (49))

$$\mathbf{A}^\dagger = \mathbf{VA}^\dagger \mathbf{U}^T = \mathbf{V}_p \mathbf{A}_p^{-1} \mathbf{U}_p^T \quad (75)$$

where  $(\mathbf{A}^\dagger)_{ii}$  is equal to  $1/\lambda_i$  if  $\lambda_i \neq 0$  and zero otherwise [3,111]. The second equality follows from the product of partitioned matrices. The generalized inverse solution is given by

$$\mathbf{x}^\dagger = \mathbf{A}^\dagger \mathbf{b} \quad (76)$$

which has the property that is a minimum-length solution (see, e. g., [3]). A more general solution is of the form

$$\mathbf{x} = \mathbf{x}^\dagger + \mathbf{z} \tag{77}$$

where

$$\mathbf{A}\mathbf{z} = \mathbf{0}. \tag{78}$$

The vectors  $\mathbf{z}$  that satisfy Eq. (78) constitute the *null space* of  $\mathbf{A}$ . The column vectors of  $\mathbf{A}$  corresponding to the zero singular value satisfy an equation similar to Eq. (78) and constitute a base for the null space of  $\mathbf{A}$ . Therefore,  $\mathbf{z}$  is a linear combination of the columns of the matrix  $\mathbf{V}_o$ .

The definition of  $\mathbf{A}^\dagger$  is based on a sharp distinction between zero and nonzero singular values. In practice, however, this clear-cut situation does not occur, with some of them nonzero yet much smaller than the largest one. In this case the solution of Eq. (76) may be strongly affected by errors in the data. This question is made more precise when the condition number,  $\kappa$ , of  $\mathbf{A}$  is introduced

$$\kappa = \frac{\lambda_{\text{largest}}}{\lambda_{\text{smallest}}} \geq 1. \tag{79}$$

Then, a perturbation  $d\mathbf{b}$  (such as errors) in the data introduces a perturbation  $d\mathbf{x}$  in the solution that satisfies

$$\frac{|d\mathbf{x}|}{|\mathbf{x} + d\mathbf{x}|} \leq \kappa \frac{|d\mathbf{b}|}{|\mathbf{b} + d\mathbf{b}|} \tag{80}$$

(e. g., [46,49]). Therefore, when  $\kappa$  is large a small change in the data may cause a significant change in the solution. If  $\kappa$  is large the matrix is said to be *ill-conditioned*; otherwise it is said to be *well-conditioned*. Although the terms “large” and “small” are obviously vague, the idea is not. For example, for a given data set and two velocity models leading to matrices  $\mathbf{A}$  having condition numbers that differ by a factor of say 10, the matrix with the larger  $\kappa$  may lead to a solution more affected by the errors in the data.

In the following, ill-conditioned problems will be considered a form of ill-posed problems and the distinction between the two will no longer be made. The question that must be addressed now is how to solve this type of problems. Two approaches are available, one based on the so-called regularization of the problem and the other based on the Bayesian statistics. The two approaches are discussed in detail below.

### Regularization Approach

As noted above, tomographic problems are likely to be ill-posed, and to get a solution it is necessary to recourse to mathematical techniques that turn them well-posed. This process is known as regularization (see,

e. g., [53,55,80,148,164]). The reader must be aware, however, that there is a price to be paid; namely one ends up with a family of solutions and to objectively select the most appropriate among them may not be easy or even possible. For a system with a matrix for which the computation of its singular value decomposition is practically feasible, the generalized inverse solution has several advantages, but for most tomographic problems the size of the matrices involved is so large that this option is not practical and will not be pursued here. Instead, we will discuss techniques that lead to systems that can be solved with computationally efficient methods.

Before proceeding we will introduce the following definitions. A square symmetric matrix  $\mathbf{C}$  is said to be *positive semidefinite* if

$$\mathbf{y}^T \mathbf{C} \mathbf{y} \geq 0; \quad \mathbf{y} \neq \mathbf{0}. \tag{81}$$

The matrix  $\mathbf{C}$  is *positive definite* if the  $\geq$  sign in Eq. (81) is replaced by  $>$ .

Let us summarize some of the properties of these matrices (see, e. g., [121]).

- (1) Let  $\mathbf{v}_i$  be an eigenvector of  $\mathbf{C}$  and  $\lambda_i$  its corresponding eigenvalue. Then

$$\mathbf{v}_i^T \mathbf{C} \mathbf{v}_i = \lambda_i \mathbf{v}_i^T \mathbf{v}_i = \lambda_i |\mathbf{v}_i|^2. \tag{82}$$

Because  $|\mathbf{v}_i| > 0$ ,  $\lambda_i \geq 0$  if  $\mathbf{C}$  is positive semidefinite, and  $\lambda_i > 0$  if  $\mathbf{C}$  is positive definite. In the second case the inverse of  $\mathbf{C}$  exists because  $\mathbf{C} = \mathbf{U}\mathbf{A}\mathbf{U}^T$ , with  $\mathbf{U}$  and  $\mathbf{A}$  similar to those in Eq. (49) and Eq. (50) (see, e. g., [103]), and  $\mathbf{C}^{-1} = \mathbf{U}\mathbf{A}^{-1}\mathbf{U}^T$ .

- (2) Any matrix of the form  $\mathbf{B}^T \mathbf{B}$  is either positive definite or semidefinite. Consider

$$\mathbf{y}^T (\mathbf{B}^T \mathbf{B}) \mathbf{y} = (\mathbf{B} \mathbf{y})^T \mathbf{B} \mathbf{y} = |\mathbf{B} \mathbf{y}|^2 \geq 0; \quad \mathbf{y} \neq \mathbf{0}. \tag{83}$$

Therefore,  $\mathbf{B}^T \mathbf{B}$  is at least positive semidefinite. In addition, if the inverse of this matrix exists, all of its eigenvalues will be positive and  $\mathbf{B}^T \mathbf{B}$  will be positive definite. If the inverse does not exist, the matrix will be positive semidefinite.

- (3) Let  $\mathbf{C}$  be a diagonal matrix with positive diagonal elements. Then  $\mathbf{C}$  is positive definite because

$$\mathbf{y}^T \mathbf{C} \mathbf{y} = \sum_i c_i y_i^2 > 0; \quad \mathbf{y} \neq \mathbf{0}, \quad c_i = (\mathbf{C})_{ii} > 0. \tag{84}$$

- (4) Let matrices  $\mathbf{B}$  and  $\mathbf{P}$  be positive semidefinite and definite, respectively, and let  $\lambda^2$  be a scalar. Then  $\mathbf{B} + \lambda^2 \mathbf{P}$  is positive definite because

$$\mathbf{y}^T (\mathbf{B} + \lambda^2 \mathbf{P}) \mathbf{y} = \mathbf{y}^T \mathbf{B} \mathbf{y} + \lambda^2 \mathbf{y}^T \mathbf{P} \mathbf{y} > 0; \quad \mathbf{y} \neq \mathbf{0}. \tag{85}$$

These results will be used to show that the typical methods of solution of Eq. (74) when  $\mathbf{A}^{-1}$  does not exist or is ill-conditioned are based on the solution of a new well-posed problem.

The simplest regularization approach is to constrain (in some sense) the length of the solution vector  $\mathbf{x}$ . Let us consider the minimization of the function

$$\bar{S}(\mathbf{x}) = w|\mathbf{Ax} - \mathbf{b}|^2 + \mathbf{x}^T \mathbf{D} \mathbf{x} \equiv wS(\mathbf{x}) + Q(\mathbf{x}) \quad (86)$$

with respect to  $\mathbf{x}$ . Here  $w$  is a positive weighting factor,  $\mathbf{D}$  is a symmetric positive definite matrix, and  $S$  and  $Q$  are defined by the identity. This problem was solved by Levenberg [87] and a similar one by Marquardt (who used  $\mathbf{D} = \mathbf{I}$ ) [92] in the context of non-linear problems. The minimization of  $\bar{S}$  leads to the following equation

$$(\mathbf{A}^T \mathbf{A} + \lambda^2 \mathbf{D}) \mathbf{x} = \mathbf{A}^T \mathbf{b}; \quad \lambda^2 = \frac{1}{w} \quad (87)$$

(see, e.g., [121]). A similar equation was introduced by Tihonov [148], who investigated the regularization of ill-posed linear problems in the context of integral equations. The  $\mathbf{x}$  that solves Eq. (87) is known as the *damped least-squares estimator*. When  $\lambda^2 = 0$ , Eq. (87) corresponds to the *ordinary* least-squares solution.

Note that Eq. (87) is a special case of Eq. (85) with  $\mathbf{B} = \mathbf{A}^T \mathbf{A}$  and  $\mathbf{P} = \mathbf{D}$ , which means that there is a value of  $\lambda^2$  that makes the matrix on the left side of Eq. (87) well conditioned. Therefore, for a given  $\mathbf{D}$ , Eq. (87) will have a family of solutions, which will depend on  $\lambda^2$ . A problem with this equation, however, is that it may be too large for tomographic problems. For this reason, Eq. (87) will be derived by consideration of the following system

$$\begin{pmatrix} \mathbf{A} \\ \lambda \mathbf{D}^{1/2} \end{pmatrix} \mathbf{x} = \begin{pmatrix} \mathbf{b} \\ \mathbf{0} \end{pmatrix}. \quad (88)$$

This new system will be solved using ordinary least squares, which requires solving

$$(\mathbf{A}^T \quad \lambda \mathbf{D}^{1/2}) \begin{pmatrix} \mathbf{A} \\ \lambda \mathbf{D}^{1/2} \end{pmatrix} \mathbf{x} = (\mathbf{A}^T \quad \lambda \mathbf{D}^{1/2}) \begin{pmatrix} \mathbf{b} \\ \mathbf{0} \end{pmatrix} \quad (89)$$

where  $\mathbf{D}^{1/2}$  is known as the square root of  $\mathbf{D}$  (see, e.g., [121]). After performing the matrix multiplications we obtain Eq. (87), which shows that the minimization of  $\bar{S}(\mathbf{x})$  is equivalent to solving Eq. (74) under the constraint

$$\lambda \mathbf{D}^{1/2} \mathbf{x} = \mathbf{0}. \quad (90)$$

Clearly, the  $\lambda$  in this equation is unnecessary, but it is introduced to indicate that it is a weighting factor. We also note that in many applications  $\mathbf{D} = \mathbf{I}$ .

Let us mention some of the properties of the damped least-squares solution in the context of nonlinear linearized inverse problems. This type of problems require an initial estimate of the solution, say  $\mathbf{x}_0$ . If this estimate is poorly chosen, it may lead to a non-convergent iterative process, and for this reason Levenberg [87] decided to minimize  $\bar{S}$ . Moreover, when dealing with linearized problems,  $S$  is the linearized version of the true residuals. Let  $s$  represent the sum of residuals whose linearization leads to  $S$ . A question investigated by Levenberg [87] regards the relations between  $\bar{S}$ ,  $S$  and  $s$ , and between the damped and ordinary least-squares solutions, indicated by  $\mathbf{x}_\lambda$  and  $\mathbf{x}^{ls}$ , respectively. Some of his results are as follows:

$$S(\mathbf{x}_\lambda) < S(\mathbf{x}_0) \quad (91)$$

$$Q(\mathbf{x}_\lambda) < Q(\mathbf{x}^{ls}) \quad (92)$$

and

$$s(\mathbf{x}_\lambda) < s(\mathbf{x}_0) \quad \text{for some } \lambda^2 \quad (93)$$

as long as  $\mathbf{x}_0$  is not a stationary point of  $s$  [87,121]. Equations (91) and (92) show that it is possible to simultaneously minimize  $S$  and  $Q$  for all  $\lambda^2$ , while Eq. (93) shows that there are values of  $\lambda^2$  that minimize  $s$ , which is the quantity of actual interest to us. Although none of these results is obvious, they are implicitly assumed when solving linearized problems. In addition, if  $\lambda^2$  is large enough, Eq. (87) becomes

$$\mathbf{x} \approx \frac{1}{\lambda^2} \mathbf{D}^{-1} \mathbf{A}^T \mathbf{b} \quad (94)$$

with  $\mathbf{x}$  going to zero as  $\lambda^2$  goes to infinity. Therefore, when solving linearized nonlinear problems  $\lambda^2$  should be relatively large in the early iterations, so that  $\mathbf{x}$  is small enough to assure convergence to a solution. Then, as the iterations proceed the value of  $\lambda^2$  should be decreased gradually until it reaches its regularization value.

Another regularization approach is to limit the ‘‘roughness’’ of the solution by constraining its Laplacian to be equal to zero ([86] and references therein). This constraint can be introduced by writing the Laplacian using a finite-difference approximation

$$6x_{i,j,k} - (x_{i-1,j,k} + x_{i+1,j,k} + x_{i,j-1,k} + x_{i,j+1,k} + x_{i,j,k-1} + x_{i,j,k+1}) = 0 \quad (95)$$

where  $x_{i,j,k}$  is the value of  $\mathbf{x}$  at a grid point (identified by the three indices) and  $x_{i\pm 1,j,k}$ ,  $x_{i,j\pm 1,k}$ , and  $x_{i,j,k\pm 1}$  are the values of  $\mathbf{x}$  at adjacent points. This 3-D constraint is



used in [15]. Equation (95) must be translated into a matrix form, which will depend on the number of grid points in the three dimensions. Let  $\mathbf{L}$  indicate this matrix. Then, the constrained system becomes

$$\begin{pmatrix} \mathbf{A} \\ \lambda \mathbf{L} \end{pmatrix} \mathbf{x} = \begin{pmatrix} \mathbf{b} \\ \mathbf{0} \end{pmatrix}. \quad (96)$$

To solve this new system by ordinary least squares we must write an expression similar to Eq. (89), which in turn gives

$$(\mathbf{A}^T \mathbf{A} + \lambda^2 \mathbf{L}^T \mathbf{L}) \mathbf{x} = \mathbf{A}^T \mathbf{b}. \quad (97)$$

In general,  $\mathbf{L}$  is positive definite (see, e. g., [5,114]) and so is  $\mathbf{L}^T \mathbf{L}$ , which means that this approach also leads to a well-conditioned system.

In practice, Eqs. (88) and (96) can be solved using methods that compute the solution iteratively, such as the ART, SIRT, and LSQR methods. As noted earlier, ART is affected by noise in the data, while SIRT has the disadvantage that it introduces an unwanted scaling of the problem. This question will not be discussed here, but it should be noted that the term SIRT does not refer to a single technique; rather it refers to a family of iterative methods based on ideas similar to those described in [48]. The earlier application to geophysical problems is described in [39], but the algorithm used there was equivalent to the solution of a problem that was a scaled version of the original one, i. e., an equation similar to Eq. (58) is multiplied by a certain diagonal matrix which is not controlled by the user [73]. A similar problem affects the original SIRT formulation [63]. A general discussion of this question can be found in [154]. The LSQR method [108] was introduced in seismology by Nolet ([104] and references therein), does not have the problems that affect ART and SIRT, and is very well suited for the tomographic method because the matrices involved are highly sparse, which helps speed the computations considerably. For these reasons, the LSQR method is used widely used in seismic tomography (see, e. g., [15,86,163]).

So far we have not discussed how the parameter  $\lambda^2$  should be chosen. This question does not have a definite answer. In their discussion of the damped least-squares method, Lawson and Hanson [83] plotted the norm (i. e. length) of the solution vector versus the norm of the vector of residuals obtained for different values of the damping parameter and noted (for specific examples) that the resulting curve had an approximate L shape. The optimal value of the damping parameter was that corresponding to the corner of the curve at which it goes from nearly vertical to nearly horizontal. The idea here is to minimize both the residual and the length of the solution. This approach

is discussed further in [55] and [56], where the logarithms of the two norms, rather than the norms themselves, are used. A somewhat related approach is based on the plot of the data variance versus the solution variance [42]. A recent example of choice of damping parameter motivated by this latter approach is provided in [78].

We end this section with the definition of *resolution*. Regardless of how Eq. (74) is solved, we can write

$$\hat{\mathbf{x}} = \hat{\mathbf{A}} \mathbf{b} \quad (98)$$

where  $\hat{\mathbf{A}}$  can be considered the inverse, in some sense, of  $\mathbf{A}$ . Now write  $\mathbf{b}$  using Eq. (74). This gives

$$\hat{\mathbf{x}} = \hat{\mathbf{A}} \mathbf{A} \mathbf{x} \equiv \mathbf{R} \mathbf{x} \quad (99)$$

where  $\mathbf{R}$  is known as the resolution matrix (see, e. g., [3]) and is defined by the identity. If  $\mathbf{R} = \mathbf{I}$ ,  $\hat{\mathbf{x}} = \mathbf{x}$ . If  $\mathbf{A}$  were known perfectly well, this result would imply that the solution vector  $\hat{\mathbf{x}}$  is exactly equal to the true solution. In practice, in seismic tomography  $\mathbf{A}$  is only poorly known, and even if  $\mathbf{R} = \mathbf{I}$  there is no reason to say that  $\hat{\mathbf{x}}$  represents the true solution, as sometimes stated. To see that, consider an extreme case. Let us suppose that a velocity model consists of just one block. In this case  $\mathbf{A}$  will have only one column and will be well conditioned, which means that the corresponding Eq. (74) can be solved by least squares. The resulting resolution matrix will be equal to  $\mathbf{I}$ , but it is obvious that the computed solution will not correspond to the true velocity of the earth (except in the unlikely case that the velocity is a constant). If the number of blocks is gradually increased to make the model more realistic, at some point  $\mathbf{A}$  will become ill-posed and a regularized solution must be introduced. In this case from Eq. (87) we get

$$\hat{\mathbf{A}} = (\mathbf{A}^T \mathbf{A} + \lambda^2 \mathbf{D})^{-1} \mathbf{A}^T \quad (100)$$

and

$$\mathbf{R} = (\mathbf{A}^T \mathbf{A} + \lambda^2 \mathbf{D})^{-1} \mathbf{A}^T \mathbf{A} = \frac{\mathbf{A}^T \mathbf{A}}{\mathbf{A}^T \mathbf{A} + \lambda^2 \mathbf{D}} \neq \mathbf{I} \quad \text{for } \lambda^2 \neq 0. \quad (101)$$

Here  $\mathbf{R}$  was written as a ratio for visual purposes only. Note that  $\mathbf{R}$  is not equal to  $\mathbf{I}$  when  $\lambda^2 = 0$  unless  $(\mathbf{A}^T \mathbf{A})^{-1}$  exists. On the other hand, from Eq. (85) we know that the inverse of the matrix in parentheses always exists because  $\mathbf{D}$  is assumed to be positive definite.

Now consider the implications of a resolution matrix not equal to the identity. Using Eq. (98) we can write the  $i$ th component of the solution vector as

$$\hat{x}_i = \sum_k R_{ik} x_k, \quad (102)$$

so that  $\hat{x}_i$  is equal to a linear combination of some of the components of  $\mathbf{x}$ , not to  $x_i$  (as would be the case if  $\mathbf{R} = \mathbf{I}$ ). This result can be interpreted as follows. As the number of parameters in a velocity model is increased (e. g., the number of blocks is increased), it is expected that it will become a more realistic representation of the actual velocity variations in the earth. However, the available data do not allow the determination of individual values of the parameters; only average values can be computed. We may summarize this situation by saying that in geophysics, as in economics, “there is no such thing as a free lunch”. Note that  $\mathbf{R}$  does not depend on the the data. Rather, it is controlled by the distribution of stations and seismic sources. Seismologists have some control on the former (within financial and logistic limits), but not on the latter, which places a strong constraint on the resolution that can be achieved in any given case. The Bayesian approach discussed below is an attempt to circumvent this intrinsic limit.

The computation of the resolution matrix when the number of velocity parameters is very large is computationally too expensive, and in such cases the resolution is estimated using synthetic arrival times using velocity models with simple patterns. An early approach [52] was based on the use of several high velocity areas extending through the whole model depth embedded in a constant velocity medium. The synthetic data were computed using the data raypaths and inverted as the actual data. Comparison of the model obtained by inversion with the model used to generate the synthetic data gives an idea of the resolving capabilities of the distribution of sources and receivers used. In a similar approach, synthetic data were generated for a checkerboard model with alternating positive and negative velocities in three dimensions [72]. In another variation only two blocks of the model had velocities different from those of the rest of the blocks [70]. The checkerboard approach has become very popular, but in principle, other synthetic data sets can be used (see Sect. “Examples”).

Finally, let us note that in addition to its mathematical definition, the term resolution is generally used (i. e., not only in seismology) to indicate level of detail. For example, given two images of the same physical object acquired using different instruments, the one showing more detail is said to have more resolution. The same idea applies to velocity models, and it is not uncommon to see references in the literature to high- or low-resolution models. These terms are useful in a qualitative way, and may not be directly related to what one would call high or low in a mathematical sense. In fact,  $\mathbf{R}$  can be made closer (and even equal) to  $\mathbf{I}$  by decreasing the number of parameters (see, e. g., [74]), which by itself will decrease the level of detail in

the model. For example, if a number of blocks in a model is replaced by a single block occupying the same volume, the velocity in the larger block will be the average (in some sense) of the velocities in the smaller blocks. As a consequence, if the small-block velocities are significantly different, the model based on the larger blocks would have lost detail and, thus, will have less resolution, although mathematically it may be larger.

### Bayesian Approach

This approach is based on probability considerations, and to discuss it we need a few basic definitions and results, based on [130]. Let  $X$  and  $Y$  be random variables and  $x$  and  $y$  be elements of a discrete set of real values. Also let  $P_X(x)$  be the probability that the event “ $X = x$ ” occurs and  $P_{X,Y}(x, y)$  be the probability that the event “ $X = x$  and  $Y = y$ ” occurs. Then the *conditional probability* that the event “ $Y = y$ ” occurs given that the event “ $X = x$ ” has occurred is given by

$$P_{Y|X}(y|x) = \frac{P_{X,Y}(x, y)}{P_X(x)}. \quad (103)$$

Similarly,

$$P_{X|Y}(x|y) = \frac{P_{X,Y}(x, y)}{P_Y(y)}. \quad (104)$$

These two equations can be combined by elimination of the common factor, giving

$$P_{X|Y}(x|y) = \frac{P_{Y|X}(y|x)P_X(x)}{P_Y(y)}. \quad (105)$$

This result is known as *Bayes’ theorem* (or rule).

When the random variables can take on values defined over a continuous interval, which is the case of interest to us, the following definitions are introduced.

(a) *Probability distribution function*  $F_X(x)$ :

$$F_X(x) = P(-\infty < X \leq x) \quad (106)$$

where  $P$  indicates the probability of the event in parentheses.

(b) *Probability density function*  $p_X(x)$ :

$$p_X(x) = \frac{dF_X(x)}{dx}. \quad (107)$$

From Eq. (107) it follows that

$$F_X(x) = \int_{-\infty}^x p_X(a) da. \quad (108)$$

A well known example of probability density function (or pdf, for short) is the Gaussian pdf, introduced below.

The definitions given above are for a single random function. When dealing with a number of them it is convenient to introduce a vector notation. For the case of  $n$  random variables  $X_1, X_2, \dots, X_n$ , we have the following two definitions

$$p_{\mathbf{X}}(\mathbf{x}) = p_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) \quad (109)$$

and

$$F_{\mathbf{X}}(\mathbf{x}) = \int_{-\infty}^{\mathbf{x}} p_{\mathbf{X}}(\mathbf{a}) d\mathbf{a} \quad (110)$$

where the subscript in  $p_{\mathbf{X}}$  and  $F_{\mathbf{X}}$  is the vector  $\mathbf{X} = (X_1 X_2 \dots X_n)^T$ , the integral symbol represents an  $n$ -fold integral, and  $d\mathbf{a} = da_1 da_2 \dots da_n$ . Then

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{\partial^n F_{\mathbf{X}}(\mathbf{x})}{\partial x_1 \partial x_2 \dots \partial x_n} \quad (111)$$

and

$$\begin{aligned} p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} &= F_{\mathbf{X}}(\mathbf{x} + d\mathbf{x}) - F_{\mathbf{X}}(\mathbf{x}) \\ &= P(\mathbf{x} < \mathbf{X} \leq \mathbf{x} + d\mathbf{x}) \end{aligned} \quad (112)$$

where  $d\mathbf{x}$  is the vector with  $i$ th component given by  $dx_i$ .

For vector random variables Bayes' theorem becomes

$$p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) = \frac{p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) p_{\mathbf{X}}(\mathbf{x})}{p_{\mathbf{Y}}(\mathbf{y})}. \quad (113)$$

In the following we will adopt the common practice of dropping the subscripts of the probability density functions.

The application of Bayes' theorem to inverse problems is based on the following interpretation of the theorem. Before an experiment is conducted, a random variable  $\mathbf{X}$  has a given a priori pdf  $p(\mathbf{x})$ . During the experiment data  $\mathbf{y}$  are collected, and as a result  $\mathbf{X}$  has an a posteriori pdf given by  $p(\mathbf{x}|\mathbf{y})$  (see, e. g. p. 36 in [10]). The pdf  $p(\mathbf{y}|\mathbf{x})$  represents the probability that an experiment would have produced the result  $\mathbf{y}$  if the value of the variable had been  $\mathbf{x}$ . To apply Bayes' theorem to our problem, we need to rewrite Eq. (74) adding a random error vector  $\mathbf{e}$

$$\mathbf{Ax} = \mathbf{b} + \mathbf{e} \quad (114)$$

and to assume that  $\mathbf{x}$  and  $\mathbf{b}$  are random variables. In this formulation the data are represented by  $\mathbf{b}$  (which plays the role of  $\mathbf{y}$ ), and  $p(\mathbf{b}|\mathbf{x})$  for a given  $\mathbf{x}$  is the probability of the error  $\mathbf{e}$  [71]. The denominator in Eq. (113) does not depend on  $\mathbf{x}$  and can be ignored. The question that remains

is the selection of the form of the probability density functions on the right side of the equation. The standard choice is a *Gaussian* pdf, given, in general, by

$$p(\mathbf{x}) = (2\pi)^{-n/2} |\mathbf{C}_{\mathbf{x}}|^{-1/2} \cdot \exp \left[ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})^T \mathbf{C}_{\mathbf{x}}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}) \right] \quad (115)$$

where  $n$  is as in Eq. (109),  $\mathbf{C}_{\mathbf{x}}$  and  $\boldsymbol{\mu}_{\mathbf{x}}$  are the covariance matrix and mean of  $\mathbf{x}$ , respectively, and the vertical bars indicate determinant. Introducing this expression and a similar one for  $p(\mathbf{e})$  in Eq. (113), and then taking logarithms on both sides of the resulting equation gives

$$\ln[p(\mathbf{x}|\mathbf{b})] = -\frac{1}{2} [(m+n) \ln(2\pi) + \ln |\mathbf{C}_{\mathbf{e}}| + \ln |\mathbf{C}_{\mathbf{x}}| + S] - \ln(p(\mathbf{y})) \quad (116)$$

where  $m$  is the number of components of  $\mathbf{b}$  and

$$S = (\mathbf{e} - \boldsymbol{\mu}_{\mathbf{e}})^T \mathbf{C}_{\mathbf{e}}^{-1} (\mathbf{e} - \boldsymbol{\mu}_{\mathbf{e}}) + (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})^T \mathbf{C}_{\mathbf{x}}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}). \quad (117)$$

Now we will make the additional assumption that  $\boldsymbol{\mu}_{\mathbf{e}} = \mathbf{0}$  and will write  $\mathbf{e} = \mathbf{Ax} - \mathbf{b}$  (using Eq. (114)). Under these conditions Eq. (117) becomes

$$S = (\mathbf{Ax} - \mathbf{b})^T \mathbf{C}_{\mathbf{e}}^{-1} (\mathbf{Ax} - \mathbf{b}) + (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})^T \mathbf{C}_{\mathbf{x}}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}). \quad (118)$$

The last step is to find the value of  $\mathbf{x}$  that maximizes  $\ln p(\mathbf{x}|\mathbf{b})$  (and thus,  $p(\mathbf{x}|\mathbf{b})$ ) and, therefore, minimizes  $S$ . Taking the derivative of  $S$  with respect to  $\mathbf{x}$  and setting it equal to zero leads to the *maximum a posteriori estimator*  $\mathbf{x}_{\text{MAP}}$ , given by

$$\mathbf{x}_{\text{MAP}} = \mathbf{P} (\mathbf{A}^T \mathbf{C}_{\mathbf{e}}^{-1} \mathbf{b} + \mathbf{C}_{\mathbf{x}}^{-1} \boldsymbol{\mu}_{\mathbf{x}}) \quad (119)$$

where

$$\mathbf{P} = (\mathbf{A}^T \mathbf{C}_{\mathbf{e}}^{-1} \mathbf{A} + \mathbf{C}_{\mathbf{x}}^{-1})^{-1}. \quad (120)$$

An alternative expression for  $\mathbf{x}_{\text{MAP}}$  is

$$\mathbf{x}_{\text{MAP}} = \boldsymbol{\mu}_{\mathbf{x}} + \mathbf{PA}^T \mathbf{C}_{\mathbf{e}}^{-1} (\mathbf{b} - \mathbf{A}\boldsymbol{\mu}_{\mathbf{x}}). \quad (121)$$

These results are based on [13]. An expression similar to Eq. (119) is given in [130]. The Bayesian approach presented here was introduced in the context of X-ray tomography under the assumption that  $\boldsymbol{\mu}_{\mathbf{x}} = \mathbf{0}$  [71]. An application can be found in [64]. In geophysics a similar approach was introduced in [141] and [142].

Equation (121) is important for two reasons. First,  $\mathbf{x}_{\text{MAP}}$  can be interpreted as the sum of two terms: one corresponding to the prior information about the value of  $\mathbf{x}$

(given by  $\boldsymbol{\mu}_x$ ) and another one that contains the new information provided by the data collected in the experiment. Second, using Eq. (121) and the fact that the mean values of  $\mathbf{x}$  and  $\mathbf{e}$  are  $\boldsymbol{\mu}_x$  and  $\mathbf{0}$ , respectively, it can be shown that the expected value of  $\mathbf{x}_{\text{MAP}}$  is  $\boldsymbol{\mu}_x$ . Therefore, the estimator is *biased* (i. e., its expected value is not  $\mathbf{x}$ ). The significance of this result is that the expected value of the estimator is independent of the data, which is not a desirable feature. Finally, we also note that the covariance matrix of the difference  $\mathbf{x}_{\text{MAP}} - \mathbf{x}$  is equal to the matrix  $\mathbf{P}$  [13].

Now let us consider two special cases. First, no a priori information is used. In this case  $S$  in Eq. (118) is equal to the first term on the right hand side and its minimization leads to the *generalized least-squares estimator*, given by

$$\hat{\mathbf{x}} = (\mathbf{A}^T \mathbf{C}_e^{-1} \mathbf{A})^{-1} \mathbf{A}^T \mathbf{C}_e^{-1} \mathbf{b} \quad (122)$$

(if the inverse exists). The *ordinary least-squares estimator* is obtained from Eq. (122) with  $\mathbf{C}_e = \sigma_e^2 \mathbf{I}$ , where  $\sigma_e$  is a measure of the standard deviation of the errors. The estimator  $\hat{\mathbf{x}}$  has the properties that it is unbiased and minimizes the determinant of its covariance matrix, and does not require a Gaussian pdf (see, e. g., [10]).

Second, let us assume that the a priori information is  $\boldsymbol{\mu}_x = \mathbf{0}$  and  $\mathbf{C}_x \neq \mathbf{O}$ . Then, using Eqs. (119) and (120) we get

$$\begin{aligned} \mathbf{x}_{\text{MAP}} &= (\mathbf{A}^T \mathbf{C}_e^{-1} \mathbf{A} + \mathbf{C}_x^{-1})^{-1} \mathbf{A}^T \mathbf{C}_e^{-1} \mathbf{b} \\ &= \mathbf{C}_x \mathbf{A}^T (\mathbf{A} \mathbf{C}_x \mathbf{A}^T + \mathbf{C}_e)^{-1} \mathbf{b}. \end{aligned} \quad (123)$$

A derivation of the second equality is given in [13]. This result agrees with similar results derived in [47] and [74] using different approaches. An alternative derivation can be found in [3]. If we further simplify the problem by letting  $\mathbf{C}_e = \sigma_e^2 \mathbf{I}$  and  $\mathbf{C}_x = \sigma_x^2 \mathbf{I}$ , the first equality in Eq. (123) gives

$$\begin{aligned} \mathbf{x}_{\text{MAP}} &= \left( \frac{1}{\sigma_e^2} \mathbf{A}^T \mathbf{A} + \frac{1}{\sigma_x^2} \mathbf{I} \right)^{-1} \frac{1}{\sigma_e^2} \mathbf{A}^T \mathbf{b} \\ &= \left( \mathbf{A}^T \mathbf{A} + \frac{\sigma_e^2}{\sigma_x^2} \mathbf{I} \right)^{-1} \mathbf{A}^T \mathbf{b}. \end{aligned} \quad (124)$$

This particular form of  $\mathbf{x}_{\text{MAP}}$  has been referred to as the *stochastic inverse* [2,4]. If we now put  $\lambda = \sigma_e / \sigma_x$  we see that this solution is formally similar to the damped least-squares solution obtained from Eq. (87) with  $\mathbf{D} = \mathbf{I}$ .

Another important result concerning  $\mathbf{x}_{\text{MAP}}$ , as given in Eq. (119), is that it can be obtained by simply treating the prior information as a constraint, i. e.,  $\mathbf{x} = \boldsymbol{\mu}_x$ , which then can be added to Eq. (114) in a way similar to what was done in Eqs. (88) and (96), and then finding the general-

ized least-squares estimator, defined by Eq. (122) [74,136]. When proceeding in this way there is no need to make the Gaussianity assumption, although without it (or some other assumption) it is not possible to derive statistical properties of the estimator.

Finally, we note that although the Bayesian approach seems to be a good way to handle the problem of low resolution that affects many seismic tomography problems, it has some serious drawbacks that should not be overlooked. As noted earlier, the Bayesian solution is biased by the prior information, and if this information is incorrect, the solution will be affected by some amount of error that cannot be quantified objectively. The assumption that the velocity adjustments can be described by a Gaussian pdf (or any other simple pdf) is introduced by mathematical convenience, not because it is likely to represent the actual 3-D variations of the velocity adjustments within the earth. This applies, particularly, to crustal areas within complicated tectonic settings. Finally, the covariance matrix is also needed, but because it is essentially unknowable, it is frequently assumed that it is diagonal. Again, this is an assumption based on convenience, rather than on scientific facts. A critique of the Bayesian approach as applied to geophysical inverse problems can be found in [109]. For a somewhat different perspective see [104]. A good overview of the Bayesian method can be found in [57]. An example of the problems that erroneous prior information may introduce is provided in Subsect. “**Effect of Inaccurate Prior Information on the Location of Aftershocks**”.

## Examples

Here we present two examples of local tomography, one from the San Fernando basin in southern California and one from Taiwan, and also give an example of the problems that can be created when erroneous prior information is used. The tomography software used was written by H. Benz [15] and solves for earthquake locations and slownesses using the formulation described in Subsect. “**Local Slowness Tomography**” – Subsect. “**Decoupling of the Earthquake Location and Tomography Problems**”. The slowness model is parameterized in terms of constant velocity blocks and the computation of arrival times is performed using the software of Podvin and Lecomte [113]. As noted in Subsect. “**Computation of Local Travel Times**”, this software has become popular because of its ability to handle sharp velocity variations accurately. As a consequence, the resulting velocity models show high resolution (i. e. a high level of detail). For the examples presented here, none of the other existing models show such resolution. Additional results derived using

Benz's software can be found in, e. g., [95,105] and [159]. The original software relocated the events with a constant  $v_p/v_s$  ratio but was modified according to Eq. (65). To regularize the solution Eq. (96) is used.

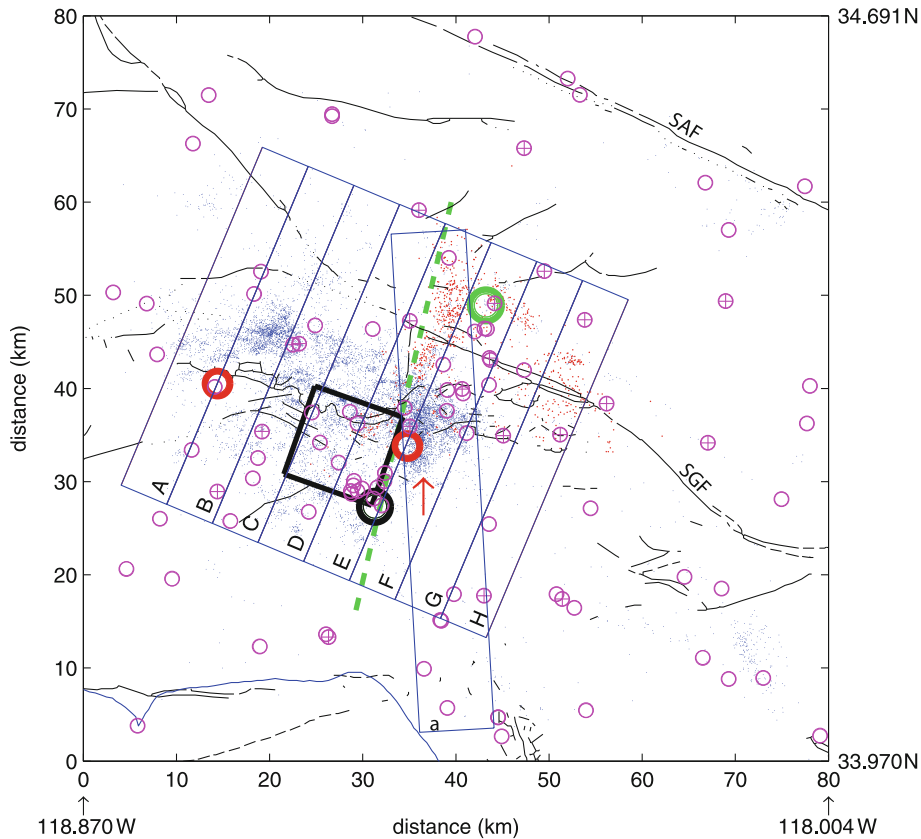
### **P Wave Velocity Model for the San Fernando, California, Area**

The Los Angeles, California, region is one of the most seismically active in the United States, and because of its extremely large population (over ten million), a large earthquake there may be catastrophic in terms of human and/or economic costs. A major source of seismic hazard there is the presence of large sedimentary basins, which have much lower velocity than their host rocks. As a consequence, they can amplify significantly the ground motion caused by earthquakes. For example, a study published in 1989 [160] showed that the ground motion in the San Fernando and Los Angeles basins can be amplified by factors of about three and four with respect to rock sites, and that the 3-D velocity models available were not capable of generating this amplification. To address this and other problems, the Southern California Earthquake Center (SCEC) supported the development of reference of 3-D  $P$  and  $S$  wave velocity models for the major basins of southern California [91]. These models were constructed using depth and age of sediments data compiled as part of oil and water exploration studies and geological studies. Empirical relations between these two types of data were used to estimate  $P$  wave velocities. Additional empirical relations between  $P$  wave velocities, density, and Poisson's ratio were used to calculate  $S$  wave velocity. For depths less than 300 m, the  $P$  and  $S$  wave velocities were constrained using borehole velocity data. Deep borehole information was used for calibration purposes. For the San Fernando basin four boreholes were available, but the deepest one was only one 3.5 km deep. Therefore, below that depth the model does not have hard constraints. For the rocks outside of the basins an existing 3-D tomographic velocity model [58] was used. This model assigns velocities to points on a grid, which has a 15 km x 15 km horizontal spacing at depths of 1, 4, 6, 10, 15, 17, 22, 31 and 33 km.

The  $P$  wave velocities in the SCEC model for the Los Angeles basin were compared to sonic log velocities in [140]. This comparison shows that the standard deviation of the velocity differences is about 440 m/s, which is up to 20% of the model velocities. An underestimation of the velocities near the center of the basin and an overestimation near the border was also observed. An additional comparison [139] produced results that are consistent with those in [140].

The 1994,  $M = 6.7$ , Northridge, California, earthquake occurred on a previously unknown blind thrust fault in the San Fernando valley, to the NW of Los Angeles. This was the costliest earthquake in the United States (about US\$ 20,000 million), although the number of deaths was small (58, [143]), thanks to building code provisions and the fact that the event occurred early in the morning (4:30 AM local time). Because of its importance, this earthquake has been extensively studied, but more than ten years after its occurrence, a number of important questions still remain partially or totally unanswered (see, e. g., [123]). One of them is the exact nature of the 3-D velocity variations in the area, as the existing velocity models (see, e. g., [59,97,118,162]) have low resolution.

The velocity model issue has been addressed in [123]. The Benz's tomography software was applied to events in the Northridge mainshock-aftershock sequence and to aftershocks of the nearby 1971,  $M = 6.6$ , San Fernando earthquake. The data used were 192,421  $P$  wave first arrivals from 12,656 events recorded during 1981–2000 by 81 permanent and temporary stations and 799 aftershocks of the San Fernando earthquake recorded by a portable network of 20 stations (Fig. 4). The velocity model covers a volume with a surface area of 80 km by 80 km and a depth of 24 km and was divided into cubic blocks with sides of 2 km. For the computation of travel times cubic blocks with sides of 1 km were used. The initial locations were computed using a standard 1-D velocity model with three layers having thicknesses of 5.5, 10.5 and 16 km, and corresponding velocities of 5.5, 6.3, and 6.7 km/s. For the inversion the initial velocity model was the 1-D model described in [60] with a small modification (the velocity in the upper 2 km was reduced from 4.8 to 4.5 km/s). A cross section of the model is shown in a subsequent figure. The number of iterations was ten, and the value of  $\lambda$  in Eq. (96) ranged between 64 (first four iterations) and 20 (last two iterations) with intermediate values in between. Because some of the stations are in the mountains, a depth of  $-2$  km was used as a reference depth for the model. The root-mean square residual for all the events was 0.17 s and 0.07 s for the first and last iterations. Representative velocity cross sections (Fig. 5) show that the resulting model has much more detail than any of the other published models. To make sure that this detail was not artificial the initial model was changed and different data subsets were used, with the result that the most important aspects of the model were robust. In addition, realistic synthetic arrival times were generated with the velocity model and event locations from the last iteration. This involved using the stations that had recorded the observed data (for each event) and applying the original weights. Then the synthetic arrivals times were



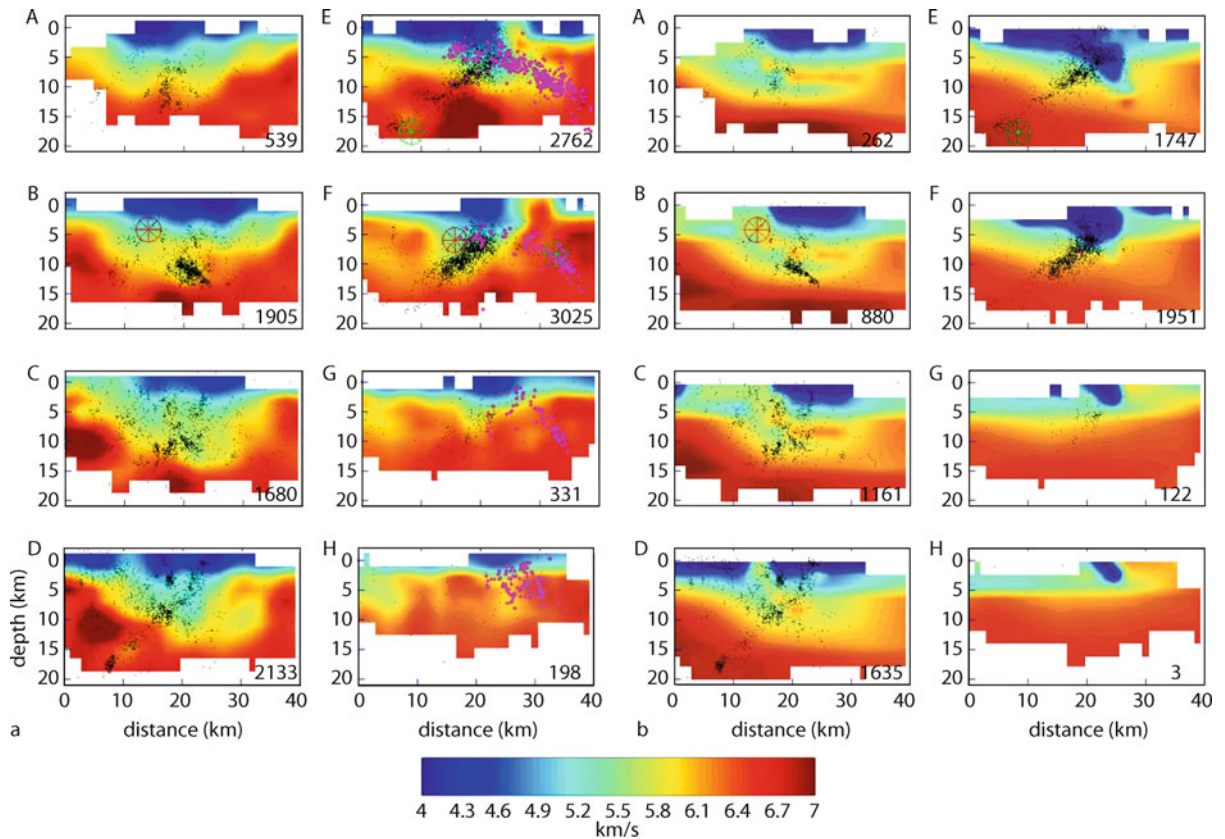
Tomography, Seismic, Figure 4

Area around the San Fernando valley, California, for which a 3-D  $P$  wave velocity model has been determined. *Blue dots* indicate the epicenters of 12,656 seismic events recorded during 1981–2000 (mostly aftershocks of the 1994,  $M = 6.7$ , Northridge earthquake). *Red dots* indicate epicenters of 799 aftershocks of the 1971,  $M = 6.6$ , San Fernando earthquake recorded during February–April 1971. The locations shown here were determined as part of the velocity determination and earthquake relocation. The large *black* and *green circles* indicate the epicenters of the Northridge and San Fernando mainshocks, respectively, the latter provided by the U.S. Geological Survey (U.S.G.S.). *Magenta open* and *crossed circles* indicate the stations that recorded the Northridge and San Fernando events, respectively. The large *red circles* indicate the epicenters of the two largest Northridge aftershocks (the location of the eastern one is from the U.S.G.S. and is poorly constrained). *Black lines* indicate faults. The San Andreas and San Gabriel faults are labeled SAF and SGF. The bold box is the projection of the Northridge earthquake fault plane determined in [68] using geodetic data. The events and velocities in boxes A through H are shown in cross section form in Fig. 5. Most of the San Fernando events are to the east of the *green dashed line* while the Northridge mainshock rupture occurred to the west of that line. The prominent band of Northridge seismicity in a roughly N-S direction (identified by the red arrow in box F) did not occur within the area that ruptured during the main shock. The events and velocity within the box labeled a are shown in cross section form in Fig. 5. From [123]

inverted as the actual data (i. e., the same initial locations and velocity model were used). The corresponding velocity model is very close to the input model (Fig. 6). A few areas have velocity differences of up to  $\pm 0.5$  km/s, but this does not detract from an overall good agreement. Regarding the hypocentral locations, the average difference between the true and computed values is 0.15 km in epicenter and 0.23 km in depth.

The inversion model is also supported by two additional pieces of evidence. One is the SCEC model, which

was sampled at the centers of the blocks in the inversion model. However, because the SCEC model has the surface as zero depth and the elevation in the area ranges between 0 and 1.6 km, a direct depth comparison is not strictly possible and for comparison purposes it was referred to a base elevation of  $-0.6$  km. As Fig. 5 shows, the inversion model has the main features of the SCEC model, and although there are some obvious differences, some of them come from deficiencies in the SCEC model, as discussed next. A second piece of evidence is the ex-



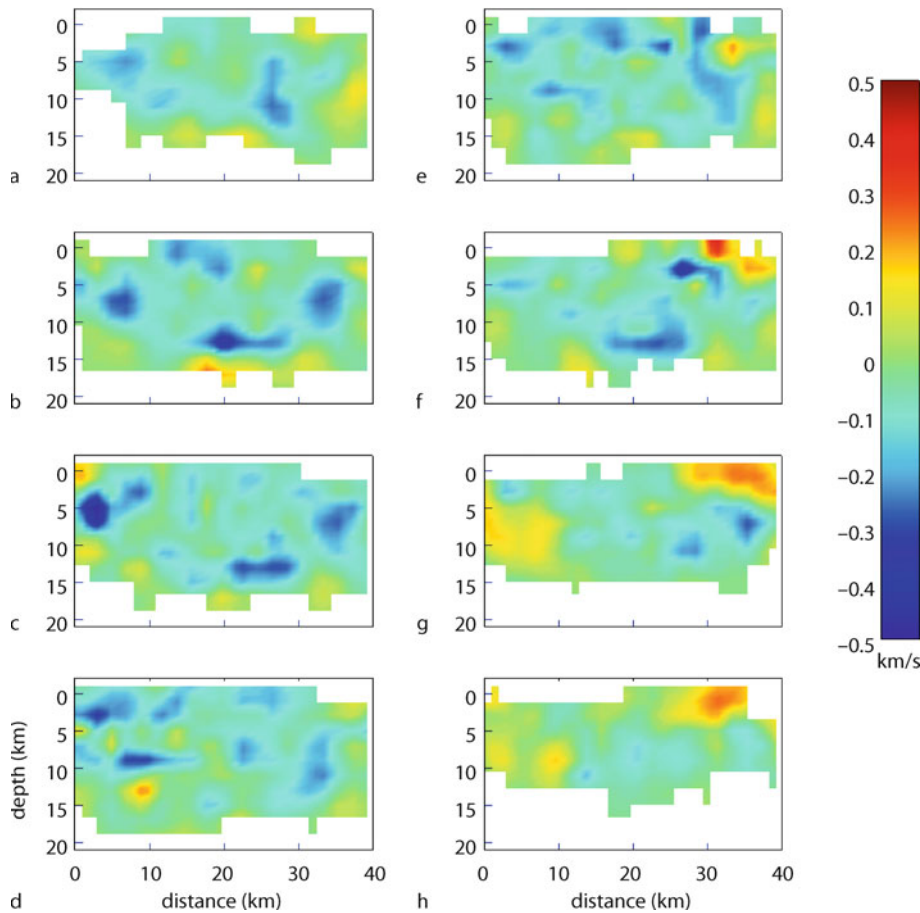
Tomography, Seismic, Figure 5

**a** Depth cross sections for the 3-D  $P$  wave velocity model determined by inversion and the events in the boxes A through H in Fig. 4. The width of the cross sections is 5.3 km. The letters are on the southern ends of the cross sections. The velocities are assigned to the centers of the model blocks and are interpolated along planes passing through the centers of the cross sections. *Black and magenta dots* indicate Northridge and San Fernando aftershocks. The large circled asterisks in cross sections E, B and F indicate the Northridge mainshock, its two largest aftershocks and the San Fernando mainshock (see Fig. 4). The number in the right lower corner of each cross section denotes the number of events. Only the velocity blocks covered by a combined ray length of 0.1 km or more are shown. Note the correlation between seismicity and velocity. The events between about 10 and 20 km in D and E are within high-velocity, basement rocks, and form narrow and well-defined lineations. These events span about 10 km horizontally and basically define the width of the fault that slipped during the main shock. The Northridge aftershocks in F correspond mostly to those indicated by the *red arrow* in Fig. 4. Most of these events are shallower than about 14 km and form a band of seismicity within and near the edge of the basin. The north-dipping events in B below about 10 km probably occurred on the Santa Susana fault. **b** Corresponding cross sections for the SCEC 3-D velocity model [91] and the events that occurred during 1994. From [123]

cellent agreement (Fig. 7) with a roughly 55 km long density model derived from the analysis of gravity data [82]. In particular, the low-velocity areas in the model correlate with the Los Angeles, San Fernando and Soledad basins, while high velocities underlie the Santa Monica and San Gabriel mountains. In contrast, the SCEC model does not match the density model equally well. Figure 7 also shows another tomographic model [59], computed for a grid with a 10 km  $\times$  10 km horizontal spacing at depths of 1, 4, 6, 8, 12, 16 and 20 km. Clearly, this model has a poor resolution and may be affected by significant artifacts, which may be

a direct consequence of the large horizontal dimensions of the blocks. Therefore, for a more direct comparison the inversion with Benz's software was repeated with blocks having horizontal sides of 10 km. As Fig. 8 shows, the corresponding results closely resembles those in Fig. 7b, which seems to indicate that the lack of resolution of the model in Fig. 7d may be due to either the software or the inversion parameters used, or both.

The results of the event relocation are also important. There are two types of differences between the initial and final locations. One is quasi-systematic, with the initial lo-



Tomography, Seismic, Figure 6

**a** Depth cross sections of the difference between the 3-D model shown in Fig. 5a and the 3-D model determined by inversion of synthetic arrival times generated using the model in Fig. 5a and the event locations determined during the inversion. See text for further details. From [123]

cations on average 1.0 km to the east and 0.5 km to the south of the inversion locations, and 1.5 km deeper. These epicentral differences are a consequence of the low velocities within the basin, which bias the event locations as noted earlier (see Fig. 3). These results are in agreement with a similar shift found using the joint hypocentral determination (JHD) technique [118], and have not been reported in other published tomographic inversion papers. A second difference between the initial and final locations is a considerable reduction in the epicentral scatter seen in the single-event locations. On average, there is a 1.5 km difference between the single-event and inversion epicentral locations.

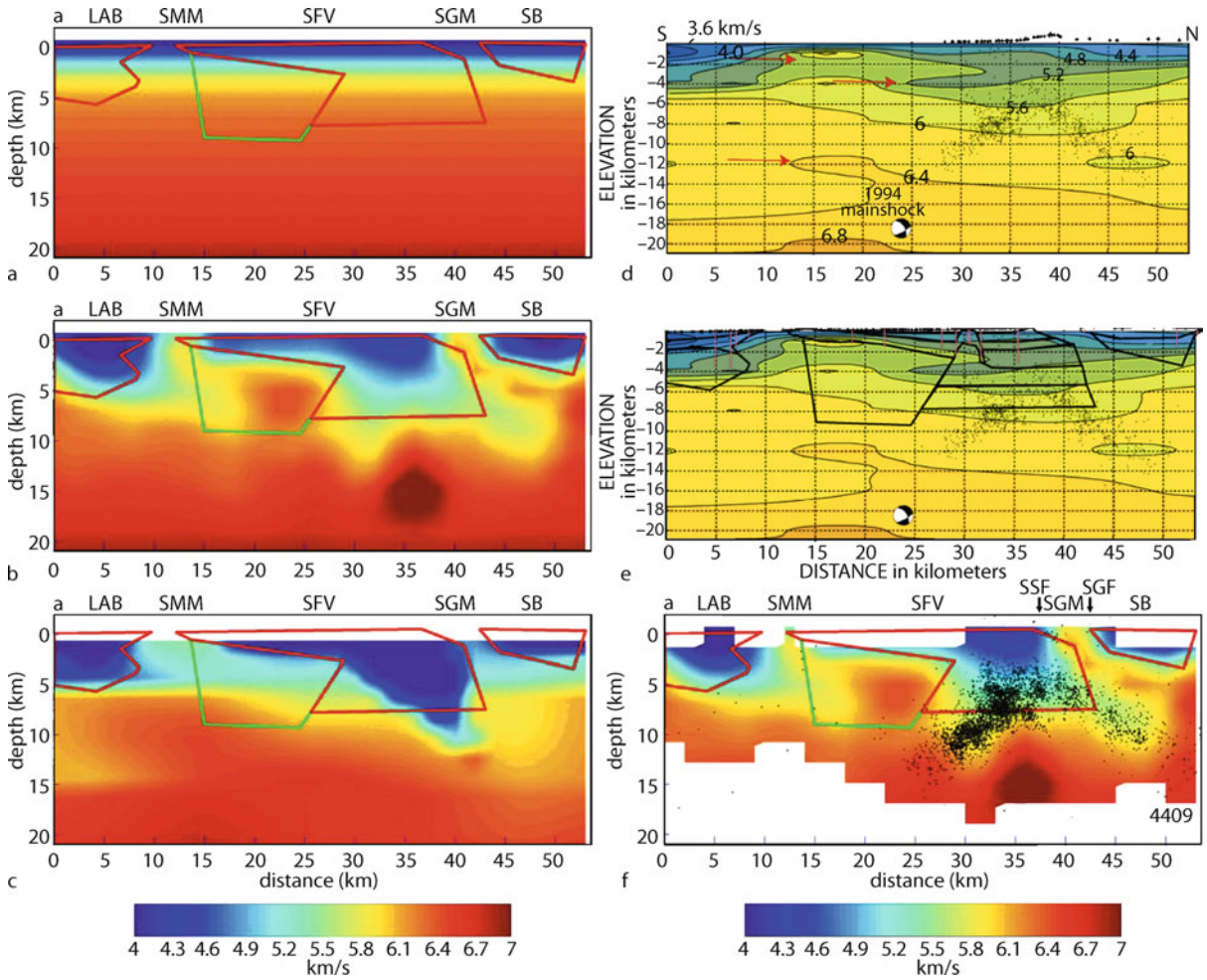
The combination of a high-resolution velocity model and improved event locations has important tectonic implications. For example, Fig. 5a shows that the most of

the seismicity occurred within the sedimentary rocks of the basin. There is little activity within the basement, mostly confined to an area around the fault plane of the Northridge earthquake. The relation between the Northridge earthquake aftershocks to the east of the main shock and San Fernando aftershocks has also been clarified; they occur on the flanks of a common high velocity block (Fig. 7f). Other significant results, not discussed here, can be found in [123].

### P and S Wave Velocity Models for Taiwan

The seismic hazard in Taiwan is also very high and for this reason a large network of seismic stations covers the island (Fig. 9). In addition, the collision and subduction processes involving the Eurasian and Philippine plates in



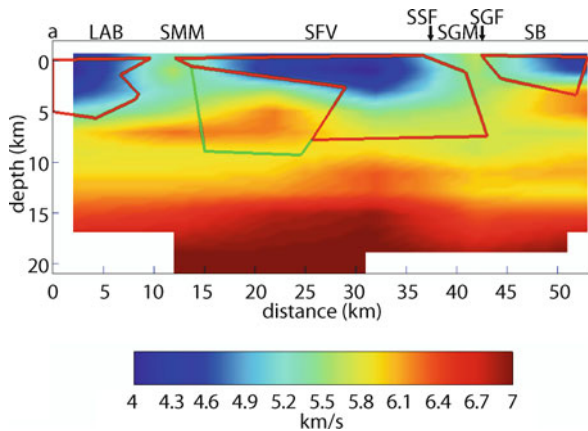


Tomography, Seismic, Figure 7

Velocity depth cross section along the center of box a in Fig. 5 for several velocity models. **a** Initial model. **b** Inversion model. All the blocks, regardless of the ray coverage, are shown. The polygons represent the bodies (simplified) used in [82] to match a gravity profile. The *red lines* bound materials with densities ranging between 2.00 and 2.55 g/cc. The area bounded by *red and green lines* corresponds to low-density basement (2.65 g/cc). Elsewhere in the figure the density is 2.71 g/cc. Note the excellent agreement between the extent of the low velocities in the Los Angeles (LAB), San Fernando (SFV) and Soledad (SB) basins and the low-density bodies, as well as the presence of high velocities in the areas of the Santa Monica (SMM) and the San Gabriel (SGM) mountains. **c** SCEC 3-D velocity model. This model does not fit the density model as well as the inversion model. **d** and **e** the 3-D velocity model described in [59] and original density model (from [82]). The contour labels in **d** indicate velocities in km/s. The *red arrows* indicate possible artifacts in the model. *Crosses* indicate the locations of gravity stations. **f** Inversion velocity model for blocks having ray coverage of 0.1 km or larger. The fact that the initial model is one-dimensional indicates that the 3-D velocity variations seen in **a** for the blocks not shown here were determined in earlier iterations. Also shown are the events within box a in Fig. 4. The box width is 8 km. Note that the Northridge and San Fernando aftershocks are underlain by a wedge of basement, with the seismicity occurring where rocks with lower velocities are present. The *arrows* labeled SSF and SGF indicate the positions of the Santa Susana and San Gabriel faults. From [123]

the Taiwan region have resulted in very strong ongoing tectonic and orogenic activities. As a consequence, Taiwan is the focus of numerous studies, past and present, aimed at getting a better understanding of the nature of the deformation processes there. Some of those studies involve

the determination of the crustal an upper mantle velocity structure, which has been investigated by a number of researchers (see, e. g., [90,126,128]). However, although the resulting information has been highly valuable, the resolution of the velocity models was relatively low. This situ-



Tomography, Seismic, Figure 8

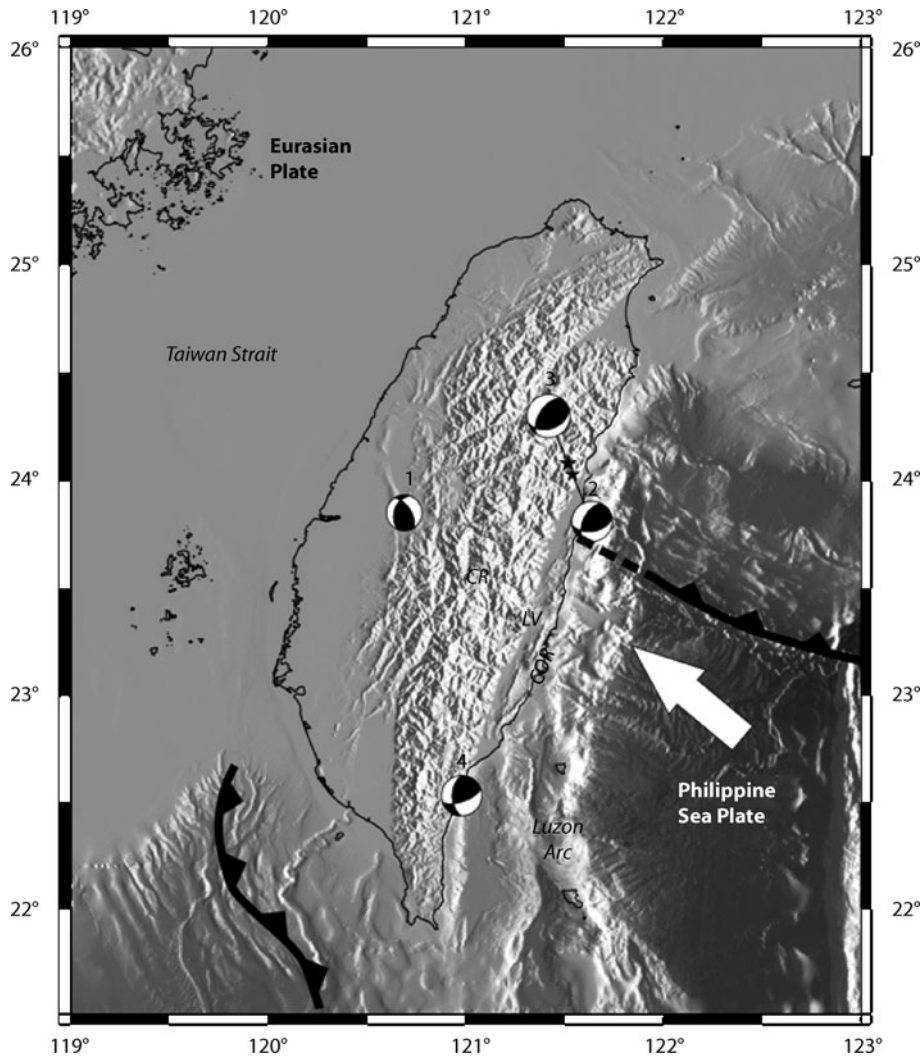
Similar to Fig. 7f for a velocity model with blocks having the two horizontal sides and the vertical side 10 km and 2 km long respectively. Comparison with Fig. 7b shows that even though the block size is horizontally much larger, the major features of the model are recovered

ation was improved by the work described in [77], which resulted in  $P$  and  $S$  wave velocity models with much higher resolutions. The main results will be described here (see also [78]). As noted above, the inversion was carried out using Benz's software. The block size was 8 km  $\times$  8 km (horizontally) and 2 km in depth for velocity and 2 km  $\times$  2 km  $\times$  2 km for travel time computation. The dataset included 69,758  $P$  wave arrivals and 42,733  $S$  wave arrivals from 6285 events distributed as uniformly as possible recorded by 78 permanent stations as well as 18,502 and 10,789  $P$  and  $S$  wave arrivals, respectively, from 1298 events recorded during two 30-station portable deployments (Fig. 10). The initial and final values of  $\lambda$  in Eq. (96) for the  $P$  and  $S$  wave inversions were 128 and 50 and 128 and 20, respectively. The number of iterations was 20 and the initial and final average root mean square residuals were 0.58 s and 0.15 s and 0.67 s and 0.21 s for the  $P$  and  $S$  waves, respectively. To avoid exceeding the limitations of the flat-earth approximation (see Subject. "Computation of Local Travel Times"), station-event pairs with epicentral distances greater than 140 km were not used in the inversion. The number of events used in the inversion was less than 5% of the total number of events recorded. Once the velocity models were determined, they were used to relocate all the locatable events using a 3-D location program based on the tomographic software [24].

Representative cross-sectional views of the computed 3-D  $P$ - and  $S$ -wave velocity models are shown in Fig. 11. As in the previous example, these models show higher resolution than other published models. The resolution of these

models was investigated with a checkerboard test, which shows that most of the original velocity pattern is well recovered to depths of 25 to 30 km for most blocks (Figs. 12, 13). Below that resolution decreases. Further evidence for the reliability of the velocity models comes from the comparison of observed station corrections and those determined using synthetic data generated with those models. This approach was introduced in [116] and [118]. The station corrections were computed using the JHD technique (see, e. g., [116,119]). The JHD corrections carry information on the lateral velocity under an array and are usually much larger than the corresponding station residuals. For example, in Taiwan they can be up to about  $\pm 1$  s and  $\pm 2$  s for  $P$  and  $S$  waves (Fig. 14), with the positive corrections corresponding to low-velocity areas (such as sedimentary basins) and the negative corrections associated with high-velocity areas. In general, if the observed and synthetic station corrections do not agree well with each other, it can be stated with confidence that the velocity model used to generate the synthetic data cannot be correct. On the other hand, a good agreement indicates that the inversion model is able to reproduce, at least, the actual velocity variations in an average sense along raypaths. For the Taiwan 3-D velocity model the agreement between the actual and synthetic  $P$  and  $S$  wave JHD corrections for two subsets of events is good (see Fig. 14 for an example), which gives confidence to the overall quality of the model.

Additional evidence in support of the  $P$  wave model is provided by the analysis of anomalous  $Pn$  waves recorded by stations along the collision suture zone in eastern Taiwan and generated by shallow eastern Taiwan events [89]. These waves can be observed at stations with epicentral distances as small as 60 km. This critical distances is much smaller than that for stations elsewhere in Taiwan and indicates the presence of an elevated Moho (i. e., a thinner crust) along the suture zone. Figure 15 shows travel times for three eastern earthquakes showing the normal and anomalous  $Pn$  waves and the corresponding Moho depths (having 36–38 km and 22–23 km depth ranges, respectively). In contrast, an earthquake in western Taiwan only shows the normal  $Pn$  waves. The presence of thinner crust along the suture zone agrees well with the 3-D velocity described here [77,78], as Fig. 16 shows. Because this zone is on the edge of the network, the resolution, although still acceptable, is not as good as under the network and the details of the model are not well resolved. However, an inversion of synthetic data similar to that described for the Northridge data showed that the main features of the velocity model on the eastern side are well recovered [77] and confirms that the profiles in Figs. 11 and 16 are representative of the actual velocity variations.



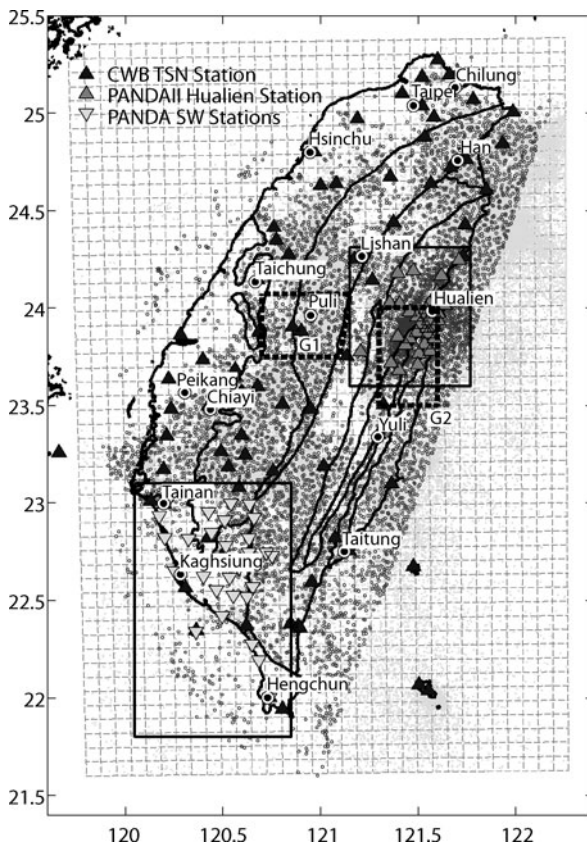
Tomography, Seismic, Figure 9

Tectonic setting of Taiwan and surrounding area. The Philippine Sea plate both subducts beneath, and collides with, the Eurasian plate, while the South China Sea sub-plate subducts beneath the Philippine Sea plate in southern Taiwan. The Longitudinal Valley (LV) is the suture zone corresponding to the collision and separates the Coastal Range (COR, part of the Philippine Sea plate) from the Central Mountain Range (CR). The beach ball symbols numbered 2–4 correspond to earthquakes with abnormal  $P_n$  waves recorded by stations along the eastern coast. Number 1 corresponds to an earthquake that does not show those  $P_n$  waves. See text for details. From [89]

### Effect of Inaccurate Prior Information on the Location of Aftershocks

The relocation of the mainshock and aftershocks of the 1989,  $M = 7.1$ , Loma Prieta, California, earthquake, offers a good example of the pitfalls that inaccurate prior information may introduce. This sequence occurred along the San Andres fault and was located using two 1-D  $P$  wave velocity models, one for stations to the southwest of the fault and one for stations to the northeast of it [37].

This division was based on differences in surface geology across the fault. The velocity of the NE model was up to about 0.1 km/s lower in the upper 1 km, and between 0.2 and 0.5 km/s higher between 1 and 9 km depth, with respect to the SE model. Below that depth the velocity differences did not exceed 0.1 km/s (Fig. 17). The two models were derived using a 1-D velocity inversion program that also included the computation of station corrections. According to [37], these models reflected the presence of elevated basement to the NE of the fault and Tertiary and



Tomography, Seismic, Figure 10

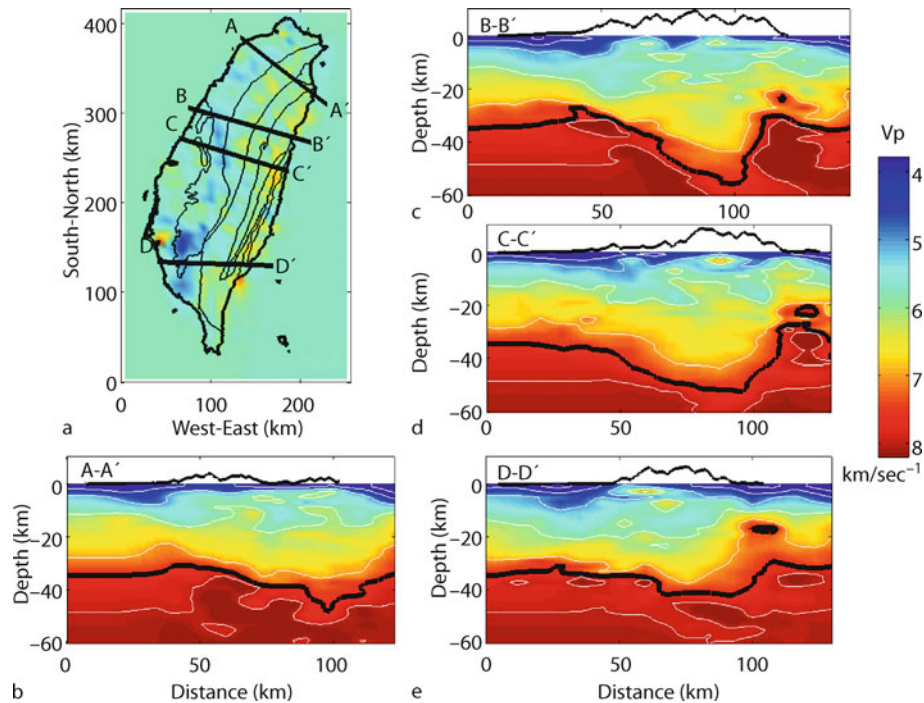
Seismic stations and events used in the 3-D tomographic study of Taiwan. *Solid triangles* indicate stations of the Taiwan Seismic Network (TSN), operated by the Central Weather Bureau (CWB). *Grey and inverted triangles* indicate stations of a portable network (PANDA) deployed in the Hualien and Pingtung areas (*solid rectangles*), respectively. The *small circles* correspond to the epicenters of the events used in the 3-D tomography. The *grey background* corresponds to the epicenters determined by the CWB between 1991 and 2002. The events within the *dashed rectangles* (labeled G1 and G2) were used with the joint hypocentral determination (JHD) technique. The  $8\text{ km} \times 8\text{ km}$  grid of the 3-D velocity is shown by the *dashed lines*. From [78]

younger sediments above 9 km depth to the SW of the fault. Note that although the prior information was not used as a quantitative constraint, it was essential to establish the boundary between the two velocity models and was implicitly used to justify the main features of the models. The seismicity was located with a single-event location program and these two velocity models and the corresponding station corrections. The mainshock fault plane inferred from the seismicity, however, had a considerable discrepancy with the fault plane determined using geodetic data, with the former consistently to the northeast of the latter by as much as 3 km. A possible explanation for

this discrepancy was the presence of lateral variations in the values of the elastic constants in the vicinity of the fault zone, which were not taken into account when the geodetic fault plane was determined [43]. A more likely explanation, however, is that the discrepancy was the result of a systematic mislocation of the events introduced by the use of incorrect velocity models. This problem was identified in [117], where the geodetic fault plane was compared to the event locations computed with the joint hypocentral determination (JHD) technique and a single velocity model, equal to the average of the two models described above [116]. With these new locations, the discrepancy was greatly reduced except in the southern end of the rupture zone, where the difference was about 1.2 m. This residual discrepancy is probably due to a significant lateral velocity contrast there. Subsequently, the analysis described in [37] was repeated with two modified velocity models [38], with the result that the new locations became similar to those in [116]. Interestingly, the new NE and SW velocity models in [38] are significantly different from their earlier counterparts (Fig. 17), with higher velocities to the SW of the San Andreas fault, rather than to the NE. Thus, these new models contradict the geological arguments used as supporting evidence for the earlier models, which in turn means that the prior information was not quite correct. Also worth noting is the fact that the locations determined using a 3-D velocity model also showed the discrepancy with the geodetic fault plane (see, e.g., [43]). Clearly, this result casts doubts on the reliability of the 3-D model used.

### Future Directions

Seismic tomography began about thirty years ago [2,4,33] and since then the field has grown steadily. However, because of the intrinsic difficulties of wave propagation in the earth and its computer simulation, the ill-posedness of the inverse problem that must be solved (particularly at the global scale), and very limited funding, progress has been slow. Fortunately, in spite of these obstacles the point has been reached where it is possible to establish with some certainty which features of the internal composition of the earth are well resolved and which ones are not, and what needs to be done to improve the existing knowledge (see, e.g. [16,41,129,135,150]). It is clear that progress will come from several fronts and will be fueled by the steady decline in the price of computers and their increased power as well as the availability of relatively inexpensive PC clusters. This development will allow the generation of more realistic synthetic seismograms (see, e.g. [81]) and the computation of more accurate inverse



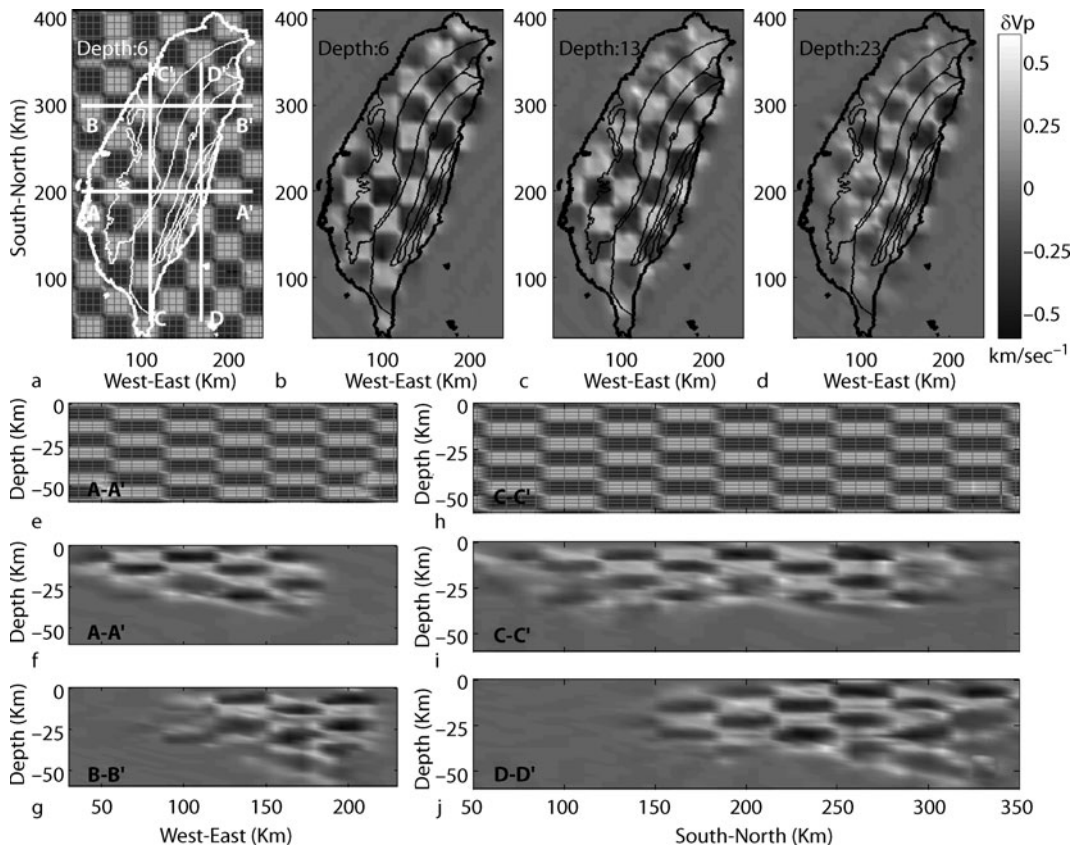
Tomography, Seismic, Figure 11

Cross-sectional views of the Taiwan 3-D  $P$  wave velocity model. The 7.8 km/s contour line (*black line*) marks the depth to the Moho (approximately). The cross section locations are shown in a. The *bold lines* below 0 depth indicate the 7.8 km/s contour line. The *bold lines* above 0 depth indicate elevation (with a vertical exaggeration of 3). From [78]

solutions as well as the computation of resolution and covariance matrices. The integration of tomographic, geodynamic, and mineral physics models will lead, hopefully, to a better understanding of the earth's interior and the processes therein. However, because oceans cover two-thirds of the earth-surface, it will be necessary to make progress in the deployment of ocean-bottom seismographs before these goals can be achieved fully.

## Bibliography

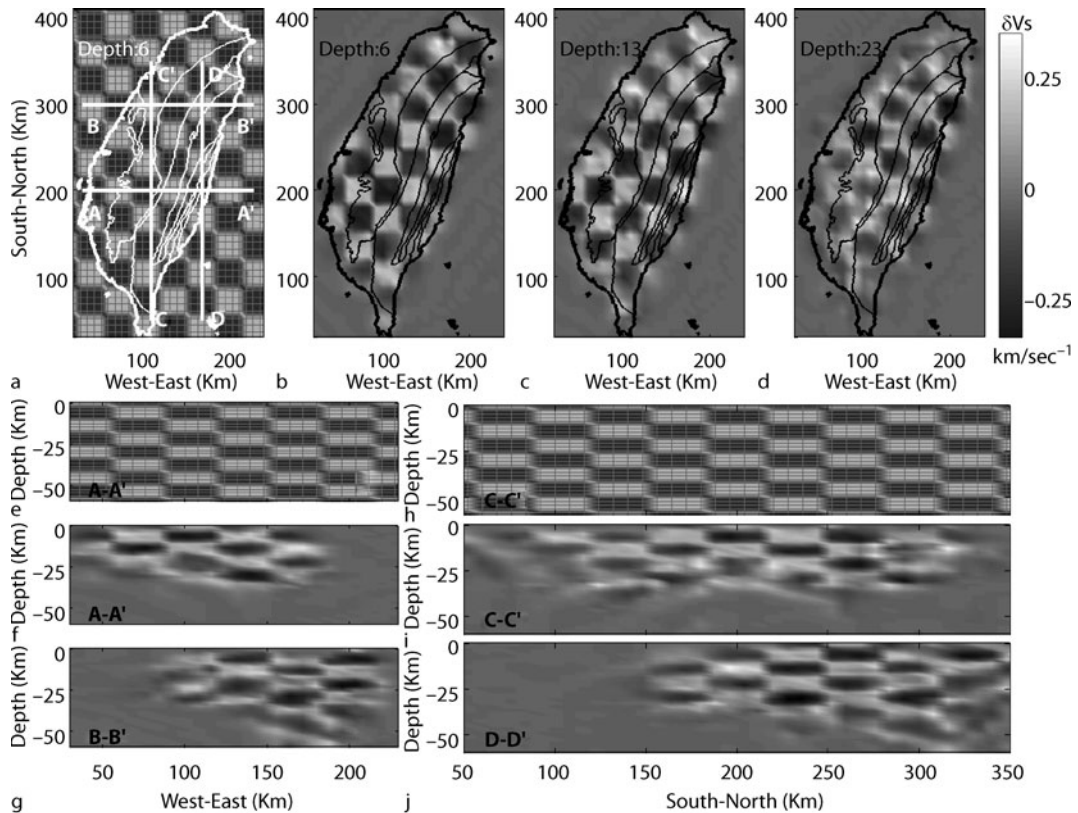
1. Aki K (1993) Overview. In: Iyer H, Hirahara K (eds) *Seismic tomography*. Chapman, London, pp 1–8
2. Aki K, Lee W (1976) Determination of three-dimensional velocity anomalies under a seismic array using first  $P$  arrival times from local earthquakes, 1. A homogeneous initial model. *J Geophys Res* 81:4381–4399
3. Aki K, Richards P (1980) *Quantitative seismology*, vol 2. Freeman, San Francisco
4. Aki K, Christofferson A, Husebye E (1977) Determination of the three-dimensional seismic structure of the lithosphere. *J Geophys Res* 82:277–296
5. Allen M, Isaacson E (1998) *Numerical analysis for applied science*. Wiley, New York
6. Ambrose J (1973) Computerized transverse axial scanning (tomography): Part 2. *Clin Appl Br J Radiol* 46:1023–1047
7. Bai C-Y, Greenhalgh S (2005) 3-D multi-step travel time tomography: imaging the local, deep velocity structure of Rabaul volcano, Papua New Guinea. *Phys Earth Planet Inter* 151:259–275
8. Bai C-Y, Greenhalgh S (2006) 3D local earthquake hypocenter determination with an irregular shortest-path method. *Bull Seism Soc Am* 96:2257–2268
9. Bai C-Y, Greenhalgh S, Zhou B (2007) 3D ray tracing using a modified shortest-path method. *Geophysics* 72(4):T27–T36
10. Bard Y (1974) *Nonlinear parameter estimation*. Academic Press, New York
11. Barret H, Hawkins W, Joy M (1983) Historical note on computed tomography. *Radiology* 147:172
12. Bates R, Peters T (1971) Towards improvements in tomography. *NZ J Sci* 14:883–896
13. Beck J, Arnold K (1977) *Parameter estimation in engineering and science*. Wiley, New York
14. Benz H, Smith R (1984) Simultaneous inversion for lateral velocity variations and hypocenters in the Yellowstone region using earthquake and refraction data. *J Geophys Res* 89:1208–1220
15. Benz H, Chouet B, Dawson P, Lahr J, Page R, Hole J (1996) Three-dimensional  $P$  and  $S$  wave velocity structure of Re-doubt Volcano, Alaska. *J Geophys Res* 101:8111–8128
16. Boschi L, Ampuero J-P, Peter D, Mai P, Soldati G, Giardini D (2007) Petascale computing and resolution in global seismic tomography. *Phys Earth Planet Inter* 163:245–250



Tomography, Seismic, Figure 12

Results of the checkerboard test for the Taiwan  $P$  wave velocity model. The checkerboard pattern is shown in a, e and h. The velocity variations across the block boundaries is 0.6 km/s. Synthetic arrival times were generated with this model and the earthquake and station locations used in the inversion of the actual data. The results of the inversion of the synthetic data are shown in map view for different depths in b–d. Cross sectional views are shown in f, g, i, and j. From [78]

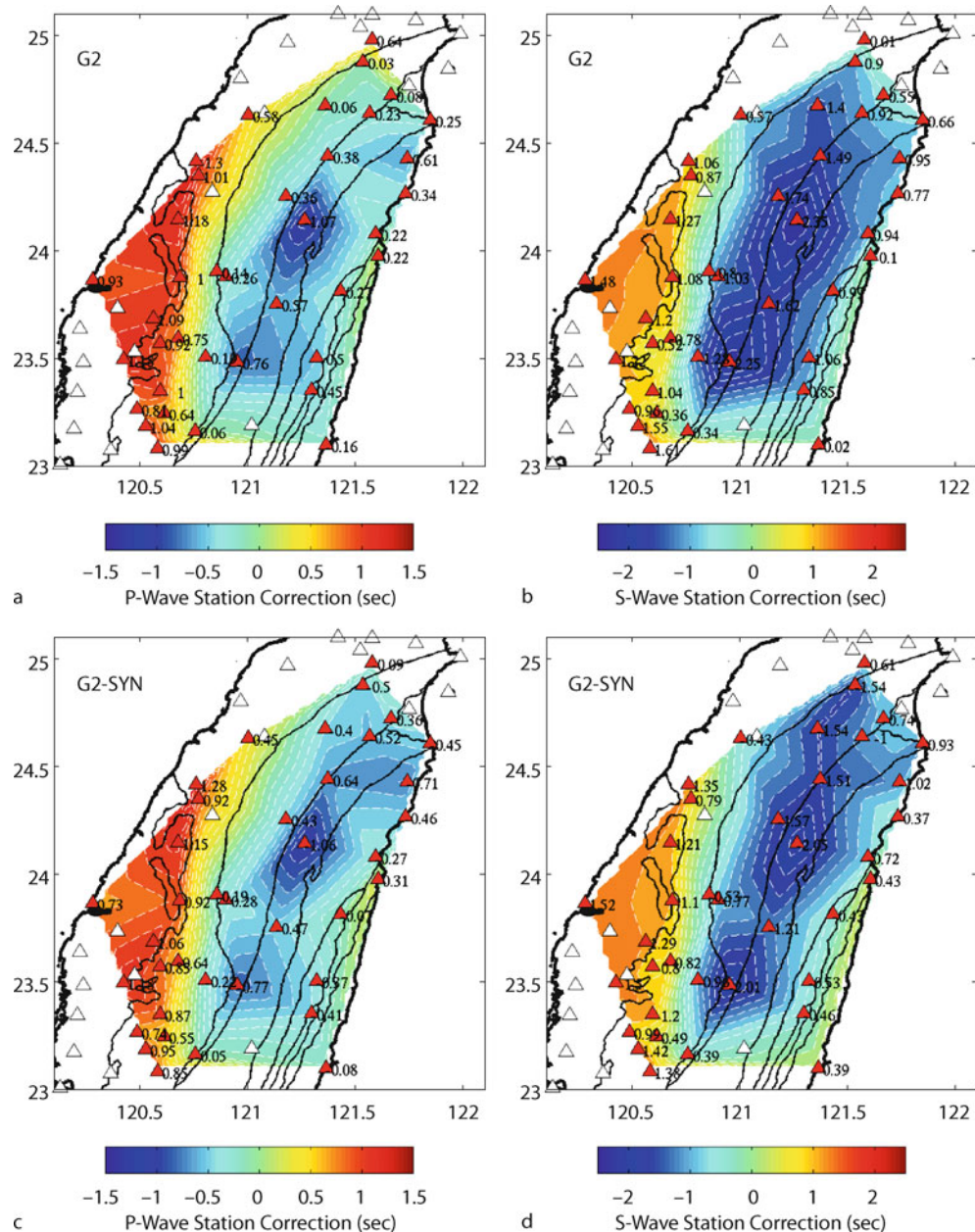
17. Bracewell R (1956) Strip integration in radio astronomy. *Aust J Phys* 9:198–217
18. Bracewell R, Riddle A (1967) Inversion of fan-beam scans in radio astronomy. *J Astrophys* 150:427–434
19. Broad W (1980) Riddle of the Nobel debate. *Science* 207: 37–38
20. Brooks R, Di Chiro G (1976) Principles of computer assisted tomography (CAT) in radiographic and radioisotopic imaging. *Phys Med Biol* 5:689–732
21. Červený V (2001) *Seismic ray theory*. Cambridge University Press, Cambridge
22. Červený V, Molotkov I, Pšenčík I (1977) *Ray method in seismology*. Charles University, Prague
23. Chapman C (1987) The Radon transform and seismic tomography. In: Nolet G (ed) *Seismic tomography*. Reidel, Dordrecht, pp 25–47
24. Chen H, Chiu J-M, Pujol J, Kim K, Chen K-C, Huang B-S, Yeh Y-H, Chiu S-C (2006) A simple algorithm for local earthquake location using 3D  $V_P$  and  $V_S$  models: test examples in the central United States and in central eastern Taiwan. *Bull Seis Soc Am* 96:288–305
25. Claerbout J (1985) *Imaging the Earth's Interior*. Blackwell Scientific Publications, Boston
26. Cormack A (1963) Representation of a function by its line integrals, with some radiological applications. *J Appl Phys* 34:2722–2727
27. Cormack A (1964) Representation of a function by its line integrals, with some radiological applications, II. *J Appl Phys* 35:2908–2913
28. Cormack A (1973) Reconstruction of densities from their projections, with applications in radiological physics. *Phys Med Biol* 18:195–207
29. Cormack A (1980) Recollections of my work with computer assisted tomography. *Mol Cell Biochem* 32:57–61
30. Cormack A (1982) Computed tomography: some history and recent developments. *Proc Symp Appl Math* 27:35–42
31. Creager K (1984) *Geometry, velocity structure, and penetration depths of descending slabs in the western Pacific*. Ph D dissertation, University of California, San Diego
32. Creager K, Boyd T (1992) Effects of earthquake mislocation on estimates of velocity structure. *Phys Earth Planet Inter* 75: 63–76



Tomography, Seismic, Figure 13

Similar to Fig. 12 for the S wave velocity model. The velocity variations across box boundaries is 0.346 km/s. From [78]

33. Crosson R (1976) Crustal structure modeling of earthquake data. 1. Simultaneous least squares estimation of hypocenter and velocity parameters. *J Geophys Res* 81:3036–3046
34. Crowther R, DeRosier D, Klug A (1970) The reconstruction of a three-dimensional structure from projections and its application to electron microscopy. *Proc R Soc Lond Ser A* 317:319–340
35. Deans S (1983) *The Radon transform and some of its applications*. Wiley, New York
36. DeRosier D, Klug A (1968) Reconstruction of three dimensional structures from electron micrographs. *Nature* 217: 130–134
37. Dietz L, Ellsworth W (1990) The October 17, 1989 Loma Prieta, California, earthquake and its aftershocks: Geometry of the sequence from high-resolution locations. *Geophys Res Lett* 17:1417–1420
38. Dietz L, Ellsworth W (1997) Aftershocks of the Loma Prieta earthquake and their tectonic implications. In: P Reasenberg (ed) *The Loma Prieta, California, earthquake of October 17, 1989 – Aftershocks and postseismic effects*. US Geol Surv Prof Pap 1550-D, D5-D47
39. Dines K, Lytle R (1979) Computerized geophysical tomography. *Proc Inst Electr Electron Eng* 67:1065–1073
40. Durrani T, Bisset D (1984) The Radon transform and its properties. *Geophysics* 49:1180–1187; Errata, 1985, 50:884–886
41. Dziewonski A (2003) Global seismic tomography: What we really can say and what we make up. *Geol Soc Am Penrose Conference, Plume IV: Beyond the Plume Hypothesis, Abstracts* (available at: [www.mantleplumes.org/Penrose/PenPDFAbstracts/Dziewonski\\_Adam\\_abs.pdf](http://www.mantleplumes.org/Penrose/PenPDFAbstracts/Dziewonski_Adam_abs.pdf))
42. Eberhart-Phillips D (1986) Three-dimensional velocity structure in northern California Coast Ranges from inversion of local earthquake arrival times. *Bull Seism Soc Am* 76: 1025–1052
43. Eberhart-Phillips D, Stuart W (1992) Material heterogeneity simplifies the picture: Loma Prieta. *Bull Seism Soc Am* 82:1964–1968
44. Eliseevnin V (1965) Analysis of rays propagating in an inhomogeneous medium. *Sov Phys Acoust* 10:242–245
45. Evans J, Achauer U (1993) Teleseismic velocity tomography using the ACH method: theory and application to continental-scale studies. In: Iyer H, Hirahara K (eds) *Seismic tomography*. Chapman, London, pp 319–360
46. Forsythe G, Malcolm M, Moler C (1977) *Computer methods for mathematical computations*. Prentice-Hall, Englewood Cliffs
47. Franklin J (1970) Well-posed extensions of ill-posed linear problems. *J Math Anal Appl* 31:682–716
48. Gilbert P (1972) Iterative methods for the three-dimensional reconstruction of an object from projections. *J Theor Biol* 36:105–117

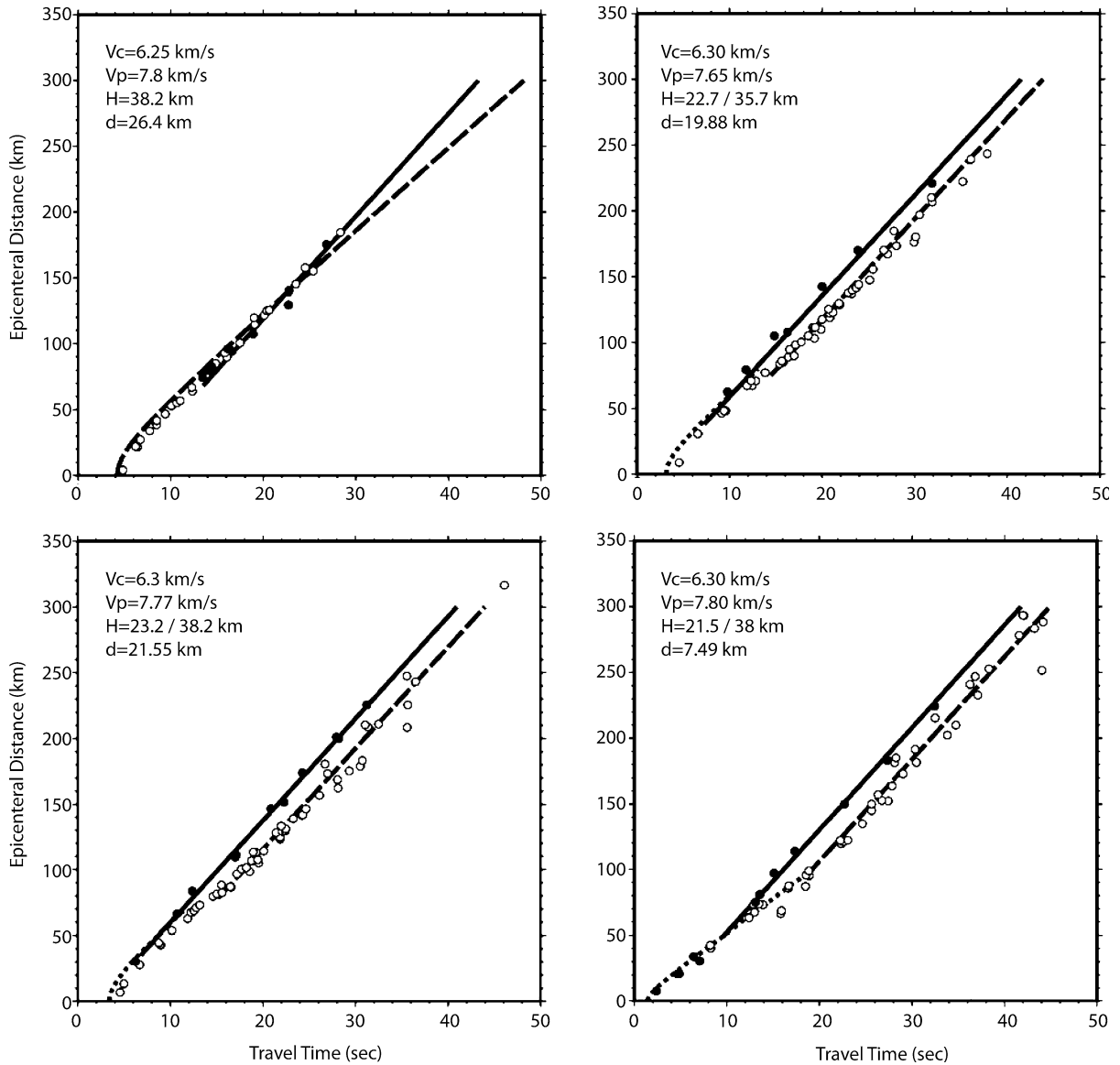


Tomography, Seismic, Figure 14

*Top:* P and S wave station corrections determined using the joint hypocentral determination technique (JHD) and the data in box G2 in Fig. 10. *Bottom:* JHD station corrections obtained using synthetic data generated with the 3-D velocity model and locations determined by tomographic inversion. From [78]

49. Gill P, Murray W, Wright M (1981) Practical optimization. Academic Press, London
50. Gordon R (1974) A tutorial on ART. Inst Electr Electron Eng Trans Nucl Sci NS-21:78–93
51. Gordon R, Bender R, Herman G (1970) Algebraic reconstruction techniques (ART) for three-dimensional electron microscopy and X-ray photography. J Theor Biol 29:471–481
52. Grand S (1987) Tomographic inversion for shear velocity beneath the North American plate. J Geophys Res 92:14065–14090
53. Groetsch C (1993) Inverse problems in the mathematical sciences. Vieweg, Braunschweig
54. Gubbins D (1981) Source location in laterally varying media. In: Husebye E, Mykkeltveit S (eds) Identification of seismic



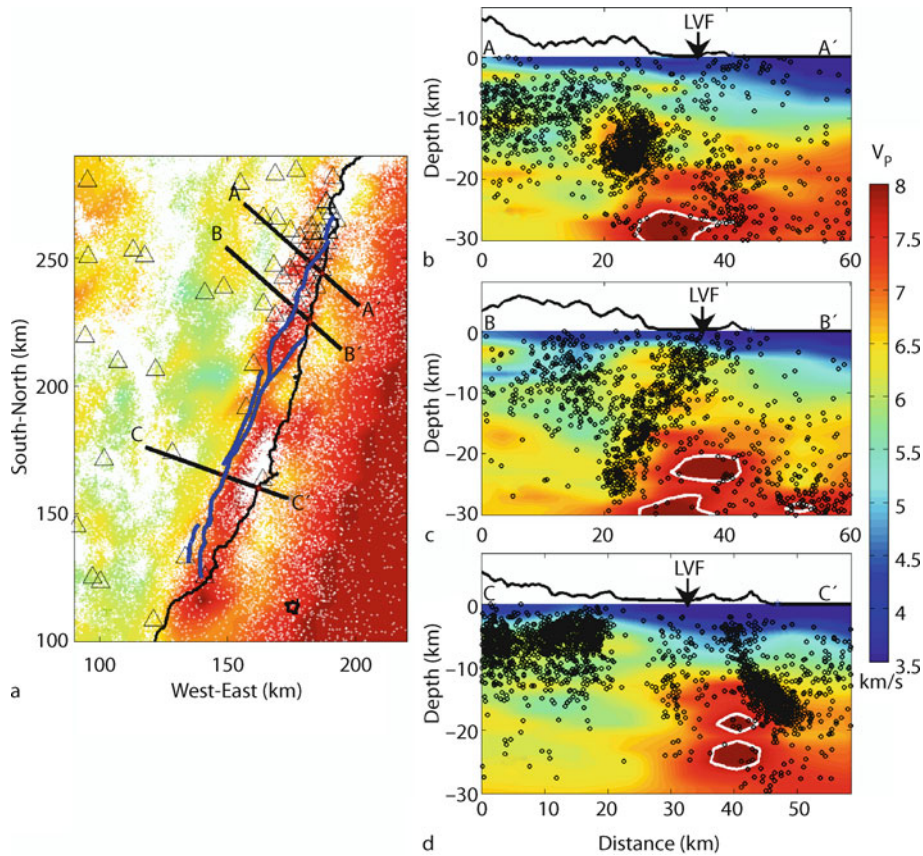


Tomography, Seismic, Figure 15

Epicentral distance versus travel time plots for the four earthquakes in Fig. 9. The plot in the *upper left corner* corresponds to the event numbered 1. The *solid* and *dashed* lines are best-fit lines corresponding to the anomalous and normal Pn waves. For events 2–4 (along Taiwan's east coast) the anomalous Pn waves are seen in the eastern stations.  $V_c$ : crustal velocity;  $V_p$ : mantle velocity;  $H$ : depth(s) to the Moho;  $d$ : event depth. From [89]

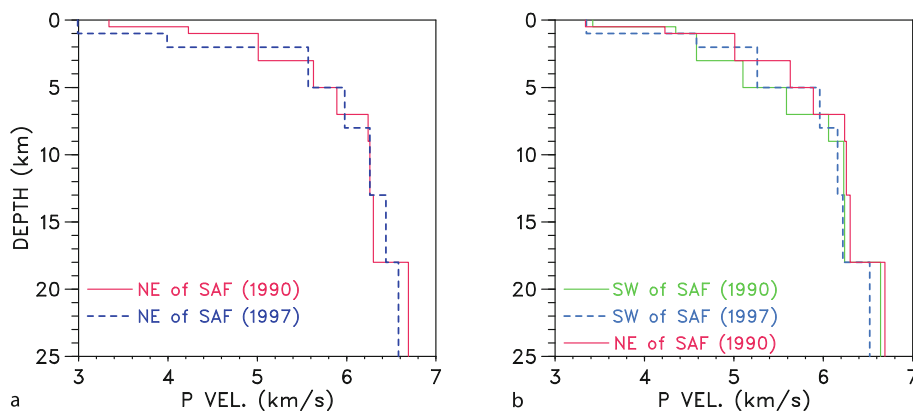
sources – Earthquake or underground explosion. Reidel, Dordrecht, pp 543–573

55. Hansen P (1992) Analysis of discrete ill-posed problems by means of the L-curve. SIAM Rev 34:561–580
56. Hansen P (1994) Regularization Tools: A Matlab package for analysis and solution of discrete ill-posed problems. Num Algorithms 6:1–35; (Software available at: <http://www2.imm.dtu.dk/~pch/Regutools/>)
57. Hanson K (1987) Bayesian and related methods in image reconstruction from incomplete data. In: Stark H (ed) Image recovery: theory and applications. Academic, Orlando, pp 79–125
58. Hauksson E (2000) Crustal structure and seismicity distribution adjacent to the Pacific and North America plate boundary in southern California. J Geophys Res 105:13875–13903
59. Hauksson E, Haase J (1997) Three-dimensional  $V_p$  and  $V_p/V_s$  velocity models of the Los Angeles basin and central Transverse Ranges, California. J Geophys Res 102:5423–5453



Tomography, Seismic, Figure 16

*Left:* Map view of the Taiwan  $P$  wave velocity model at a depth of 20 km (from [78]). The coordinate system is as in Fig. 11a. *Blue lines:* surface trace of active faults. *White dots:* epicentral locations. *Right:* Cross-sectional views of  $P$  wave velocity across the Longitudinal Valley (LVF) and seismicity. The locations of the cross sections are shown on the left. Note the elevated high-velocity oceanic upper mantle along the entire collision suture zone. The 7.8 km/s contour is indicated by the white line around the dark-red areas. From [89]



Tomography, Seismic, Figure 17

$P$  wave velocity models used in [37] and [38] to locate the 1989,  $M = 7.1$ , Loma Prieta, California, earthquake and its aftershocks. *Left:* Models for stations to the NE of the San Andreas fault. *Right:* Models for stations to the SW of the San Andreas fault. For a comparison, the NE 1990 model is also shown. After [119]

60. Hauksson E, Jones L, Hutton K (1995) The 1994 Northridge earthquake sequence in California: seismological and tectonic aspects. *J Geophys Res* 100:12335–12355
61. Hawley B, Zandt G, Smith R (1981) Simultaneous inversion for hypocenters and lateral velocity variations: an iterative solution with a layered model. *J Geophys Res* 86:7073–7086
62. Herman G (1980) *Image reconstruction from projections*. Academic Press, New York
63. Herman G, Lent A (1976) Iterative reconstruction algorithms. *Comput Biol Med* 6:273–294
64. Herman G, Hurwitz H, Lent A, Lung H-P (1979) On the Bayesian approach to image reconstruction. *Inform Contr* 42:60–71
65. Hounsfield G (1973) Computerized transverse axial scanning (tomography): Part 1. Description of system. *J Br Radiol* 46:1016–1022
66. Hounsfield G (1979) Computed medical imaging, Nobel Lecture. (available at: [nobelprize.org/nobel\\_prizes/medicine/laureates/1979/hounsfield-lecture.pdf](http://nobelprize.org/nobel_prizes/medicine/laureates/1979/hounsfield-lecture.pdf))
67. Hounsfield G (1980) Autobiography. In: Wilhelm O (ed) *The Nobel Prizes 1979*. The Nobel Foundation, Stockholm (available at: [nobelprize.org/nobel\\_prizes/medicine/laureates/1979/hounsfield-autobio.html](http://nobelprize.org/nobel_prizes/medicine/laureates/1979/hounsfield-autobio.html))
68. Hudnut K et al (1996) Co-seismic displacements of the 1994 Northridge, California, earthquake. *Bull Seism Soc Am* 86(1B):S19–S36
69. Hudson J (1980) *The excitation and propagation of elastic waves*. Cambridge University Press, Cambridge
70. Humphreys E, Clayton R (1988) Adaptation of back projection tomography to seismic travel times problems. *J Geophys Res* 93:1073–1085
71. Hurwitz H (1975) Entropy reduction in Bayesian analysis of measurements. *Phys Rev A* 12:698–706
72. Inoue H, Fukao Y, Tanabe K, Ogata Y (1990) Whole mantle P-wave travel time tomography. *Phys Earth Planet Inter* 59: 294–328
73. Ivansson S (1983) Remark on an earlier proposed iterative tomographic algorithm. *J Geophys R Astr Soc* 75:855–860
74. Jackson D (1979) The use of a priori data to resolve non-uniqueness in linear inversion. *J Geophys R Astr Soc* 57: 137–157
75. Julian B, Gubbins D (1977) Three-dimensional seismic ray tracing. *J Geophys* 43:95–113
76. Kak A, Slaney M (1988) *Principles of computerized tomographic imaging*. Inst Electr Electron Eng Press, New York
77. Kim K-H (2003) Subsurface structure, seismicity patterns, and their implication to tectonic evolution in Taiwan. Ph D dissertation, University of Memphis, Memphis
78. Kim K-H, Chiu J-M, Pujol J, Chen K-C, Huang B-S, Yeh Y-H, Shen P (2005) Three-dimensional  $V_P$  and  $V_S$  structural models associated with the active subduction and collision tectonics in the Taiwan region. *J Geophys Int* 162:204–220
79. Koch M (1985) Non-linear inversion of local seismic travel times for the simultaneous determination of 3D-velocity structure and hypocenters – application to the seismic zone Vrancea. *J Geophys* 56:160–173
80. Koch M (1993) Simultaneous inversion for 3-D crustal structure and hypocenters including direct, refracted and reflected phases – I Development, validation and optimal regularization of the method. *J Geophys Int* 112:385–412
81. Komatitsch D, Tsuboi S, Tromp J (2005) The spectral-element method in seismology. In: Levander A, Nolet G (eds) *Seismic earth: array analysis of broadband seismograms*. Geophysical Monograph Series, vol 157. Am Geophys Union, Washington DC, pp 205–227
82. Langenheim V, Griscom A, Jachens R, Hildenbrand T (2000) Preliminary potential-field constraints on the geometry of the San Fernando basin, southern California. US Geol Survey Open-File Report 00–219
83. Lawson C, Hanson R (1974) *Solving least squares problems*. Prentice-Hall, Englewood Cliffs
84. Lee W, Pereyra V (1993) *Mathematical introduction to seismic tomography*. In: Iyer H, Hirahara K (eds) *Seismic tomography*. Chapman, London, pp 9–22
85. Lee W, Stewart S (1981) *Principles and applications of microearthquake networks*. Academic Press, New York
86. Lees J, Crosson R (1989) Tomographic inversion for three-dimensional velocity structure at Mount St. Helens using earthquake data. *J Geophys Res* 94:5716–5728
87. Levenberg K (1944) A method for the solution of certain nonlinear problems in least squares. *Quart Appl Math* 2:164–168
88. Lewitt R (1983) Reconstruction algorithms: transform methods. *Proc Inst Electr Electron Eng* 71:390–408
89. Liang W-T, Chiu J-M, Kim K (2007) Anomalous Pn waves observed in eastern Taiwan: implications of a thin crust and elevated oceanic upper mantle beneath the active collision-zone suture. *Bull Seism Soc Am* 97:1370–1377
90. Ma K-F, Wang J-H, Zhao D (1996) Three-dimensional seismic velocity structure of the crust and uppermost mantle beneath Taiwan. *J Phys Earth* 44:85–105
91. Magistrale H, Day S, Clayton R, Graves R (2000) The SCEC southern California reference three-dimensional seismic velocity model version 2. *Bull Seism Soc Am* 90(6B):S65–S76
92. Marquardt D (1963) An algorithm for least-squares estimation of nonlinear parameters. *J Soc Ind Appl Math* 11:431–441
93. Martin M, Ritter J, CALIXTO Working Group (2005) High-resolution teleseismic body-wave tomography beneath SE Romania – I. Implications for three-dimensional versus one-dimensional crustal correction strategies with a new crustal velocity model. *J Geophys Int* 162:448–460
94. Meskò A (1984) *Digital filtering: applications in geophysical exploration for oil*. Wiley, New York
95. Monna S, Filippi L, Beranzoli L, Favali P (2003) Rock properties of the upper-crust in Central Apennines (Italy) derived from high-resolution 3-D tomography. *Geophys Res Lett* 30(61): 1–4 doi:10.1029/2002GL016780
96. Morelli A (1993) Teleseismic tomography: core-mantle boundary. In: Iyer H, Hirahara K (eds) *Seismic tomography*. Chapman, London, pp 163–189
97. Mori J, Wald D, Wesson R (1995) Overlapping fault planes of the (1971) San Fernando and 1994 Northridge, California earthquakes. *Geophys Res Lett* 22:1033–1036
98. Moser T (1991) Shortest path calculation of seismic rays. *Geophysics* 56:59–67
99. Moser T, Nolet G, Snieder R (1992) Ray bending revisited. *Bull Seismol Soc Am* 82:259–288
100. Moser T, Van Eck T, Nolet G (1992) Hypocenter determination in strongly heterogeneous earth models using the shortest path method. *J Geophys Res* 97:6563–6572
101. Nakanishi I, Yamaguchi K (1986) A numerical experiment on nonlinear image reconstruction from first-arrival times for two-dimensional island arc structure. *J Phys Earth* 34:195–201

102. Nelson G, Vidale J (1990) Earthquake locations by 3-D finite-difference travel times. *Bull Seism Soc Am* 80:395–410
103. Noble B, Daniel J (1977) *Applied linear algebra*. Prentice-Hall, Englewood Cliffs
104. Nolet G (1993) Solving large linearized tomographic problems. In: Iyer H, Hirahara K (eds) *Seismic tomography*. Chapman, London, pp 227–247
105. Okubo P, Benz H, Chouet B (1997) Imaging the crustal magma sources beneath Mauna Loa and Kilauea Volcanoes, Hawaii. *Geology* 25:867–870
106. Oldendorf W (1961) Isolated flying spot detection of radio density discontinuities – Displaying the internal structural pattern of a complex object. *IRE Trans Biomed Elec BME-8*: 68–72
107. Oransky I (2004) Obituary. Sir Godfrey N Hounsfield. *Lancet* 364:1032
108. Paige C, Saunders M (1982) LSQR: An algorithm for sparse linear equations and sparse least square problems. *ACM Trans Math Softw* 8:43–71
109. Parker R (1994) *Geophysical inverse theory*. Princeton University Press, Princeton
110. Pavlis G, Booker J (1980) The mixed discrete-continuous inverse problem: application to the simultaneous determination of earthquake hypocenters and velocity structure. *J Geophys Res* 85:4801–4810
111. Penrose R (1955) A generalized inverse for matrices. *Proc Camb Phil Soc* 51:406–413
112. Pereyra V, Lee W, Keller H (1980) Solving two-point seismic-ray tracing problems in a heterogeneous medium. *Bull Seism Soc Am* 70:79–99
113. Podvin P, Lecomte I (1991) Finite difference computation of traveltimes in very contrasted velocity models: a massively parallel approach and its associated tools. *J Geophys Int* 105:271–284
114. Press W, Teukolsky S, Vetterling W, Flannery B (1992) *Numerical Recipes*. Cambridge University Press, Cambridge
115. Prothero W, Taylor W, Eickemeyer J (1988) A fast, two-point, three-dimensional raytracing algorithm using a simple step search method. *Bull Seism Soc Am* 78:1190–1198
116. Pujol J (1995) Application of the JHD technique to the Loma Prieta, California, mainshock-aftershock sequence and implications for earthquake location. *Bull Seism Soc Am* 85: 129–150
117. Pujol J (1996) Comment on: “The 1989 Loma Prieta earthquake imaged from inversion of geodetic data” by Thora Árnadóttir and Paul Segall. *J Geophys Res* 101:20133–20136
118. Pujol J (1996) An integrated 3D velocity inversion – joint hypocentral determination relocation analysis of events in the Northridge area. *Bull Seism Soc Am* 86(1B):S138–S155
119. Pujol J (2000) Joint event location – The JHD technique and applications to data from local seismic networks. In: Thurber C, Rabinowitz N (eds) *Advances in seismic event location*. Kluwer, Dordrecht, pp 163–204
120. Pujol J (2003) *Elastic wave propagation and generation in seismology*. Cambridge University Press, Cambridge
121. Pujol J (2007) The solution of nonlinear inverse problems and the Levenberg-Marquardt method. *Geophysics* 72(4): W1–W16
122. Pujol J et al (1989) 3-D P- and S-wave velocity structure of the Andean foreland in San Juan, Argentina, from local earthquakes. *Eos Trans Am Geoph Union* 70(43):1213
123. Pujol J, Mueller K, Peng S, Chitupolu V (2006) High-resolution 3D P-wave velocity model for the East Ventura–San Fernando basin, California, and relocation of events in the Northridge and San Fernando aftershock sequences. *Bull Seism Soc Am* 96:2269–2280
124. Ramachandran G, Lakshminarayanan A (1971) Three-dimensional reconstruction from radiographs and electron micrographs: application of convolutions instead of Fourier transforms. *Proc Natl Acad Sci USA* 68:2236–2240
125. Ratchkovsky N, Pujol J, Biswas N (1997) Relocation of earthquakes in the Cook Inlet area, south central Alaska, using the joint hypocenter determination method. *Bull Seism Soc Am* 87:620–636
126. Rau R-J, Wu F (1995) Tomographic imaging of lithospheric structures under Taiwan. *Earth Planet Lett* 133:517–532
127. Robinson E (1982) Spectral approach to geophysical inversion by Lorentz, Fourier, and Radon transforms. *Proc Inst Electr Electron Eng* 70:1039–1054
128. Roecker S, Yeh Y, Tsai Y (1987) Three-dimensional P and S wave velocity structures beneath Taiwan: deep structure beneath an arc-continent collision. *J Geophys Res* 92:10547–10570
129. Romanowicz B (2003) Global mantle tomography: Progress status in the past 10 years. *Annu Rev Earth Planet Sci* 31: 303–328
130. Sage A, Melsa J (1971) *Estimation theory with applications to communications and control*. McGraw-Hill, New York
131. Sandoval S, Kissling E, Anson J, Svekopalapko Seismic Tomography Working Group (2003) High-resolution body wave tomography beneath the Svekopalapko array: I, A priori three-dimensional crustal model and associated traveltime effects on teleseismic wave fronts. *J Geophys Int* 153:75–87
132. Shepp L, Kruskal J (1978) Computerized tomography: the new medical X-ray technology. *Am Math Mon* 85:420–439
133. Shepp L, Logan B (1974) The Fourier reconstruction of a head section. *Inst Electr Electron Eng Trans Nucl Sci NS-21*: 21–43
134. Snoko J, Lahr J (2001) Locating earthquakes: at what distances can the Earth no longer be treated as flat? *Seism Res Lett* 72:538–541
135. Soldati G, Boschi L, Piersanti A (2006) Global seismic tomography and modern parallel computers. *Ann Geophys* 49: 977–986
136. Sorenson H (1980) *Parameter estimation: principles and problems*. Dekker, New York
137. Spakman W (1993) Iterative strategies for non-linear travel time tomography using global earthquake data. In: Iyer H, Hirahara K (eds) *Seismic tomography*. Chapman, London, pp 190–226
138. Stanton L (1969) *Basic medical radiation physics*. Appleton-Century-Crofts, New York
139. Stewart J, Choi Y, Graves R, Shaw J (2005) Uncertainty of southern California basin depth parameters. *Bull Seism Soc Am* 95:1988–1993
140. Süss M, Shaw J (2003) P wave seismic velocity structure derived from sonic logs and industry reflection data in the Los Angeles basin, California. *J Geophys Res* 108(13):1–18 doi:10.1029/2001JB001628
141. Tarantola A, Valette B (1982) Inverse problems = quest for information. *J Geophys* 50:159–170

142. Tarantola A, Valette B (1982) Generalized nonlinear inverse problems solved using the least squares criterion. *Rev Geophys Space Phys* 20:219–232
143. Teng T-L, Aki K (1996) Preface to the 1994 Northridge earthquake special issue. *Bull Seism Soc Am* 86(1B):S1–S2
144. Thurber C (1983) Earthquake locations and three-dimensional crustal structure in the Coyote Lake area, central California. *J Geophys Res* 88:8226–8236
145. Thurber C (1992) Hypocenter-velocity structure coupling in local earthquake tomography. *Phys Earth Planet Inter* 75:55–62
146. Thurber C (1993) Local earthquake tomography: velocities and  $V_p/V_s$  – theory. In: Iyer H, Hirahara K (eds) *Seismic tomography*. Chapman, London, pp 563–583
147. Thurber C, Kissling E (2000) Advances in travel-time calculations for three-dimensional structures. In: Thurber C, Rabinowitz N (eds) *Advances in seismic event location*. Kluwer, Dordrecht, pp 71–99
148. Tihonov A (1963) Solution of incorrectly formulated problems and the regularization method. *Sov Math* 4:1035–1038; (Note: a more common transliteration of this author's Russian name is Tikhonov.)
149. Titchmarsh W (1948) *Introduction to the theory of Fourier integrals*. Oxford University Press, Oxford
150. Trampert J, Van der Hilst R (2005) Towards a quantitative interpretation of global seismic tomography. In: Van der Hilst R, Bass J, Matas J, Trampert J (eds) *Earth's deep mantle: structure, composition, and evolution*. Geophysical Monograph Series, vol 160. Am Geophys Union, Washington DC, pp 47–62
151. Tryggvason A, Bergman B (2006) A travelttime reciprocity discrepancy in the Podvin & Lecomte time3d finite difference algorithm. *J Geophys Int* 165:432–435
152. Um J, Thurber C (1987) A fast algorithm for two-point seismic ray tracing. *Bull Seism Soc Am* 77:972–986
153. Van der Hilst R, Engdahl E (1992) Step-wise relocation of ISC earthquake hypocenters for linearized tomographic imaging of slab structure. *Phys Earth Planet Inter* 75:39–53
154. Van der Sluis A, Van der Vorst H (1987) Numerical solution of large, sparse linear algebraic systems arising from tomographic problems. In: Nolet G (ed) *Seismic tomography*. D Reidel, Dordrecht, pp 49–83
155. Van Tiggelen R (2002) In search for the third dimension: from radiostereoscopy to three-dimensional imaging. *JBR-BTR* 85:266–270
156. Van Tiggelen R, Pouders E (2003) Ultrasound and computed tomography: spin-offs of the World Wars. *JBR-BTR* 86:235–241
157. Vidale J (1988) Finite-difference calculation of travel times. *Bull Seism Soc Am* 78:2062–2076
158. Vidale J (1990) Finite-difference calculation of travel times in three dimensions. *Geophysics* 55:521–526
159. Villaseñor A, Benz H, Filippi L, De Luca G, Scarpa R, Patanè G, Vinciguerra S (1998) Three-dimensional *P*-wave velocity structure of Mt. Etna, Italy. *Geophys Res Lett* 25:1975–1978
160. Wald D, Graves R (1998) The seismic response of the Los Angeles basin, California. *Bull Seism Soc Am* 88:337–356
161. Woodhouse J, Dziewonski A (1989) Seismic modelling of the Earth's large-scale three-dimensional structure. *Phil Trans Roy Soc Lond A* 328:291–308
162. Zhao D, Kanamori H (1995) The 1994 Northridge earthquake: 3-D crustal structure in the rupture zone and its relation to the aftershock locations and mechanisms. *Geophys Res Lett* 22:763–766
163. Zhao D, Hasegawa A, Horiuchi S (1992) Tomographic imaging of *P* and *S* wave velocity structure beneath northeastern Japan. *J Geophys Res* 97:19909–19928
164. Zhdanov M (2002) *Geophysical inverse theory and regularization problems*. Elsevier, Amsterdam

## Tsunami Earthquakes

JASCHA POLET<sup>1</sup>, H. KANAMORI<sup>2</sup>

<sup>1</sup> Geological Sciences Department, California State Polytechnic University, Pomona, USA

<sup>2</sup> Seismological Laboratory, Caltech, Pasadena, USA

### Article Outline

Glossary

Definition of the Subject

Introduction

Characteristics of Tsunami Earthquakes

Factors Involved in the Seismogenesis and Tsunamigenesis of Tsunami Earthquakes

A Model for Tsunami Earthquakes

Future Directions

Bibliography

### Glossary

**$m_b$**  body wave magnitude, based on the amplitude of the direct P wave, period of the measurement: 1.0–5.0 s. Also see: Seismic Magnitude.

**$M_S$**  surface wave magnitude, based on the amplitude of surface waves, period of the measurement: 20 s. Also see: Seismic Magnitude.

**$M_w$**  moment magnitude, determined from the seismic moment of an earthquake, typical period of the measurement: > 200 s. Also see: Seismic Magnitude.

**Magnitude saturation** due to the shape of the seismic source spectrum, relatively short period measurements of seismic magnitude will produce similar magnitudes for all earthquakes above a certain size. The value of this threshold earthquake size depends on the period of the measurement: magnitude measurements using shorter period waves will saturate at lower values than magnitude measurements using longer period waves.  $M_w$  will not saturate.

**Run-up height** difference between the elevation of maximum tsunami penetration (inundation line) and the sea level at the time of the tsunami.

**Tsunami earthquake** an earthquake that directly causes a regional and/or teleseismic tsunami that is greater in amplitude than would be expected from its seismic moment magnitude.

**Tsunami magnitude** a scale for the relative size of tsunamis generated by different earthquakes,  $M_t$  in particular is calculated from the logarithm of the maximum amplitude of the tsunami wave measured by a tide gauge distant from the tsunami source, corrected

for the distance to the source (also see: Satake, this volume).

**Seismic magnitude** a scale for the relative size of earthquakes. Many different scales have been developed, almost all based on the logarithmic amplitude of a particular seismic wave on a particular type of seismometer, with corrections for the distance between source and receiver. These measurements are made for different wave types at different frequencies, and thus may lead to different values for magnitude for any one earthquake.

**Seismic moment** the product of the fault surface area of the earthquake, the rigidity of the rock surrounding the fault and the average slip on the fault.

### Definition of the Subject

The original definition of “tsunami earthquake” was given by Kanamori [37] as “an earthquake that produces a large size tsunami relative to the value of its surface wave magnitude ( $M_S$ )”. Therefore, the true damage potential that a tsunami earthquake represents may not be recognized by conventional near real-time seismic analysis methods and may only become apparent upon the arrival of the tsunami waves on the local coastline. Although tsunami earthquakes occur relatively infrequently, the effect on the local population can be devastating, as was most recently illustrated by the July 2006 Java tsunami earthquake. This event (moment magnitude  $M_w = 7.8$ ) was quickly followed by tsunami waves two to seven meters high, traveling as far as two kilometers inland and killing at least 668 people ([http://www.searo.who.int/en/Section23/Section1108/Section2077\\_11956.htm](http://www.searo.who.int/en/Section23/Section1108/Section2077_11956.htm)).

It is important to note that the definition of “tsunami earthquake” is distinct from that of “tsunamigenic earthquake”. A tsunamigenic earthquake is any earthquake that excites a tsunami. Tsunami earthquakes are a specific subset of tsunamigenic earthquakes, which we will later in this article more precisely define as earthquakes that directly cause a regional and/or teleseismic tsunami that is greater in amplitude than would be expected from their seismic moment magnitude.

### Introduction

Shallow oceanic earthquakes may excite destructive tsunamis. Truly devastating tsunamis occur only infrequently, but as the natural disaster of the tsunami following the 2004 Sumatra–Andaman Islands earthquake has shown, may cause widespread damage (in this case in the region of the Indian Ocean) and lead to hundreds of thousands of casualties. In general, tsunamis are caused by shal-

low earthquakes beneath the ocean floor displacing large volumes of water. Thus, the magnitude of the earthquake plays an important role in determining its tsunamigenic potential. However, a particular subclass of shallow subduction zone earthquakes: “tsunami earthquakes”, poses a special problem.

For the purpose of this article, we will define the term “tsunami earthquake” as follows: an earthquake that directly causes a regional and/or teleseismic tsunami that is greater in amplitude than would be expected from its seismic moment magnitude. With this definition we specifically exclude seismic events that were followed by tsunamis directly caused by slides or slumps resulting from the original earthquake (as was the case for the 1992 Flores [28,32] and the 1998 Papua New Guinea earthquakes [22,73] for example). We further exclude events that only very locally caused large tsunamis as a result of, for example, focusing effects due to features of the ocean floor bathymetry (e.g. [63]) or directivity effects combined with the shape of the coastline, as was the case for the tsunamis that hit the harbor of Crescent City after the 1964 Aleutian Islands earthquake [12] and the November 15, 2006 Kurile Island event ([http://www.usc.edu/dept/tsunamis/california/Kuril\\_2006/](http://www.usc.edu/dept/tsunamis/california/Kuril_2006/)). Furthermore, this definition compares the size of the tsunami with the moment magnitude of the earthquake and not its surface wave magnitude, slightly modifying the definition given by Kanamori [37], in order to exclude great events for which the surface wave magnitude saturates.

Our primary objective in this article is to describe the characteristics of tsunami earthquakes and the possible factors involved in the anomalously strong excitation of tsunamis by these events. We will also discuss a possible model for these infrequent, but potentially very hazardous, events. The earthquakes listed in Table 1 and plotted in Fig. 1 are considered tsunami earthquakes, according to our modified definition presented in the previous paragraph, by the majority of the community of earthquake and tsunami researchers. However, we note that the interpretation of the 1994 Java and the 1946 Aleutian Islands earthquakes varies with investigators. The 1994 Java earthquake occurred off the southeastern coast of this island, near the east end of the Java Trench in the Indian Ocean, at 1:18 am local time. It generated a devastating tsunami that took the lives of more than 200 East Java coastal residents. Run-up measured along the southeastern Java coast ranged from 1 to 14 m, while run-up measured along the southwestern coast of Bali ranged from 1–5 m [74,82]. Although the anomalously high tsunami excitation of the 1994 event is not in doubt, its earthquake source characteristics have been debated [3,59]. The 1946

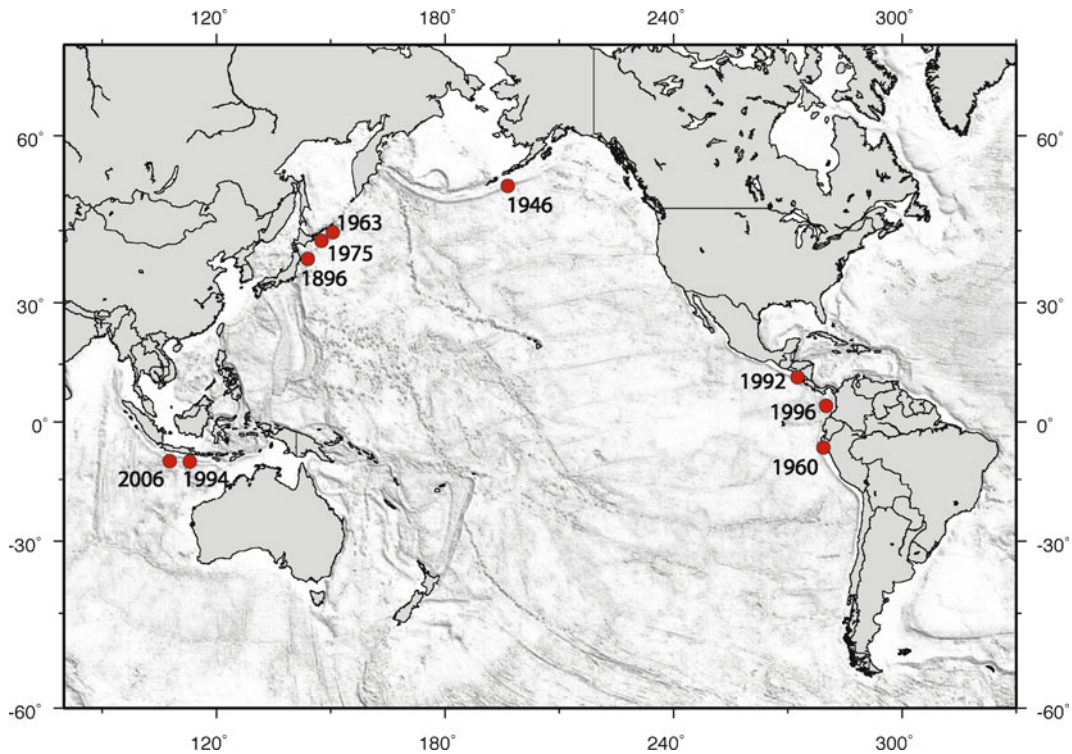
Tsunami Earthquakes, Table 1

Tsunami Earthquakes (references for most entries are listed in main text, others are from the National Geophysical Data Center Global Tsunami Database ([http://www.ngdc.noaa.gov/seg/hazard/tsu\\_db.shtml](http://www.ngdc.noaa.gov/seg/hazard/tsu_db.shtml)) and the Centennial Earthquake Catalog [18])

Date	Geographical Region	$M_w$	$m_b$	$M_S$	$M_t$	Deaths
1896/06/15	Japan			7.2	8.0	26360
1946/04/01	Aleutian Islands	8.2		7.3	9.3	165
1960/11/20	Peru	7.6	7.0	7.0		66
1963/10/20	Kurile Islands	7.8	7.1	7.2		
1975/06/10	Kurile Islands	7.5	5.6	7.0		
1992/09/02	Nicaragua	7.7	5.4	7.2		179
1994/06/02	Java	7.8	5.7	7.1		250
1996/02/21	Peru	7.5	5.8	6.6		12
2006/07/17	Java	7.7	6.2	7.2		668

Aleutian Islands earthquake off Unimak Island produced one of the largest trans-Pacific tsunamis and had a tsunami magnitude of 9.3 [1], but its moment magnitude is only  $M_w = 8.2$ , making it a tsunami earthquake [36]. Some of the great tsunami heights measured (exceeding 30 m in height on Unimak Island and 16 m in run-up at the Hawaiian islands; [78]) can be attributed to slumping [19]. However, its anomalously high tsunamis are probably primarily due to the seismic source directly [47].

The other, less controversial, tsunami earthquakes listed in our table are: the 1896 Sanriku event near the coast of Japan, two events near the Kurile Islands: one in 1963 and the other in 1975, the 1992 Nicaragua earthquake, the Peru earthquake 4 years later, as well as an earlier event in this region in 1960 and most recently the 2006 Java earthquake. The June 15, 1896 Sanriku earthquake generated devastating tsunamis with a maximum run-up of 25 m and caused the worst tsunami disaster in the history of Japan with over 20,000 deaths, despite its moderate surface wave magnitude ( $M_S = 7.2$ ) and weak seismic intensity [2,26,76]. The November 20, 1960 Peru earthquake excited a tsunami that was anomalously large for an earthquake of moderate magnitude [57], resulting in 66 fatalities (from the tsunami event database of the National Geophysical Data Center, <http://www.ngdc.noaa.gov/nndc/struts/form?t=101650&s=70&d=7>). The October 20, 1963 Kurile earthquake was an aftershock to the great Kurile Islands underthrusting earthquake ( $M_w = 8.5$ ) of October 13, 1963 and produced a maximum run-up height of 10–15 m at Urup Island, much larger than the height of the main shock tsunami of 5 m [2]. The 1975 earthquake occurred south of the Kurile Islands and was weakly felt along the entire southern part of the Kurile Islands. Like the 1963 tsunami earthquake, this event



Tsunami Earthquakes, Figure 1

Map of tsunami earthquakes (listed in Table 1). Location for 1896 earthquake from [2] and for 2006 earthquake from the Global CMT catalog. All other earthquake locations from the Centennial Earthquake Catalog [18]

can be considered an aftershock [21] of a larger event ( $M_S = 7.7$ ) that occurred essentially at the same location on June 17, 1973. The maximum run-up height was 5 m on Shikotan Island, while the main shock had a run-up height measured at 4.5 m. A fairly strong tsunami was also recorded on tide gauges in Alaska and Hawaii (from the tsunami event database of the National Geophysical Data Center <http://www.ngdc.noaa.gov/nndc/struts/form?t=101650&s=70&d=7>). After a time period of almost two decades without a significant tsunami earthquake, the 1992 Nicaragua earthquake was the first tsunami earthquake to be captured by modern broadband seismic networks. This tsunami caused 179 deaths (from the Emergency Disasters Database, <http://www.em-dat.net/disasters/list.php>) and significant damage to the coastal areas of Nicaragua and Costa Rica, reaching heights of up to 8 m. The 1996 Peru earthquake struck at 7:51 am local time, approximately 130 km off the northern coastal region of Peru. Approximately one hour after the main shock, a damaging tsunami reached the Peruvian coast, with run-up heights of 1 to 5 meters along a coastline of 400 km [27], resulting in twelve deaths [11]. Finally, the 2006 Java earthquake was located only about 600 km

west-northwest of the tsunami earthquake that occurred 12 years earlier in the same subduction zone. The Ministry of Health reported that approximately 668 people died and 65 are missing ([http://www.searo.who.int/en/Section23/Section1108/Section2077\\_11956.htm](http://www.searo.who.int/en/Section23/Section1108/Section2077_11956.htm)) due to a tsunami that had a maximum run-up height of 15.7 m along the coast of central Java [45].

Several other earthquakes in the last few years have produced damaging tsunamis and have been mentioned as possible tsunami earthquakes. Seno and Hirata [69] suggest that the great 2004 Sumatra–Andaman earthquake also likely involved a component of tsunami earthquakes, because tsunamis larger than expected from seismic slip occurred, possibly due to slow slip in the shallow subduction boundary. It has also been proposed that the Kurile Islands earthquake of November 15, 2006 [35] may have exhibited some characteristics of tsunami earthquakes and, even more recently, the Solomon Island earthquake of April 1, 2007 excited large tsunamis, at least locally (<http://soundwaves.usgs.gov/2007/04/>). However, the disparity between seismic and tsunami excitation by these events is not nearly as large as for the events in Table 1, and we do not list these events as tsunami earthquakes.



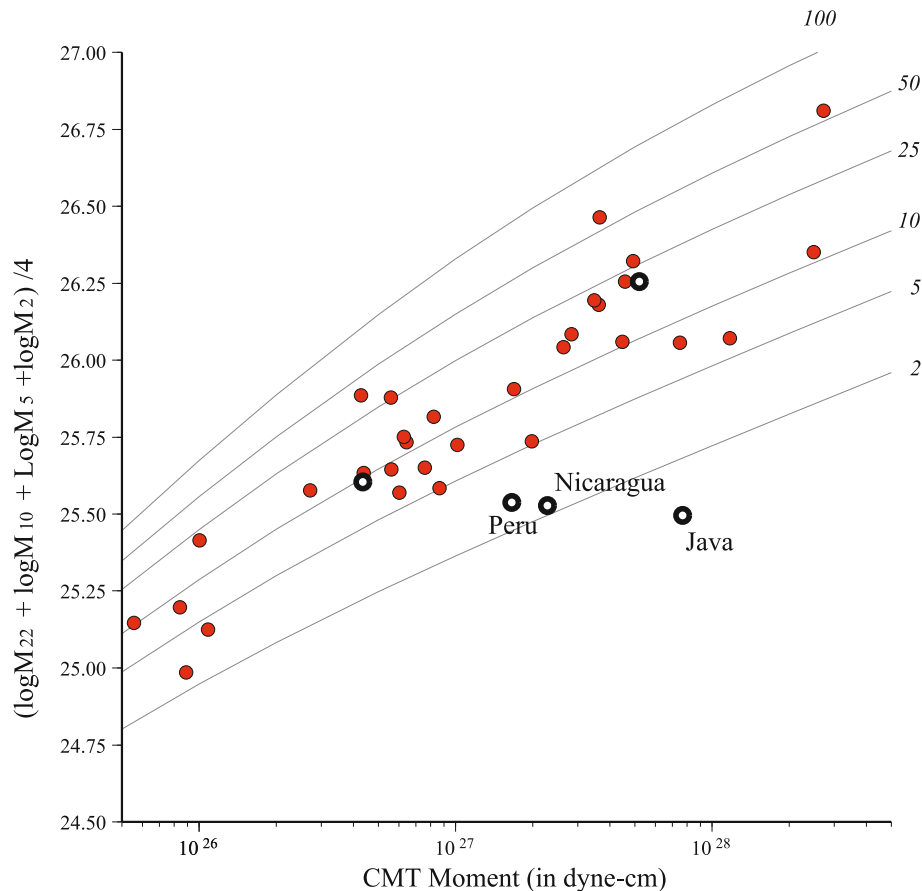
### Characteristics of Tsunami Earthquakes

Fortunately, tsunami earthquakes occur only infrequently. Fewer than ten of these events have occurred in the last three decades since the global installation of seismic broadband instruments and tide gauges and easy availability of their data were established. However, from the detailed investigations of the most recent events and comparisons with the limited data available for the older earthquakes in Table 1, several characteristics of these earthquakes clearly emerge.

#### Slow Character

The slow character of tsunami earthquakes manifests itself in several different, yet related, ways. One well-established

characteristic of tsunami earthquakes is the discrepancy between the determined values of the different seismic magnitude types, calculated from various kinds of seismic waves or waves of different frequency ranges. A typical comparison is that of the body wave magnitude,  $m_b$ , or the surface wave magnitude,  $M_S$ , with the moment magnitude of the earthquake,  $M_w$ . For tsunami earthquakes  $m_b$  is typically much smaller than the other two magnitudes,  $M_S$  and  $M_w$ , and  $M_w$  typically exceeds  $M_S$ . The discrepancy between these different magnitudes is more pronounced than for regular subduction zone earthquakes with similar moment magnitudes. For example: for the 1992 Nicaragua earthquake:  $m_b = 5.4$ ;  $M_S = 7.2$ ,  $M_w = 7.7$  [18]; for the 1994 Java tsunami earthquake:  $m_b = 5.7$ ;  $M_S = 7.1$ ,  $M_w = 7.8$  [59]; for the 2006 Java earthquake:



Tsunami Earthquakes, Figure 2

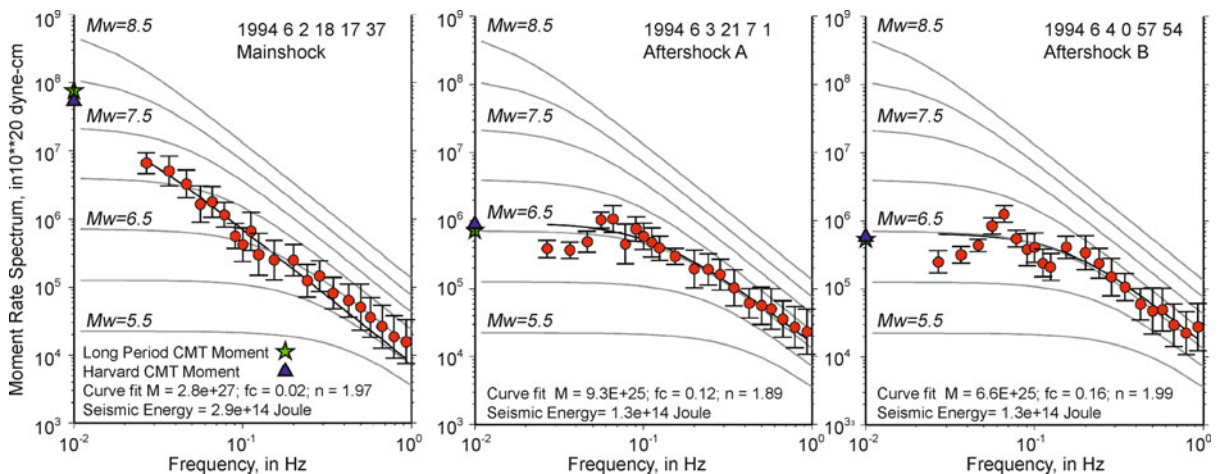
Average of log of moment rate spectrum at four periods (2, 5, 10, 22 s) as a function of the seismic moment as determined by CMT inversion of long-period surface waves. Reference curves were calculated for an  $\omega^2$  model [13]. Values next to the grey curves indicate the stress drop used to calculate the reference curve. Events shown are all shallow subduction zone earthquakes from 1992 to 1996 with moment magnitude 7.0 or greater. Earthquakes followed by anomalously large tsunamis are indicated with open circles. Of these events, only the 1992 Nicaragua, 1994 Java and 1996 Peru earthquakes are slow tsunami earthquakes, as is shown in this figure by their relatively low moment rate spectrum at shorter periods. Adapted from [59]

$m_b = 6.2$ ;  $M_S = 7.2$ ,  $M_w = 7.7$  [5]. Since the body wave magnitude is calculated from short period P-waves, the surface wave magnitude is determined by the amplitude of surface waves with a period of 20 seconds and the moment magnitude is generally based on longer periods for big events, this consistent discrepancy is an indication of the relatively greater seismic energy release at longer periods (or the “slow” character) of these tsunami events. Similarly, investigations of teleseismic P-waves [5,59] have shown that their source spectra are depleted in high frequency energy at periods shorter than 20 seconds as compared to other shallow subduction zone earthquakes (see Fig. 2, from [59]), as well as their own aftershocks (see Fig. 3, adapted from [60]). Modeling of the rupture processes shows that the rupture velocities for tsunami earthquakes are slower than for most other subduction zone earthquakes (for several events: [58] and [33]; for Aleutians 1946: [47]; for Nicaragua 1992: [40] and [30]; for Peru 1996: [31]; for Java 2006: [5]). Correspondingly, the centroid times and source durations or rise times determined for these events are also relatively large with respect to other large subduction zone earthquakes ([59]; for Kurile Islands 1975: [71]; for Peru 1960: [57]; for Java 2006: [5] and [25]), although they may not be anomalous relative to other, smaller, subduction zone earthquakes at very shallow depth [9]. The energy that is radiated by these slow rupture processes is also anomalously low, as is shown by analyses of the radiated energy to moment ratio (for several recent tsunami earthquakes: [51]; for the

1946 Aleutian Islands earthquake: [47]) and radiation efficiency [83].

Unfortunately, the slow character of tsunami earthquakes also means that local residents are not warned by strong ground shaking of the possibility of an impending tsunami. Field surveys and first-person accounts describe the motion of tsunami earthquakes more as a weak “rolling motion” than the usual impulsive character of local events. In the case of the Nicaragua earthquake, some felt a very feeble shock before the tsunami, but most did not feel the earthquake at all [30]. For the 1994 Java event earthquake-induced ground shaking was not noticed by the coastal residents interviewed in Bali and Java [74]. Interviews with local residents carried out for the 2006 Java earthquake [50] also indicate that they felt little or no shaking.

Most designs for tsunami earthquake discriminators and early warning systems make use of a number of the manifestations of the unusually slow character of tsunami earthquakes listed above and many incorporate the use of long period seismic waves for robust estimation of the size of the event. For example, the pulse width of the P-wave, used to calculate moment magnitude  $M_{wp}$ , can give an accurate estimate of source duration time. The combination of  $M_{wp}$  and the source duration can provide an effective tool to issue early tsunami warnings [81]. A slightly later arrival on a seismogram, the W-phase, is a distinct ramp-like long-period (up to 1000 s) phase that begins between P and S waves on displacement seismograms and is particu-



Tsunami Earthquakes, Figure 3

Moment rate spectra for the Java thrust mainshock (left) and two of its largest aftershocks, tensional events in the outer rise (right panels). The stars indicate the Harvard CMT moment, the triangles the moment determined using very long period surface waves. Grey reference curves were calculated for an  $\omega^2$  model [13] with a stress drop of 30 bars and an S-wave velocity of 3.75 km/s. The moment rate for the main shock is similar to that of its much smaller (in terms of  $M_w$ ) aftershocks for periods shorter than 10 seconds. Adapted from [60]

larly pronounced for slow earthquakes; thus it can be used for identification of these types of events [38]. Another method for fast regional tsunami warning uses the ratio of the total seismic energy to the high-frequency energy (between 1 and 5 Hz), computed from the seismograms [70]. Similarly, the detection of deficient values of seismic energy-to-seismic moment ratio can be accomplished in automated, real-time mode [51].

### Location: Close to Trench

The hypocenters of the recent tsunami earthquakes are located relatively close to the trench, as compared to regular subduction zone earthquakes. The Global CMT and other [59] centroid locations for several of these events are located even on the seawards side of the trench (also see Fig. 4). It may be possible that the inversion process mislocates the centroid of the event due to the unusually long duration of the seismic source for its moment and thus its unusually late centroid time. Inversions using seismic and/or tsunami waveforms and other waveform investigations for the 1896 Sanriku event [76], the 1946 Aleutian Islands earthquake [36,78], the 1960 Peru earthquake [58], the 1963 Kurile earthquake [7,58,86], the 1975 Kurile earthquake [58,86], the 2006 Java earthquake [5,20], the Nicaragua earthquake [40,41,62] and the 1996 Peru event [31,64] indicate the presence of concentrated slip in a narrow region near the trench (see Fig. 4). Although in many of these inversions only 1-D Green's functions were used, preliminary research results using a more realistic velocity model for the shallow subduction zone [54] shows a similar picture for the 2006 Java tsunami earthquake.

### Aftershocks

The aftershock sequences of (recent) tsunami earthquakes are unusual in the preponderance of events not located on the interface between overriding and subducting plates [59]. Some of these aftershocks are located in the outer rise according to their Global CMT centroid locations and, in the case of the Java 2006 aftershocks, relocations using a 2.5-D model of the subduction zone [54] confirm this location. Others are located in the overriding plate ([8] for the 2006 Java earthquake), with some deeper within the accretionary prism (for the Java 2006 earthquake: [54]). The low number, or non-existence, of large (greater than magnitude 5.5) interplate earthquakes suggests that the main shock almost completely relieved the stress on the interface or may be related to the frictional properties of the fault. Several explanations have been proposed for the anomalously high number of intraplate earthquakes following tsunami earthquakes. Because

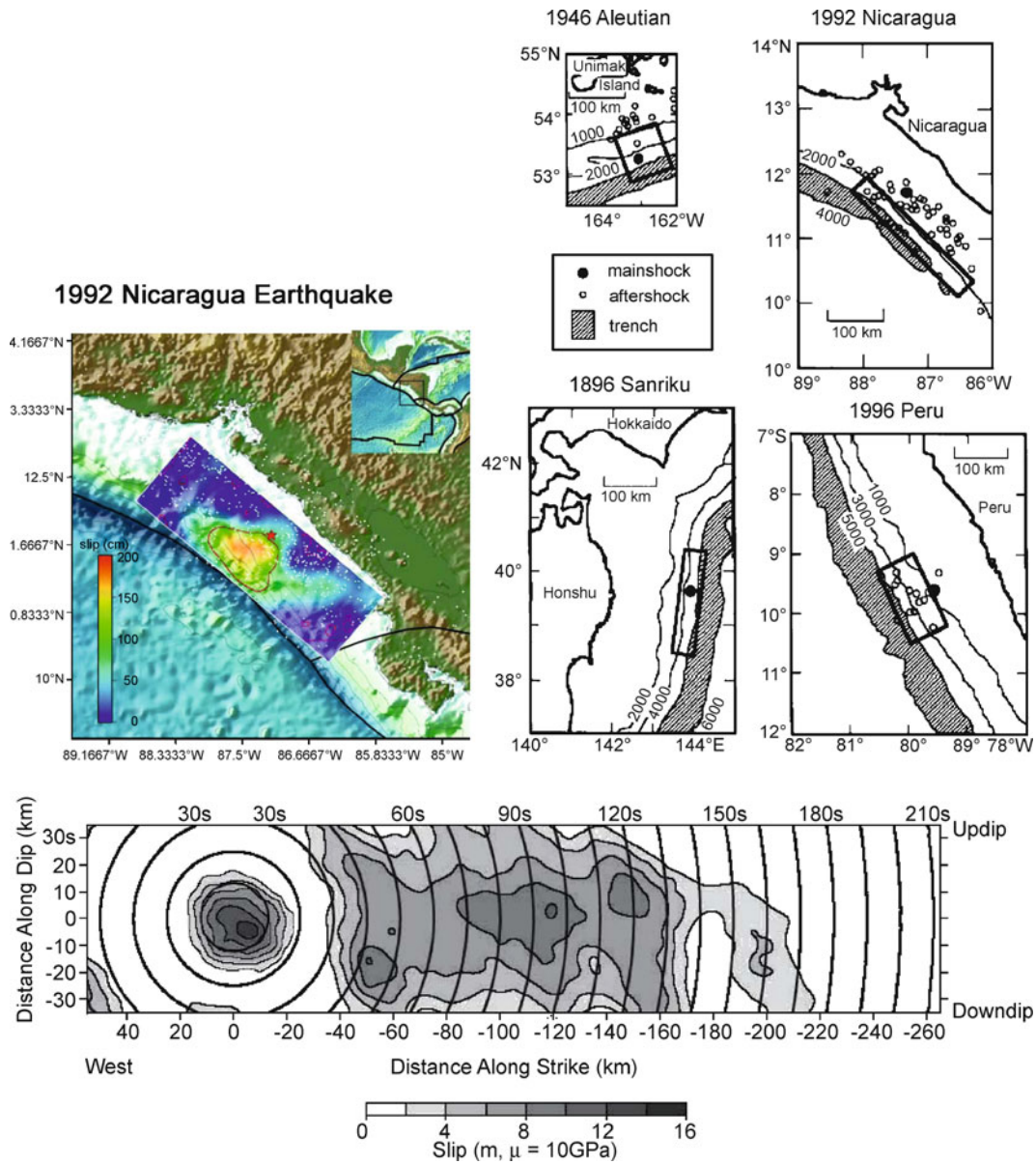
of the proximity of the areas of high slip to the trench, the stress-change in the outer rise and trench area due to a tsunami earthquake are greater than for the "standard" subduction zone earthquake. Several modeling studies of stress alterations caused by large subduction earthquakes suggest that the subduction slip will act to increase the tensional stress and favor normal events in zones towards the ocean from the upper limit of the rupture [16,80]. The subducting plate also seems to have a highly-broken up or rough character in many areas in which tsunami earthquakes have occurred ([49,75]; for Java: [48]; for Peru: [29] and [44]), which suggests the presence of more pervasive pre-existing weak zones, due to, for example, seafloor spreading related fabric. These weak zones may be re-activated in outer rise, or deeper intraplate, earthquakes following a tsunami earthquake.

### Factors Involved in the Seismogenesis and Tsunamigenesis of Tsunami Earthquakes

Based on the consistent characteristics of tsunami earthquakes, as described in the previous section, and observations of their tectonic environments, hypotheses have been developed as to the cause of their extraordinary tsunami excitation and unusual seismic source process. In this section, we will document the factors that have been proposed to affect the seismo- and tsunami-genesis of tsunami earthquakes. Some are associated with the numerical prediction of tsunami wave-heights based on observed seismic waveforms, and others with possible unusual conditions of the tectonic environment in which these events occur; many of these factors are closely or at least somewhat related.

### Slow Character May Lead to Underestimation of Earthquake Size

Prior to the installation of the broadband Global Seismic Network, the magnitudes of earthquakes were often determined from the amplitude of their teleseismic P-waves only. In the case of tsunami earthquakes, using this technique to determine their magnitude would lead to an initial underestimation of their true size, since their source spectra are depleted in the relatively high frequency energy that usually dominates the direct P-wave signals of regular earthquakes [59]. A similar issue would occur, although to a lesser degree, when using surface waves of periods of 20 seconds to determine the surface wave magnitude of these slow events [58]. With the advent of broadband sensors in the past several decades, it has now however become possible to investigate seismic waves to very long periods, hundreds or even thousands of seconds. Using more sophis-



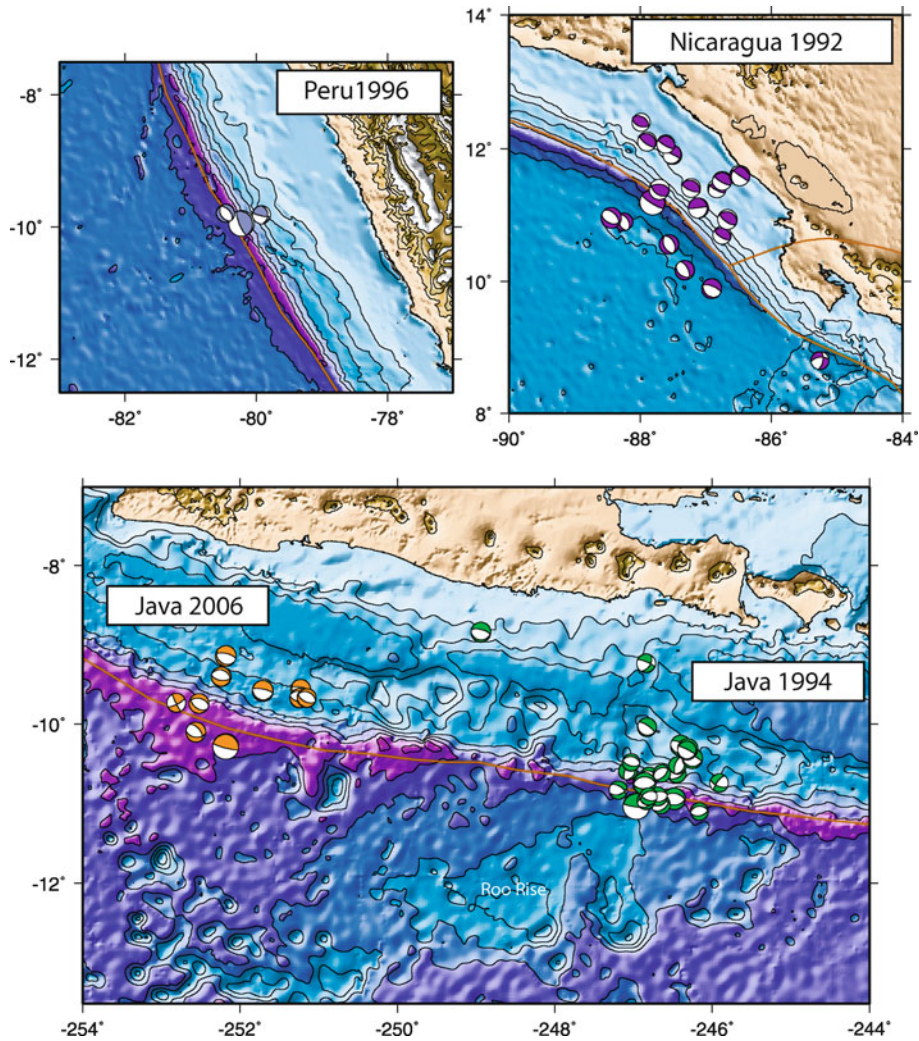
Tsunami Earthquakes, Figure 4  
 Slip or fault models determined for various tsunami earthquakes. The models shown are for: Nicaragua 1992 (top left, from [33]) Java 2006 (bottom, from [5]) and four tsunami earthquakes (top right, from [64]; the main shock and aftershock epicenters are shown and the hatched area indicates the trench). These models all show the presence of substantial slip close to the trench

ticated techniques and the waveforms from technologically advanced sensors, the source spectrum of the recent tsunami earthquakes can now be modeled to these very long periods. Thus, no long period seismic energy that would excite tsunami waves should be “hidden” from the view of seismologists in the computation of moment magnitude or full rupture models using long period surface and body waves. However, for the earthquakes discussed

in this paper, the observed tsunami are still larger than would be expected, even for their moment magnitude.

**Effect of the Presence of Weak Materials with Low Shear Modulus**

Most earthquake source inversions (either for full rupture or Centroid Moment Tensor parameters) implement



Tsunami Earthquakes, Figure 5

Aftershock sequences (seismicity in the region of the main shock within a period of 4 years after its occurrence) for several tsunami earthquakes, from the Global CMT catalog. Relatively few interplate aftershocks occur on the fault plane after a tsunami earthquake occurs, but a preponderance of normal-fault, probably intraplate, aftershocks is apparent

a simple one-dimensional velocity and rigidity model to compute synthetic seismograms and model the recorded waveforms. If tsunami earthquakes are unusually shallow and/or involve sediments of low seismic wave speed, this is probably a poor approximation of the actual structure near the source region [23,64]. Some authors have attempted to rectify this error by using moment or slip distributions determined by seismic inversions in a structural model with significantly reduced shear modulus, more appropriate for the shallow trench region, and forward modeling the tsunami waves (e.g. [23]). This approach, however, is not satisfactory because a seismic inversion for moment or slip using this more appropriate rigidity model would

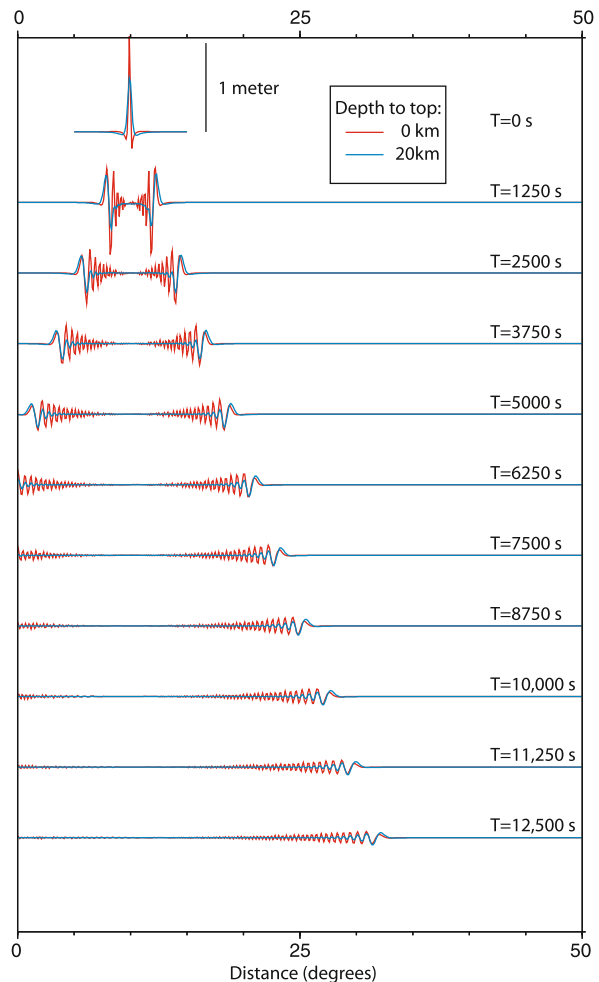
produce a different distribution of moment or slip. Thus, a simple “correction” for the use of an inappropriate value of rigidity cannot be carried out after the rupture model has already been computed. If available, a correct rigidity model should be part of the seismic inversion process itself. To do this correctly, Green’s functions should be computed for a three-dimensional (or possibly two dimensional, if the velocity structure is relatively uniform in the trench-parallel direction) velocity structure of the shallow subduction zone and incorporated in the modeling of the seismic waveforms. However, such sophisticated models are currently only available for very few subduction zones and the computational power required for these calcula-

tions (for body wave frequencies) is substantial. Unfortunately, whichever approach is chosen to go from recordings of seismic waveforms of tsunami earthquakes to the prediction of tsunami wave heights, a good model of the velocity and elastic properties of the shallow subduction zone is an unavoidable requirement.

Even when the moment distribution or moment magnitude of a tsunami earthquake has been determined using an appropriate model for the velocity and elastic parameters, there exists also the issue of enhanced tsunami excitation in material with weaker elastic properties, such as sedimentary layers. Modeling suggests that an event for which 10% of the moment is in sediments generates a tsunami 10 times larger than its seismic moment would suggest [52] mainly because the slip in this material would be much greater than that for the same seismic moment in a stronger material (moment is the product of slip, area and rigidity after all). Therefore, the moment of a tsunami earthquake, even if determined correctly, may not directly reflect its tsunamigenic potential when low velocity sediments are present in the rupture zone. Since tsunami wave heights are mainly determined by the vertical displacement of the ocean floor, which in turn is primarily controlled by the slip on the fault plane, the slip (distribution) is more directly indicative of tsunamigenic potential. Since variations in shear modulus of a factor of five are not uncommon in shallow subduction zones [23], earthquakes with similar moments can result in substantially different slip models and tsunami excitation.

### Shallow Depth of Slip Causes Relatively Great Displacement of Ocean Floor

Shallower earthquakes produce greater and shorter wavelength vertical displacement of the ocean floor, and thus greater and shorter wavelength tsunami waves right above the source region. However, higher frequency waves travel more slowly than longer period waves and, after a few hundred or thousand kilometers of travel, they drift to the back of the wave train and do not contribute to the maximum amplitude. Beyond about 2000 km distance, any earthquake at a depth less than 30 km appears to be equally efficient in tsunamigenesis ([85] and Fig. 6). Therefore, the exact depth of the slip in a shallow earthquake is not a key factor in determining its teleseismic tsunami wave-heights. Although the teleseismic tsunami wave-heights for a shallow slip event may not be significantly greater in amplitude than those for a somewhat deeper slip event [85], at local and regional distances the depth of the slip is an important factor. Therefore, for tsunami earthquakes, which have anomalously great tsunami height at



**Tsunami Earthquakes, Figure 6**  
Cross sections of an expanding tsunami from a M7.5 thrust earthquake. The fault strikes north south (into the page) and the sections are taken east west. Elapsed time in seconds is given at the left and right sides. Red lines are for a fault that breaks the surface and blue lines for a fault with its top at a depth of 20 km. Deeper earthquakes make smaller and longer wavelength tsunamis at relatively short distances

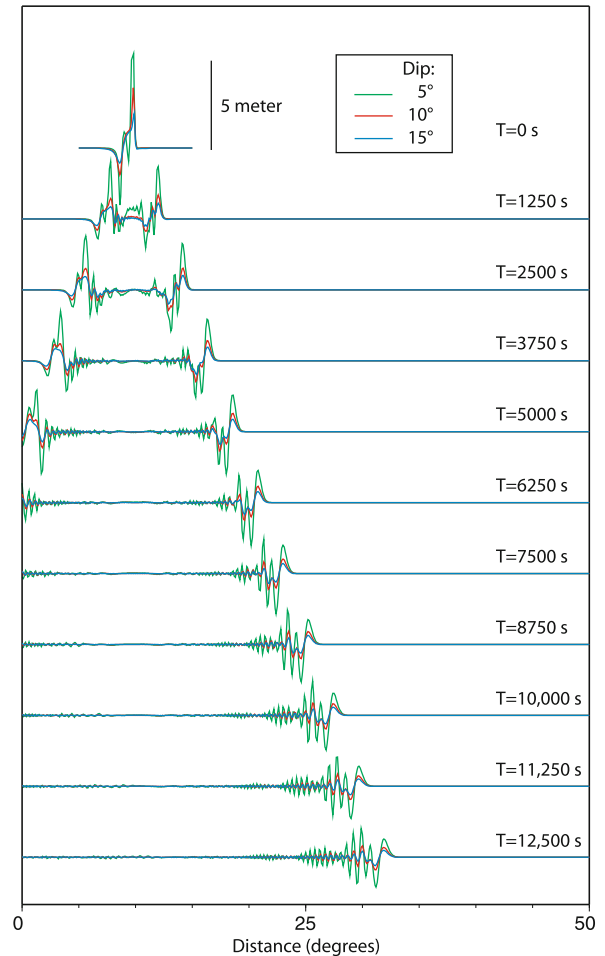
mainly local and regional distances close to the rupture, the depth of the rupture should be a significant factor in their tsunamigenesis (see Fig. 6). Furthermore, modeling of shallow subduction zone earthquakes using a specific crack model [24] indicate that a rupture intersecting the free surface results in approximately twice the average slip. However, under the assumption of other specific frictional and crack models, the modeling of subduction zone earthquakes by Wang and He [84] produces slip models that have less vertical displacement of the ocean floor when the slip reaches the surface, due to a different slip distribution and the curvature of the subduction interface.

### Shallow Fault Dip May Lead to Underestimation of Slip from Seismic Waves

It is notoriously difficult to resolve moment,  $M_0$ , and dip,  $\delta$ , independently for shallow thrust earthquakes. The excitation functions of Rayleigh waves for shallow thrust earthquakes show that only the product of dip and moment (more precisely  $M_0 \sin(2\delta)$ ) can be resolved [39] when using these surface waves in inversions for a source mechanism. If Love wave data are also included in the inversion, it may be possible to add constraints by concentrating on fitting the amplitudes of those Love waves recorded at azimuths corresponding to the along strike direction [34], however, this approach may be complicated by directivity effects. The polarity of body waves can only be used to constrain the focal mechanism of an earthquake at a limited range of incidence angles and thus also cannot provide any additional constraints on the dip of the shallowly dipping plane, unless assumptions are made about the rake angle. It is thus possible to (severely) underestimate the amount of slip in the earthquake, if the dip of the mechanism is poorly constrained and the inversion leads to a value for dip that is too high. This could be particularly important for very shallow subduction earthquakes, since the dip is expected to be small for these events. Thus, a difference between a dip of 3 or 6 degrees in a CMT solution may not seem significant, but it could lead to a difference in moment (and thus slip) of a factor of two. To illustrate the importance of this issue for very shallowly dipping thrust events, we show in Fig. 7 three different homogeneous slip models (only slip was changed, fault surface area and rigidity were kept constant) that will produce similar surface waves because the product of their slip and dip is identical. However, the vertical deformation of the ocean floor and the tsunami waveforms resulting from these three different earthquakes would be significantly different in amplitude.

### Horizontal Deformation of the Ocean Floor May Lead to Great Displacement of Water, Yet Is Neglected in Tsunami Modeling

Most tsunami modelers only consider a water surface displacement identical to the vertical deformation of the ocean bottom due to faulting when computing the tsunami height resulting from an earthquake and neglect the effect of horizontal deformation. However, when the tsunami source is located close to a steep slope and the horizontal displacement is large relative to the vertical displacement, which is generally the case for tsunami earthquakes due to their mechanism and shallow depth,



**Tsunami Earthquakes, Figure 7**  
Tsunami wave heights as a function of time for three different slip models for a shallow thrust earthquake, which will produce similar surface wave recordings (because the product of dip and moment is held constant). Thus, it is difficult to resolve between these different models using an inversion of surface waves, yet they produce very different vertical displacement of the ocean floor and thus very different tsunamis

the effect of horizontal deformation may become significant [77].

Furthermore, it has been suggested that the lateral collision force of a continental slope into the ocean due to faulting could also play an important part in the tsunami-generation of these events [72]. This type of dynamic excitation of tsunami waves would be particularly important for very shallowly dipping, shallow, thrust events, which would have a large component of horizontal motion. However, this concept has still been debated. This problem will be resolved if tsunami excitation is computed

using the three-dimensional bathymetry in the source region with the displacement of the sea-floor as the initial condition.

### **Subduction of Bathymetric Features May Enable Seismic Slip in Usually Aseismic Region and Be Related to Unusual Aftershock Sequences**

Sandbox experiments show the pervasive influence on the geomorphology of the shallow subduction zone margin when a seamount on the subducting plate is being subducted [17]. Subduction and underplating of relatively undeformed and water-laden sediments beneath the rear part of the margin could, together with the dense fracture network generated by seamount subduction, modify the fluid pressure and introduce significant variations of the effective basal friction and thus the local mechanical plate coupling. More directly, subduction of a seamount may increase the normal stress across the subduction interface and hence enhance seismic coupling [15,67]. Unusual earthquakes have been documented in regions where ridges, seamounts or other bathymetric features are being subducted (e. g. [14,43]) and investigations of rupture characteristics of large underthrusting earthquakes provide evidence that seamounts can be subducted to seismogenic depths and that variations in seafloor bathymetry of the subducting plate may strongly influence the earthquake rupture process [10,61]. In the case of the Java tsunami earthquake of 1994, the bathymetry of the area landwards of the trench suggests that a local high is in the process of being subducted close to the area of maximum slip ([3], Fig. 4). A bathymetric high in the form of the Roo Rise can also be found just seaward of the trench region. In the bathymetry of the area around the 2006 Java event no such pronounced local feature can be found (Fig. 4), but the regional bathymetry south of the Java trench region is distinguished by an overall rough character. Similarly, the Nicaragua subduction zone is characterized by a highly developed horst and graben structure in the subducting plate, but no large-scale features, like a subducting seamount, are obvious. However, in case of the Peru earthquakes, both the 1996 and 1960 tsunami earthquakes occur at the intersection of the trench with major topographic features on the Nazca plate: the Mendaña fracture zone and the Trujillo trough, respectively [53].

Thus, subduction of either pronounced local bathymetric features or more regional seafloor roughness or horst-and-graben structures, which may modify the local coupling between subducting and overriding plates, has been documented in or near the rupture zone of many tsunami earthquakes.

### **Accretionary Prism: Uplift, Slides or Splay Faulting May Displace a Relatively Great Volume of Water**

Tsunami earthquakes may involve seismic slip along the normally aseismic basal decollement of the accretionary prism [58,76]. Sediments near the toe of an inner trench slope may be scraped off by a large horizontal movement over the decollement due to an earthquake and thus cause an additional inelastic uplift, which could have a large effect on tsunami generation (for the 1896 Sanriku earthquake: [79]; for the 1946 Aleutian Islands earthquake: [78]).

The existence of splay faulting, which would be more effective in exciting tsunamis due to their steeper dip, within the accretionary prism itself has been suggested to be a cause of the large tsunami excitation for the 1994 Java earthquake [3], the 2004 Sumatra mega-thrust event [46] and the 1963/1975 Kurile earthquakes [21]. Splay faulting can further promote extensive vertical deformation of the ocean floor, and hence large tsunamis, through partitioning or branching of a rupture upwards from the interface along multiple splay faults leading up to the surface [21,56].

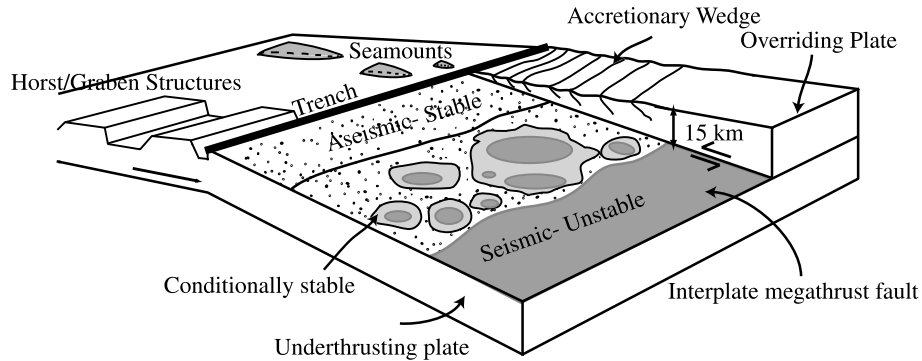
### **Presence of Fluids Influences Seismic Behavior**

In subduction zones, fluids expelled from the subducting plate play an important role in many different subduction related phenomena such as volcanism, metamorphism and seismogenesis. Zones of high pore fluid pressure in the shallow subduction zone would change the effective normal stress significantly, possibly extending the region in which seismic slip is possible to shallower depths and generate slip of a slow nature. The presence of such zones has been suggested to be related to the occurrence of silent slip in the Nankai trough region [42]. In the case of tsunami earthquakes, the presence of large zones of elevated fluid pressure has been proposed to cause fairly rapid seismic slip close to the trench axis, following the breakage of asperities [68].

### **A Model for Tsunami Earthquakes**

Most explanations for the slowness of tsunami earthquakes involve the presence of low velocity, low strength and low rigidity sediments in the accretionary prism and between the overriding and subducting plate in the shallow subduction zone. Rupture through these slow sediments is thought to promote a slow rupture velocity. In Scholz's [65,66] model of the typical subduction zone (see Fig. 8 for an interpretation of this model from [9]), three possible stability regimes exist. In the stable zone, seismic





Tsunami Earthquakes, Figure 8

Cartoon illustrating frictional conditions of the subduction interface between subducting and overriding plate. Individual unstable sliding contact areas (dark gray) can provide the nucleation sites for rupture in the shallow subduction zone environment, which is typically a stable (stippled) or conditionally stable (light gray) frictional region. From [9]

slip cannot be supported and aseismic creep releases all strain. In the unstable zone episodic seismic slip occurs. In the conditionally stable zone, in between these two zones, slip can be abrupt if it experiences loading from a nearby slip patch. This conditionally stable zone may be very heterogeneous due to roughness of the thrust fault [75] creating isolated asperities or due to permeability changes [55] from the subduction of low permeability materials or the presence of fluids [68]. Tsunami earthquakes may represent a rupture of one or several large “unstable” asperities, which then propagates for a significant distance in the conditionally stable sedimentary materials [9,59].

Alternatively, but consistently, we could interpret tsunami earthquakes in the context of fracture mechanics as displaying a lack of radiated energy and a low rupture speed in a high  $G_C$  (fracture energy) environment [5,83]. As stated above, if these ruptures also involve localized patches of relatively strong unstable friction (that would be associated with high rupture speed and low  $G_C$ ) it would allow the rupture to propagate seismically, instead of as a continuous creep process. From this point of view, tsunami earthquakes dissipate a large amount of energy during the fracture process and are left with little energy to radiate. It is possible that the highly faulted trench and deformed sediments result in larger energy dissipation during failure due to an excessive amount of branching and bifurcation of cracks which gives rise to inelastic behavior and hence a large dissipation of energy [6,83], possibly involving the branching of the rupture into multiple splay faults in the accretionary prism.

Tsunami earthquakes therefore would represent slip at unusually shallow depths that would typically be dominated by creep-processes. Their nucleation would be made possible through the existence of localized asperities or

patches of unstable friction in a typically stable or conditionally stable region. The presence of compartments of elevated fluid pressure may aid in the propagation of the seismic slip by creating zones of nearly zero friction surrounding the asperities [68]. These asperities may be created by the subduction of bathymetric features like seamounts or ridges or by the broken up nature of the subducting plate itself, creating a horst-and-graben system, which would act as buckets for sediment subduction [59,75]. The stress-release on these asperities would be near complete, and any additional unloading of stress on the plate interface due to the rupture may occur mostly through creep. This would result in a relatively low number of aftershocks occurring on the interface between overriding and subducting plate. However, the static stress change in the outer rise area would be significant due to the shallow nature of most of the slip and thus normal faulting outer rise earthquakes would be more likely to be triggered [16,29,80]. The subduction of a bathymetric feature would also likely result in fracturization of the margin [17] in the overriding plate, thus further promoting the occurrence of intraplate aftershocks in this area of the shallow subduction zone. If the subducting plate itself is highly broken-up and thus contains pre-existing weak zones, this may facilitate further faulting within the subducting plate, in particular close to the lower edge of the rupture where the stress change due to the main shock is relatively large.

In this model for tsunami earthquakes discussed above, the unusually high effectiveness in the excitation of tsunami waves can be attributed to several factors, with the shallowness of the slip as the main underlying cause. Other important concerns coming into play are the possible failure of seismological techniques to provide an accurate estimate of the slip due to complexities associated with a very

shallowly propagating rupture in a subduction zone, the possible failure of tsunami modeling to determine accurate wave heights due to similar complexities and the possible involvement of splay faulting or uplift of sediments near the trench in the accretionary prism.

### Future Directions

The Sumatra–Andaman earthquake and tsunami of 2004 renewed interest in the development of near real-time methods to estimate the true size of large earthquakes and the tsunamis that might follow them, and the installation of instrumentation that will facilitate these measurements. Because the time between the earthquake and the arrival of the first tsunami waves at the local coastline is short, it is still unclear how effective these types of early warning systems are for saving lives at short distances from the tsunami source, but they will be useful at large distances.

New technologies and surveys will enhance our knowledge of the geomorphology and velocity structure of the shallow subduction zone. Ocean bottom seismometers, tide gauges, buoys and other seafloor monitoring devices will provide high quality data, which will enable us to place better constraints on where exactly the slip in shallow earthquakes occurs and in what tectonic environment. Although tsunami earthquakes occur relatively infrequently and thus may be difficult to capture, comprehensive characterizations of their rupture processes placed in the context of detailed three-dimensional models of the shallow subduction zones they occurred in will be an important next step in understanding their unusual seismic and tsunami-genic processes.

### Bibliography

#### Primary Literature

- Abe K (1979) Size of great earthquakes of 1873–1974 inferred from tsunami data. *J Geophys Res* 84:1561–1568
- Abe K (1989) Quantification of tsunamigenic earthquakes by the  $M_f$  scale. *Tectonophysics* 166:21–34
- Abercrombie RE, Antolik M, Felzer K, Ekstrom G (2001) The 1994 Java tsunami earthquake- Slip over a subducting seamount. *J Geophys Res* 106:6595–6608
- Ammon CJ, Ji C, Thio HK, Robinson D, Ni S, Hjorleifsdottir V, Kanamori H, Lay T, Das S, Helmsberg D, Ichinose G, Polet J, Wald D (2005) Rupture Process of the 2004 Sumatra–Andaman Earthquake. *Science* 308:1133. doi:10.1126/science.1112260
- Ammon CJ, Kanamori H, Lay T, Velasco AA (2006) The 17 July 2006 Java tsunami earthquake. *Geophys Res Lett* 33:24. doi:10.1029/2006GL028005
- Barragan BE, Giaccio GM, Zerbino RL (2001) Fracture and failure of thermally damaged concrete under tensile loading. *Mater Struct* 34:312–319
- Beck SL, Ruff LJ (1987) Rupture process of the great 1963 Kuril Islands earthquake sequence: asperity interaction and multiple event rupture. *J Geophys Res* 92:14123–14138
- Bilek SL, Engdahl ER (2007) Rupture characterization and relocation aftershocks of the 1994 and 2006 tsunami earthquakes in the Java subduction zone. *Geophys Res Lett* 34:L20311. doi:10.1029/2007GL031357
- Bilek SL, Lay T (2002) Tsunami earthquakes possibly widespread manifestations of frictional conditional stability. *Geophys Res Lett* 29:18–1. doi:10.1029/2002GL01521
- Bilek SL, Schwartz SY, Deshon HR (2003) Control of seafloor roughness on earthquake rupture behavior. *Geology* 31:455–458. doi:10.1130/0091-7613(2003)031
- Bourgeois J, Petroff C, Yeh H, Titov V, Synolakis CE, Benson B, Kuroiwa J, Lander J, Norabuena E (1999) Geologic Setting, Field Survey and Modeling of the Chimbote, Northern Peru, Tsunami of 21 February 1996. *Pure Appl Geophys* 154: 513–540
- Brown DL (1964) Tsunami activity accompanying the alaskan earthquake of 27 March 1964. US Army Engr Dist, Alaska, 20 pp
- Brune JN (1970) Tectonic stress and spectra of seismic shear waves from earthquakes. *J Geophys Res* 75:4997–5009
- Chung WY, Kanamori H (1978) Subduction process of a fracture zone and aseismic ridges – the focal mechanism and source characteristics of the New Hebrides earthquake of 1969 January 19 and some related events. *Geophys J Int* 54(1):221–240. doi:10.1111/j.1365-246X.1978.tb06764.x
- Cloos M (1992) Thrust-type subduction-zone earthquakes and seamount asperities; a physical model for seismic rupture. *Geology* 20:601–604
- Dmowska R, Zheng G, Rice JR (1996) Seismicity and deformation at convergent margins due to heterogeneous coupling. *J Geophys Res* 101:3015–3029
- Dominguez S, Malavieille J, Lallemand SE (2000) Deformation of accretionary wedges in response to seamount subduction: insight from sandbox experiments. *Tectonics* 19:182–196
- Engdahl ER, Villaseñor A (2002) Global seismicity: 1900–1999. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology*. Academic Press, Amsterdam, Part A, chapt 41, pp 665–690
- Fryer GJ, Watts P, Pratson LF (2004) Source of the great tsunami of 1 April 1946: a landslide in the upper Aleutian forearc. *Mar Geol* 203:201–218
- Fujii Y, Satake K (2006) Source of the July 2006 West Java tsunami estimated from tide gauge records. *Geophys Res Lett* 33:L24317.1–L24317.5. doi:10.1029/2006GL028049
- Fukao Y (1979) Tsunami earthquakes and subduction processes near deep-sea trenches. *J Geophys Res* 84:2303–2314
- Geist EL (2000) Origin of the 17 July 1998 Papua New Guinea tsunami: Earthquake or landslide? *Seism Res Lett* 71:344–351
- Geist EL, Bilek SL (2001) Effect of depth-dependent shear modulus on tsunami generation along subduction zones. *Geophys Res Lett* 28:1315–1318
- Geist EL, Dmowska R (1999) Local tsunamis and distributed slip at the source. *Pure Appl Geophys* 154:485–512
- Hara T (2006) Determination of earthquake magnitudes using duration of high-frequency energy radiation and maximum displacement amplitudes: application to the July 17, 2006 Java

- earthquake and other tsunami earthquakes. *Eos Trans AGU* 87(52):Fall Meet Suppl, Abstract S21A-0132
26. Hatori T (1967) The generating area of the Sanriku earthquake of 1896 and its comparison with the tsunami of 1933. *J Seismol Soc Jap Ser 2*, 20:164–170
  27. Heinrich P, Schindele F, Guibourg S, Ihmlé PF (1998) Modeling of the February 1996 Peruvian tsunami. *Geophys Res Lett* 25:2687–2690
  28. Hidayat D, Barker JS, Satake K (1995) Modeling the seismic source and tsunami generation of the December 12, 1992 Flores island, Indonesia, earthquake. *Pure Appl Geophys* 144: 537–554
  29. Hilde TWC (1983) Sediment subduction versus accretion around the Pacific. *Tectonophysics* 99:381–397
  30. Ide S, Imamura F, Yoshida Y, Abe K (1993) Source characteristics of the Nicaraguan tsunami earthquake of September 2, 1992. *Geophys Res Lett* 20:863–866
  31. Ihmlé PF, Gomez JM, Heinrich P, Guibourg S (1998) The 1996 Peru tsunamigenic earthquake: Broadband source process. *Geophys Res Lett* 25:2691–2694
  32. Imamura F, Gica E, Takahashi T, Shuto N (1995) Numerical simulation of the 1992 Flores tsunami: Interpretation of tsunami phenomena in northeastern Flores Island and damage at Babi Island. *Pure Appl Geophys* 144:555–568
  33. Ji C (2006) A comparison study of 2006 Java earthquake and other Tsunami earthquakes. *Eos Trans AGU* 87(52):Fall Meet Suppl, Abstract
  34. Ji C (2006) Resolving the trade-off between the seismic moment and fault dip of large subduction earthquakes and its impact on tsunami excitation. *Tsunami Sources Workshop*. Menlo Park
  35. Ji C, Zeng Y, Song AT (2007) Rupture process of the 2006 Mw 8.3 Kuril Island Earthquake inferred from joint inversion of teleseismic body and surface waves. *SSA meeting*. Kona
  36. Johnson JM, Satake K (1997) Estimation of seismic moment and slip distribution of the April 1, 1946, Aleutian tsunami earthquake. *J Geophys Res* 102:11765–11774
  37. Kanamori H (1972) Mechanism of tsunami earthquakes. *Phys Earth Planet Inter* 6:346–359
  38. Kanamori H (1993) W phase. *Geophys Res Lett* 20:1691–1694
  39. Kanamori H, Given JW (1981) Use of long-period surface waves for rapid determination of earthquake-source parameters. *Phys Earth Planet Inter* 27:8–31
  40. Kanamori H, Kikuchi M (1993) The 1992 Nicaragua earthquake – A slow tsunami earthquake associated with subducted sediments. *Nature* 361:714–716
  41. Kikuchi M, Kanamori H (1995) Source characteristics of the 1992 Nicaragua tsunami earthquake inferred from teleseismic body waves. *Pure Appl Geophys* 144:441–453
  42. Kodaira S, Iidaka T, Kato A, Park JO, Iwasaki T, Kaneda Y (2004) High pore fluid pressure may cause silent slip in the Nankai trough. *Science* 304:1295–1298. doi:10.1126/science.1096535
  43. Kodaira S, Takahashi N, Nakanishi A, Miura S, Kaneda Y (2000) Subducted seamount imaged in the rupture zone of the 1946 Nankaido earthquake. *Science* 289:104–106. doi:10.1126/science.289.5476.104
  44. Kulm LD, Prince RA, French W, Johnson S, Masias A (1981) Crustal structure and tectonics of the central Peru continental margin and trench. In: Kulm LD, Dymond J, Dasch EJ, Hussong DM (eds) *Nazca Plate: Crustal formation and Andean Convergence*. *Geol Soc Am Mem* 154:445–468
  45. Lavigne F, Gomes C, Giffo M, Wassmer P, Hoebreck C, Mardiatno D, Priyono J, Paris R (2007) Field observations of the 17 July 2006 Tsunami in Java. *Nat Hazards Earth Syst Sci* 7: 177–183
  46. Lay T, Kanamori H, Ammon CJ, Nettles M, Ward SN, Aster RA, Beck SL, Bilek BL, Brudzinski MR, Butler R, DeShon HR, Ekström G, Satake K, Sipkin S (2005) The great Sumatra–Andaman earthquake of 26 December 2004. *Science* 308:1127–1133. doi:10.1126/science.1112250
  47. Lopez AM, Okal EA (2006) A seismological reassessment of the source of the 1946 Aleutian ‘tsunami’ earthquake. *Geophys J Int* 165(3):835–849. doi:10.1111/j.1365-246X.2006.02899.x
  48. Masson DG, Parson LM, Milsom J, Nichols G, Sikumbang N, Dwiyanto B, Kallagher H (1990) Subduction of seamounts at the Java trench – a view with long-range sidescan sonar. *Tectonophysics* 185:51–65
  49. McAadoo BG, Capone MK, Minder J (2004) Seafloor geomorphology of convergent margins: Implications for Cascadia seismic hazard. *Tectonics* 23:TC6008. doi:10.1029/2003TC001570
  50. Mori J, Mooney WD, Afnimar Kurniawan S, Anaya AI, Widiyantoro S (2007) The 17 July 2006 tsunami earthquake in west Java, Indonesia. *Seismol Res Lett* 78:291
  51. Newman AV, Okal EA (1998) Teleseismic estimates of radiated seismic energy: The  $E/M_0$  discriminant for tsunami earthquakes. *J Geophys Res* 103:26885–26898
  52. Okal EA (1988) Seismic parameters controlling far-field tsunami amplitudes: A review. *Nat Haz* 1:67–96
  53. Okal EA, Newman AV (2001) Tsunami earthquakes: The quest for a regional signal. *Phys Earth Planet Inter* 124:45–70
  54. Okamoto T, Takenaka H (2006) Source process of the July 17, 2006 off Java island earthquake by using a fine crustal structure model of the Java trench and a 2.5D FDM computations. *Eos Trans AGU* 87(52):Fall Meet Suppl, Abstract
  55. Pacheco JF, Sykes LR, Scholz CH (1993) Nature of seismic coupling along simple plate boundaries of the subduction type. *J Geophys Res* 98:14133–14159
  56. Park J-O, Tsuru T, Kodaira S, Cummins PR, Kaneda Y (2002) Splay Fault branching along the Nankai subduction zone. *Science* 297:1157–1160
  57. Pelayo AM, Wiens DA (1990) The November 20, 1960 Peru tsunami earthquake: Source mechanism of a slow event. *Geophys Res Lett* 17:661–664
  58. Pelayo AM, Wiens DA (1992) Tsunami earthquakes – Slow thrust-faulting events in the accretionary wedge. *J Geophys Res* 97:15321–15337
  59. Polet J, Kanamori H (2000) Shallow subduction zone earthquakes and their tsunamigenic potential. *Geophys J Int* 142:684–702. doi:10.1046/j.1365-246x.2000.00205.x
  60. Polet J, Thio HK (2003) The 1994 Java Tsunami earthquake and its “Normal” Aftershocks. *Geophys Res Lett* 30:27–1. doi:10.1029/2002GL016806
  61. Robinson DP, Das S, Watts AB (2006) Earthquake rupture stalled by a subducting fracture zone. *Science* 312:1203–1205. doi:10.1126/science.1125771
  62. Satake K (1994) Mechanics of the 1992 Nicaragua tsunami earthquake. *Geophys Res Lett* 21:2519–2522
  63. Satake K, Kanamori H (1991) Abnormal tsunamis caused by the June 13, 1984, Torishima, Japan, earthquake. *J Geophys Res* 96:19933–19939

64. Satake K, Tanioka Y (1999) Sources of tsunami and tsunamigenic earthquakes in subduction zones. *Pure Appl Geophys* 154:467–483. doi:10.1007/s000240050240
65. Scholz CH (1990) *The mechanics of earthquakes and faulting*. Cambridge Univ Press, New York
66. Scholz CH (1998) Earthquakes and friction laws. *Nature* 391: 37–42
67. Scholz CH, Small C (1997) The effect of seamount subduction on seismic coupling. *Geol* 25:487–490
68. Seno T (2002) Tsunami earthquakes as transient phenomena. *Geophys Res Lett* 29(10):58.1–58.4. doi:10.1029/2002GL014868
69. Seno T, Hirata K (2007) Did the 2004 Sumatra–Andaman earthquake involve a component of tsunami earthquakes? *Bull Seismol Soc Am* 97:S296–S306. doi:10.1785/0120050615
70. Shapiro NM, Singh SK, Pacheco J (1998) A fast and simple diagnostic method for identifying tsunamigenic earthquakes. *Geophys Res Lett* 25:3911–3914
71. Shimazaki K, Geller RJ (1977) Source process of the Kurile Islands tsunami earthquake of June 10, 1975. *Eos Trans Am Geophys Union* 58:446
72. Song Y, Fu L, Zlotnicki V, Ji C, Hjorleifsdottir V, Shum C, Yi Y (2006) Horizontal motions of faulting dictate the 26 December 2004 tsunami genesis. *Eos Trans AGU* 87(52):Fall Meet Suppl, Abstract U53C-02
73. Synolakis CE, Bardet JP, Borrero JC, Davies HL, Okal EA, Silver EA, Sweet S, Tappin DR (2002) The slump origin of the 1998 Papua New Guinea Tsunami. *Proc Royal Soc A Math Phys Eng Sci* 458(2020):763–789. doi:10.1098/rspa.2001.0915
74. Synolakis CE, Imamura F, Tsuji Y, Matsutomi H, Tinti S, Cook B, Chandra YP, Usman M (1995) Damage, conditions of East Java tsunamis of 1994 analyzed. *EOS* 76:26
75. Tanioka Y, Ruff L, Satake K (1997) What controls the lateral variation of large earthquake occurrence along the Japan trench. *Isl Arc* 6:261–266
76. Tanioka Y, Satake K (1996) Fault parameters of the 1896 Sanriku tsunami earthquake estimated from tsunami numerical modeling. *Geophys Res Lett* 23:1549–1552
77. Tanioka Y, Satake K (1996) Tsunami generation by horizontal displacement of ocean bottom. *Geophys Res Lett* 23:861–864
78. Tanioka Y, Seno T (2001) Detailed analysis of tsunami waveforms generated by the 1946 Aleutian tsunami earthquake. *Nat Haz Earth Syst Sci* 1:171–175
79. Tanioka Y, Seno T (2001) Sediment effect on tsunami generation of the 1896 Sanriku tsunami earthquake. *Geophys Res Lett* 28:3389–3392
80. Taylor MAJ, Zheng G, Rice JR, Stuart WD, Dmowska R (1996) Cyclic stressing and seismicity at strong coupled subduction zones. *J Geophys Res* 101:8363–8381
81. Tsuboi S (2000) Application of  $M_{wp}$  to tsunami earthquake. *Geophys Res Lett* 27:3105–3108
82. Tsuji Y, Imamura F, Matsutomi H, Synolakis CE, Nanang PT, Jumadi, Harada S, Han SS, Arai K, Cook B (1995) Field survey of the east Java earthquake and tsunami of June 3, 1994. *Pure Appl Geophys* 144(3–4):839–854
83. Venkataraman A, Kanamori H (2004) Observational constraints on the fracture energy of subduction zone earthquakes. *J Geophys Res* 109:B05302.1–05302.20. doi:10.1029/2003JB002549
84. Wang K, He J (2008) Effects of frictional behavior and geometry of subduction fault on coseismic seafloor deformation. *Bull Seismol Soc Am* 98(2):571–579
85. Ward SN (2002) Tsunamis. In: Meyers RA (ed) *The Encyclopedia of Physical Science and Technology*, vol 17. Academic Press, San Diego, pp 175–191
86. Wiens D (1989) Bathymetric effects on body waveforms from shallow subduction zone earthquakes and application to seismic processes in the Kurile Trench. *J Geophys Res* 94: 2955–2972

### Books and Reviews

- Bebout G, Kirby S, Scholl D, Platt J (eds) (1996) *Subduction from Top to Bottom*. American Geophysical Union Monograph, no 96. American Geophysical Union, Washington DC
- Satake K, Imamura F (1995) Tsunamis 1992–1994. Special Issue of *Pure Appl Geophys* 144(3–4):373–890

## Tsunami Forecasting and Warning

OSAMU KAMIGAICHI

Japan Meteorological Agency, Tokyo, Japan

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Complexity Problem in Tsunami Forecasting](#)

[Components of a Tsunami Early Warning System \(TEWS\)](#)

[Tsunami Early Warning System in Japan](#)

[Future Outlook](#)

[Acknowledgments](#)

[Bibliography](#)

### Glossary

**Tsunami early warning system** It consists of four components, namely, 1) Seismic network, 2) Seismic data processing system, 3) Tsunami forecast system, and 4) Sea level data monitoring system. In a broader sense, warning transmission system (downlink to disaster management organizations and public) is also included.

**Tsunami amplitude** Amplitude is measured from undisturbed sea level to peak or trough of the wave. By definition, it can be positive or negative. Tsunami amplitude can be measured in real-time by instruments like tide gauge, pressure sensor, etc., and can be reproduced by numerical tsunami propagation simulation. One needs not to be confused with the term 'run-up height'. It is the maximum height of inundation on land, and measured in post-tsunami field surveys from traces of tsunami (i. e. damage of constructions, vegetative markers, etc.).

**Simulation point** In this article, it is defined as the surface projection location of the hypothetical earthquake fault center. The vertical component of ocean bottom deformation due to earthquake fault dislocation calculated by elastic theory gives the initial tsunami waveform for the numerical tsunami propagation simulation.

**Forecast point** In this article, it is defined as the location of the offshore point where tsunami amplitude is evaluated by using numerical tsunami propagation simulation.

**Intergovernmental coordination group** The group established under UNESCO/IOC to facilitate interna-

tional cooperation for the tsunami disaster mitigation. There are four ICGs as of now (Pacific Ocean, Indian Ocean, Caribbean Sea, North Eastern Atlantic Ocean and Mediterranean Sea). One of the most important characteristics of tsunami is that it can cause huge disaster even after long distance propagation due to amplification near the coast. Therefore, international cooperation, especially the prompt data and information exchange, is essential for the disaster mitigation.

**Centroid moment tensor solution** One of the representation of seismic source process. It is represented by the moment tensor, which is a combination of six independent equivalent force couples, and is a weighted average of the source process in time and space. Since it represents an overall image of the source process, it is suitable to evaluate tsunamigenic potential of the earthquake.

**Earthquake early warning** Earthquake Early Warning is to enable countermeasures in advance for strong motion disaster by detecting seismic *P* wave at stations near the epicenter, quickly estimate seismic intensity and arrival time of *S* wave, and transmit these estimation before the *S* wave arrival. The Japan Meteorological Agency (JMA) started to provide EEW to the general public in October, 2007. This technique is applicable to quicken tsunami warning dissemination.

### Definition of the Subject

A tsunami is, along with strong motion, one of the two major disasters caused by earthquake. To mitigate tsunami disaster, it is important to integrate software countermeasures like tsunami forecast to enable timely evacuation from area at risk before tsunami strikes the coast, as well as to intensify hardware countermeasures particularly in vulnerable coastal areas like building banks and water-gates. Tsunami disaster mitigation can be achieved effectively by the appropriate combination of the software and hardware countermeasures. Also, improving people's awareness on the tsunami disaster, necessity of spontaneous evacuation when they notice an imminent threat of tsunami on their own (feeling strong shaking near the coast, seeing abnormal sea level change, etc.) and how to respond to a tsunami forecast, and conducting tsunami evacuation drill are very important issues for disaster mitigation.

In this article, a tsunami forecast, as the most typical software countermeasure that a national organization can provide, is mainly described. Recent progress in science has deepened our understanding of the tsunami-generating source mechanisms, tsunami propagation and inundation process and enabled the development of sophis-

ticated numerical simulation programs. But at the same time, complexities of focal process and tsunami behavior, especially near the coast, have also been revealed. Therefore, careful consideration of the complicated nature of tsunami phenomena is essential in order to make tsunami forecast contents and the dissemination of warnings effective for disaster mitigation.

## Introduction

A tsunami is an oceanic gravity wave generated by submarine fault dislocation or other origins such as mud or rock slumps on steep continental margin slopes, marine volcanic eruptions and others. A large tsunami may cause disasters along densely populated or built-up coasts and sometimes also by the inundation of low land areas, up to several km inland. Most tsunamis are generated by large earthquakes occurring in oceanic areas, and it is possible to estimate tsunami generation to a certain extent from seismic wave analysis. By taking advantage of the propagation velocity difference between the much faster seismic and the slower tsunami waves, it is possible to mitigate tsunami disasters by issuing tsunami forecast before the tsunami arrives at the coast, thus enabling evacuation and other countermeasures. For other origins of tsunami, it is still difficult to quantitatively forecast tsunami generation until it is observed actually by the sea level change sensors.

In earlier years, an empirical method had been used to estimate tsunami amplitude via a regression formula that relates tsunami amplitude to the magnitude of the earthquake and its epicentral distance from the coast of interest. Recently, significant progress has been made in the understanding of the tsunami characteristics. This permits numerical simulation of the tsunami propagation once the initial tsunami wave distribution in the source area is correctly known, and bathymetry data are available with sufficient spatial resolution. In Japan, numerical simulation technique has been used in the operational tsunami forecasting procedure in the Japan Meteorological Agency (JMA) since 1999. Efforts have been made also in other countries to incorporate numerical simulation technique in tsunami forecasting. In this article, mainly based on JMA experience, procedures are introduced how to conduct tsunami forecast service by using numerical simulation techniques and giving due consideration to the complexity of this problem.

## Complexity Problem in Tsunami Forecasting

The complexity of tsunami forecasting with numerical methods is twofold and mainly due to the following:

- The uncertainty of the initial tsunami wave distribution in the source area
- The complexity of bathymetry and coastal topography

### Uncertainty of the Initial Tsunami Wave Distribution

For a local tsunami, only limited time is left from the generation of tsunami to its arrival at the coast. In order to assure maximum lead time for evacuation, this necessitates the tsunami forecast to be based on the earliest available data from seismic wave analysis. On the other hand, in the case of a distant tsunami sources, a relatively long lead time is left until the tsunami strikes the coast, and data of sea level change can be recorded in near real-time by tidal stations on the way from the source to the coast. This allows the tsunami forecast to be based on the actual observation of generated tsunami waves and thus to reduce the uncertainty in the initial spatial distribution of the tsunami wave in the source area. Also for local events it is possible to improve the accuracy and reliability of the tsunami forecast in a step-by-step manner, if more detailed data and reliable analysis results become available after the first forecast has been issued. Therefore, the most practical approach in view of the disaster mitigation is to assure the rapid issuance of the first forecast based on the seismic wave analysis, and then to update it with improved data.

There are several uncertainties when the initial tsunami wave distribution is inferred from seismological data analysis alone.

### Uncertainty of the Relative Location of the Hypocenter in the Rupture Area

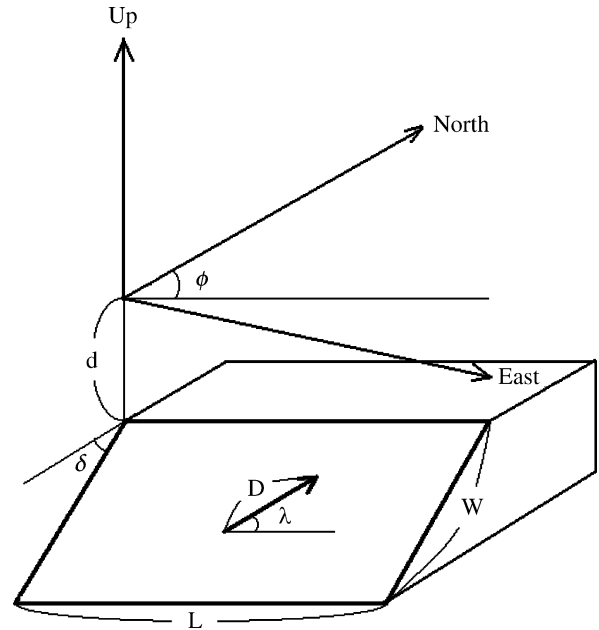
The hypocenter of an earthquake is merely the location from where the rupture starts. It is calculated from first arriving seismic waves. Depending on the direction of rupture propagation (e.g., unilateral, bilateral, radial) and the complexity/irregularity of the actual fault rupture, the seismologically determined hypocenter may neither lie in the center of the rupture area nor coincide with the area of maximum fault displacement. The bigger the earthquake and thus the larger the spatial extent of the rupture area, the less representative the seismologically derived hypocenter as a parameter for characterizing the earthquake rupture is. To reduce this uncertainty, some methods have been developed, like a rupture process inversion by using teleseismic broadband seismogram [20], or near-field strong motion data [11]. A fault model inversion by using a co-seismic step estimated by real-time GPS data analysis is under development [5,22,25]. And the spatial extent of the tsunami source area can also be constrained by using inverse refraction diagrams from tidal stations [1] or deep ocean wa-

ter pressure sensor (DART (Deep-ocean Assessment and Reporting of Tsunamis))-type stations [6] that recorded tsunami arrivals as well as by satellite sea surface altimetry data taken over the source area [7,10].

In general, the rupture process is complicated, and the slip distribution on the fault is not uniform. Non-uniformity of the slip distribution can make the initial tsunami wave distribution much more complicated, especially in its short wavelength component, and may strongly affect the tsunami behavior on the coast. To estimate the slip distribution, in addition to the seismological and geodetic methods mentioned above, fault slip distribution inversion techniques have been developed which compare actually observed tsunami waveforms with pre-calculated theoretical tsunami waveforms from unit fault dislocation segments as Green functions [26] ► **Tsunamis, Inverse Problem of.**

**Uncertainty of the Magnitude** The maximum amplitude of a seismic body or surface wave group is usually used for estimating the earthquake magnitude (see this volume) as a measure of earthquake size in a certain period range, which is often limited by the bandwidth of the seismometer response. Therefore, the tsunamigenic potential may be underestimated, especially for gigantic earthquakes (such as the Sumatra-Andaman Mw9.3 earthquake of 26 December 2004) and “tsunami earthquakes” that generate bigger tsunami than expected from its shaking potential and magnitude values in the short-period range. To overcome this difficulty, some fast body-wave methods have been developed that consider besides single maximum amplitudes also multiple rupture amplitudes over the rupture duration [2], ► **Earthquake Magnitude**, integrated seismic energy and/or total rupture duration time [8,21], while others use successfully long-period spectral mantle surface-wave amplitudes in real-time applications in the context of tsunami warning [38], and  $W$ -phase for quick estimation of CMT solution for a practical tsunami warning service [17]. Besides these seismological methods for estimating earthquake magnitude, geodetic methods that use the co-seismic step or ultra-long period component observed by crustal deformation sensors [5,13,18] play an increasing role for estimating the size of great earthquakes.

**Uncertainty of the Fault Parameters** It is difficult to estimate fault geometry promptly after the earthquake occurrence. Sea floor deformation due to fault motion is usually assumed to be identical to initial tsunami wave distribution. Accordingly, the latter is significantly affected by the specifics of fault geometry, fault and slip orienta-



Tsunami Forecasting and Warning, Figure 1  
Definitions of fault parameters.  $L$ ,  $W$ ,  $D$  and  $d$  denote the fault length, width, average slip amount and depth of the top margin of the fault respectively. The fault dip  $\delta$  is measured down from horizontal, strike  $\phi$  clockwise round from North and slip (rake)  $\lambda$  anti-clockwise round from strike direction along the fault plane (motion of the hanging wall relative to the footwall)

tion in space as well as rupture complexity. These complexities are usually neglected and the real fault rupture is roughly approximated by a dislocation model that is represent by a rectangular fault that is described by 6 parameters, namely the fault length  $L$ , width  $W$ , average slip amount  $D$ , and the three angles of fault dip  $\delta$ , strike  $\phi$  and slip (rake)  $\lambda$  (Fig. 1).

**Fault Length, Width and Average Slip Amount** Different scaling laws exist between  $L$ ,  $W$  and  $D$  with earthquake magnitude (e. g., [3,4,16,37,39]). They allow us to roughly estimate these values from determined magnitudes. Such scaling relations are based on the assumption of more or less constant stress drop (i. e. High stress drop earthquake has larger  $D$  for the same  $L$  and  $W$  of low stress drop earthquake). But the stress drop is different between inter-plate thrust events and intra-plate events. In general, intra-plate event has higher stress drop than inter-plate event. Such differences may cause significant errors when estimating these fault parameters via scaling relations with magnitudes. Best estimates of these parameters, that characterize the earthquake size, are possible via the determination of the seismic moment  $M_0 = \mu DA$  with  $\mu$  being the rigid-

ity of the crustal/lithosphere material in the rupture area  $A = L \times W$ , and  $L$  and  $W$  estimation by rupture process analysis described in Sect. “Uncertainty of the Relative Location of the Hypocenter in the Rupture Area”.

*Fault Dip, Strike and Slip Angle* Representative values for dip, strike and slip angles can be set based on the analysis of past great earthquakes having occurred in each region. The fault plane of a newly occurred earthquake is likely to be close to either one of the two nodal planes derived from centroid moment tensor (CMT) solutions. But to select from these two candidates the real acting fault plane, precise aftershock location or rupture process analysis as described in Subsect. “Uncertainty of the Relative Location of the Hypocenter in the Rupture Area” are necessary.

### Complexity due to Complicated Bathymetry

If the spatial distribution of the initial tsunami wave is given correctly, the tsunami propagation can be forecast in a deterministic manner by using numerical simulation technique. In a long-wave approximation, which considers only wavelengths of the tsunami that are much larger than the sea depth, the propagation velocity  $v$  of the tsunami is represented by  $v = \sqrt{gd}$ , where  $g = 9.81 \dots \text{m/s}^2$  is the Earth’s gravity acceleration and  $d$  the sea depth in m. This approximation is valid for most tsunami cases. As wavelength is the product of velocity times period, tsunami wavelength is proportional to the square root of sea depth. Accordingly, the wavelength of the tsunami become shorter as it approaches the coast because the sea depth becomes shallower, and the wave becomes much more sensitive to finer bathymetry changes and coastal feature. Therefore, a finer mesh of bathymetry and coastal features is necessary near the coast in order to represent the tsunami behavior correctly. But, even if very fine mesh bathymetry data are available, it is not appropriate to use it in an early stage of tsunami propagation simulation when the initial tsunami wave distribution is still uncertain as mentioned in Subsect. “Uncertainty of the Initial Tsunami Wave Distribution”.

Further, very fine-mesh simulations require a substantially long time for the conduct of simulation. Time constraints are crucial when incorporating numerical simulation techniques in an operational tsunami forecast procedure, except for distant events. Therefore, especially for local events, it is most practical to conduct tsunami simulations for a variety of scenarios in advance, to store these results in a database, and to conduct tsunami forecast by retrieving the most appropriate case for the determined hypocenter. As described later, the JMA has adopted this

way. Even in that case, the mesh size to be used in the simulations must be carefully examined taking into account the required accuracy and spatial resolution in the tsunami forecast for guiding disaster mitigation efforts and the need to complete the simulations for all scenarios in a realistic time span.

When very fine mesh is used, depending on the complexity of the bathymetry and topography near the coast, tsunami amplitude distribution along the coast shows significant scatter, and sometimes, extremely large tsunami amplitude will be estimated very locally. Therefore, it must be clearly defined in advance as to what kind of statistical average value, using a finite number of estimated tsunami amplitudes for the area of interest, should be adopted in the tsunami forecast.

### Components of a Tsunami Early Warning System (TEWS)

In general, a tsunami early warning system consists of the following constituents:

1. Seismic network (seismometers and real-time data transmission link)
2. Real-time seismic data processing system for hypocenter and magnitude determination
3. Tsunami forecast system (including warning criteria, assembling of the text, and dissemination)
4. Sea level data monitoring system (tide gauge/tsunami-meter and real-time/near real-time data transmission link)

Such a composition is adopted in Japan, USA, Russia, Chile, Australia, French Polynesia, New Caledonia and in other countries around the Pacific Ocean. The up-to-date status of the TEWS of these countries can be consulted on the UNESCO/Intergovernmental Oceanographic Commission’s (IOC) website as based on the national reports submitted to the latest meeting of the Intergovernmental Coordination Group/Pacific Tsunami Warning System (ICG/PTWS). Also in Indian Ocean countries, after the tremendous tsunami disaster brought by the Great Sumatra Earthquake in 2004, efforts have been made to establish tsunami early warning systems of the similar composition in different countries (e. g. GITEWS: German Indonesian Tsunami Early Warning System, see <http://www.gitews.org/>). Similar efforts are under way in the Caribbean, North Eastern Atlantic and Mediterranean regions too, inspired by the Sumatra event and taking into account potential tsunami risk assessments for these areas based on historical records.



To accomplish tsunami warning for local events, local seismic networks (see Subsect. “**Seismic Network**”) and real-time seismic data processing systems (see Subsect. “**Real-Time Seismic Data Processing System**”) are indispensable. For a distant event, one can utilize the hypocenter parameters contained in the international tsunami watch information provided by the Pacific Tsunami Warning Center (PTWC), West Coast/Alaska Tsunami Warning Center (WC/ATWC) of US and North West Pacific Tsunami Advisory Center (NWPTAC) of the JMA.

As a typical example, technical details of Japan’s Tsunami Early Warning System, in which all four constituents are fully deployed, is explained in the following section.

As for the item 4 (sea level data monitoring system), the US and also GITEWS are now deploying DART type of buoys [6] system. This is to measure a sea surface vertical displacement by observing the related pressure change at the ocean bottom. DART buoys are placed in far offshore regions where bathymetry is simple, and very simple tsunami waveforms can be observed without the influence of complicated bathymetry near a coast. Such measured data are preferable for a comparison with simulated waveforms. And by placing such buoys with proper spacing at a certain distance from the tsunamigenic zone over the deep ocean basin, tsunami waves, that are sufficiently separated in time from seismic waves, can be observed early enough after the earthquake occurrence to be useful for tsunami EW from distant sources.

The Pacific Marine Environmental Laboratory (PMEL) of the National Oceanic and Atmospheric Administration (NOAA) is developing a tsunami forecasting system named SIFT (Short-term Inundation Forecasting) [33] based on the DART buoy data. Like in the JMA’s tsunami simulation database, described in detail below, tsunami simulations are conducted for many different unit faults located along the tsunamigenic zone, and the calculated tsunami waveforms at DART buoy locations are stored in a database as Green functions, together with simulated waveforms at coastal points. In the case that a tsunami wave is observed by DART buoys, the observed waveforms are represented by a linear combination of several Green functions. Then, tsunami amplitudes at coastal points can be estimated from the linear combination coefficients and the simulated waveforms at coastal points (see web-page of NOAA/PMEL (<http://nctr.pmel.noaa.gov/>) for applications of their method to recent actual tsunami events). This system is planned to be introduced in the actual tsunami forecasting procedures of the US tsunami warning centers.

In Japan, a similar study is in progress at Tohoku-University [35]. They conduct tsunami simulations originat-

ing from unit vertical sea surface displacements in a unit sea area, and the resulting simulated waveforms at offshore tsunami meters are stored as Green functions in a database. They do not adhere to a specific fault model. The procedure for the estimation of tsunami waveform at coastal points is similar to that of SIFT.

### Tsunami Early Warning System in Japan

Before describing the present status of the JMA’s TEWS, a historical review is briefly given.

The JMA started tsunami warning service in 1952. Since then, the JMA has been making an effort to integrate its TEWS, and this is exactly a history of ‘fight against time’.

At the commencement of the tsunami warning service, the JMA had 46 seismometers at meteorological observatories (basically near populated area on a sedimental layer). In case of earthquake occurrence, seismograms were read at observatories, and their results (*P* and *S* arrival times and maximum amplitudes) were transmitted to the headquarters in a telegram format. At the headquarters, hypocenter and magnitude were determined manually, and the tsunami warning grade was determined by using an empirical chart based on the relation between tsunami amplitude, earthquake magnitude and epicentral distance to a coast of the past events. It took about 15 to 20 minutes to disseminate a tsunami warning.

From the late 1960s to early 1980s, telemetry technology to transmit seismic waveform data from observatories to the headquarters and a processing computer were introduced. Detection capability of the earthquake was improved by monitoring collected seismic waveforms at one place, and the *P* and *S* picking precision was also improved by introducing a digitizer. By these, dissemination time was reduced to 12 to 13 minutes.

In 1983, the Mid-Japan-Sea earthquake (Mjma7.7) occurred. The JMA disseminated a tsunami warning in 14 minutes, but in about 7 minutes, the tsunami struck the nearest coast, killing 104 people. After this event, the JMA deployed a more sophisticated computer system for the seismic waveform processing. A graphical man-machine interface was introduced for more accurate and quicker phase reading and hypocenter and magnitude calculation. Still, the empirical method was used for the tsunami amplitude estimation. Dissemination time was reduced to between 7 and 8 minutes.

In 1993, the Southwest off Hokkaido earthquake (Mjma7.8) occurred. The JMA disseminated a tsunami warning in 5 minutes, but the tsunami struck Okushiri-Island in a few minutes, and 230 people were killed or lost by

tsunami. After this event, the JMA totally replaced the seismic network. All seismometers were installed in remote un-manned sites on hard rock, and the total number of the site increased to about 180. Dissemination time was reduced to 3 to 5 minutes.

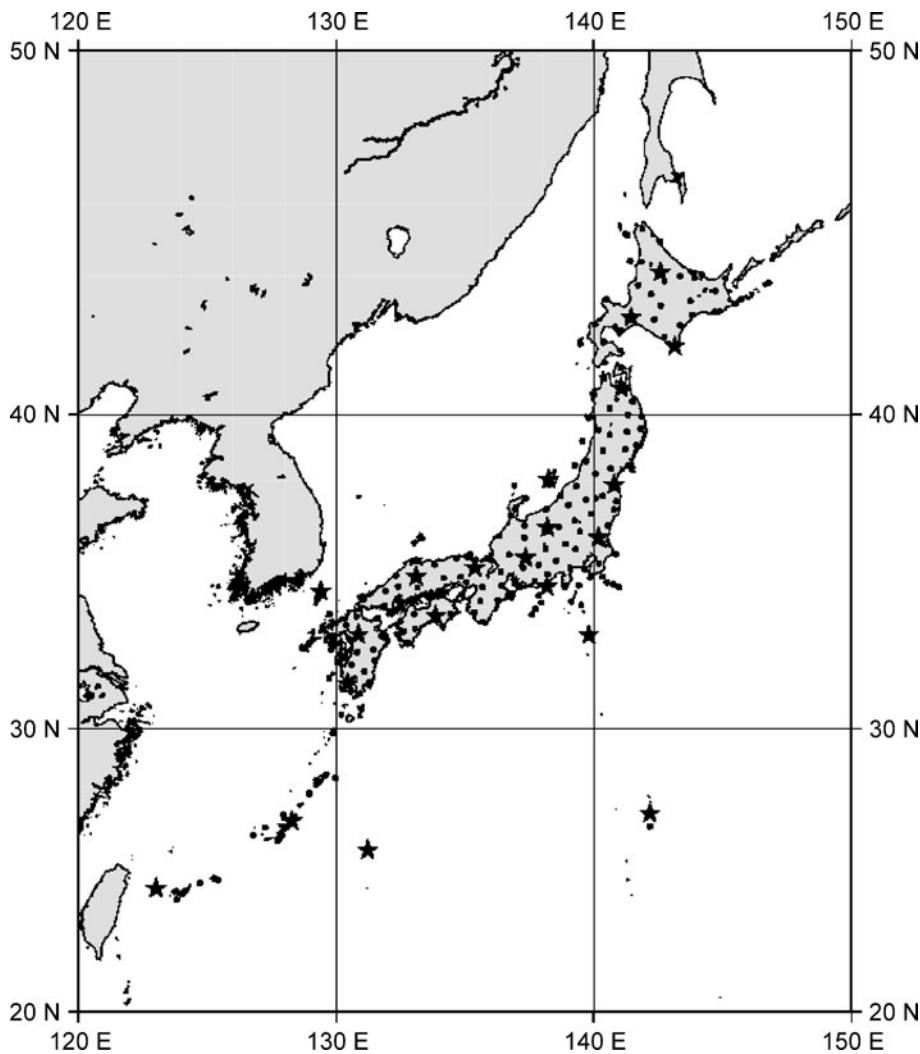
The present status of the JMA's TEWS is as follows.

### Seismic Network

The JMA operates about 180 seismometers installed in Japan (Fig. 2). The seismic waveform data are sent to JMA

continuously on a real-time basis through dedicated telephone lines.

There are two kinds of sensors deployed at each station: A short-period velocity sensor and an accelerometer. The records of the short-period sensor are mainly used for precise picking of the onset times of *P* and *S* phases which are required for an accurate hypocenter determination. Accelerometer records are used for the calculation of magnitude of large earthquakes in the case that the ground motion amplitude exceeds the dynamic range of the short-period sensor.



Tsunami Forecasting and Warning, Figure 2

Seismic network of the JMA used for tsunami forecasting. *Solid circles* and *stars* denote the locations of seismic stations operated by the JMA. The average spacing between the about 180 stations is 50 to 60 km. *Stars* denote the location of stations where STS-2 velocity broadband seismometers have been installed in addition to the Japanese short-period velocity-type seismometers and accelerometers

STS-2 broad-band seismometers have additionally been installed at 20 stations. These broadband velocity seismic records are mainly used for obtaining CMT solutions and moment magnitudes  $M_w$ .

### Real-Time Seismic Data Processing System

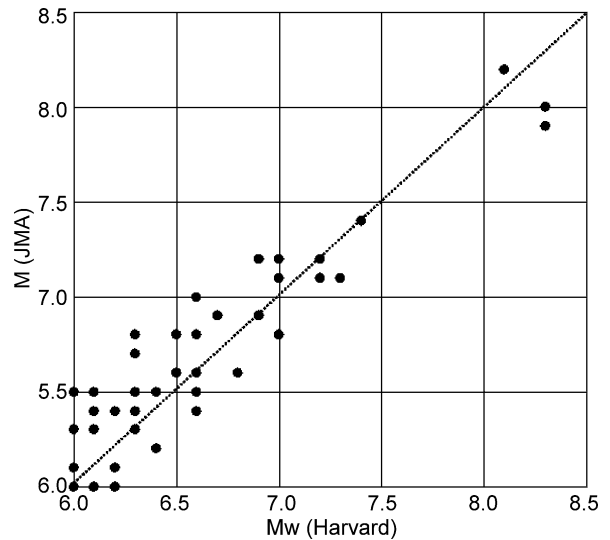
The JMA applies the short term average/long term average (STA/LTA) method as trigger criteria. It is most widely adopted in the seismological community and used to limit seismic recordings to relevant seismic event waveforms. When the ratio of STA/LTA exceeds the trigger threshold, it is supposed that a significant signal, possibly an earthquake signal, has been detected by the considered station.

STA/LTA trigger may occur due to some troubles or local noise at one station. To avoid “false trigger”, the JMA also applies other criteria including the group trigger criterion. A group is set to include several neighboring stations. When the STA/LTA trigger turns on at a certain number of stations in one group, the system considers it to be caused by an earthquake.

For the precise picking of the  $P$  onset, an auto regressive (AR) model is applied to represent seismogram time series. Akaike’s information criterion (AIC) is used to search the optimal time to separate the seismogram into two sub-stationary time series (noise part and signal part), which are represented by AR models, respectively, in a certain time window around trigger time [32,40].

The usual least square method is used in the hypocenter calculation. As for the velocity structure, the Japan local standard model JMA2001 [36] is applied. To better locate the hypocenter,  $S$  arrival times are used in addition to  $P$  arrival times. First, a preliminary hypocenter is calculated by using only  $P$  arrival times picked by an automatic picker. Then,  $S$  arrival times are picked automatically by applying the AR model and AIC in a time window centered at the theoretical  $S$  arrival times estimated from the preliminary hypocenter location and origin time. The final hypocenter is then calculated by using both the  $P$  and  $S$  arrival times.

For magnitude calculation, one of the local magnitude definitions in Japan, the JMA magnitude ( $M_{jma}$ ) according to Katsumata [19] is used, in which average amplitude decay characteristic in Japan is taken as a decay correction term of the formula. Maximum amplitude of the horizontal ground displacement is the measured input value in the formula. The displacement waveform is calculated by applying real-time recursive filter that integrates twice the accelerometer data and applies a high-pass filter to avoid computational instability. Due to historical reasons, the long-period cut-off of the high-pass filter is set at



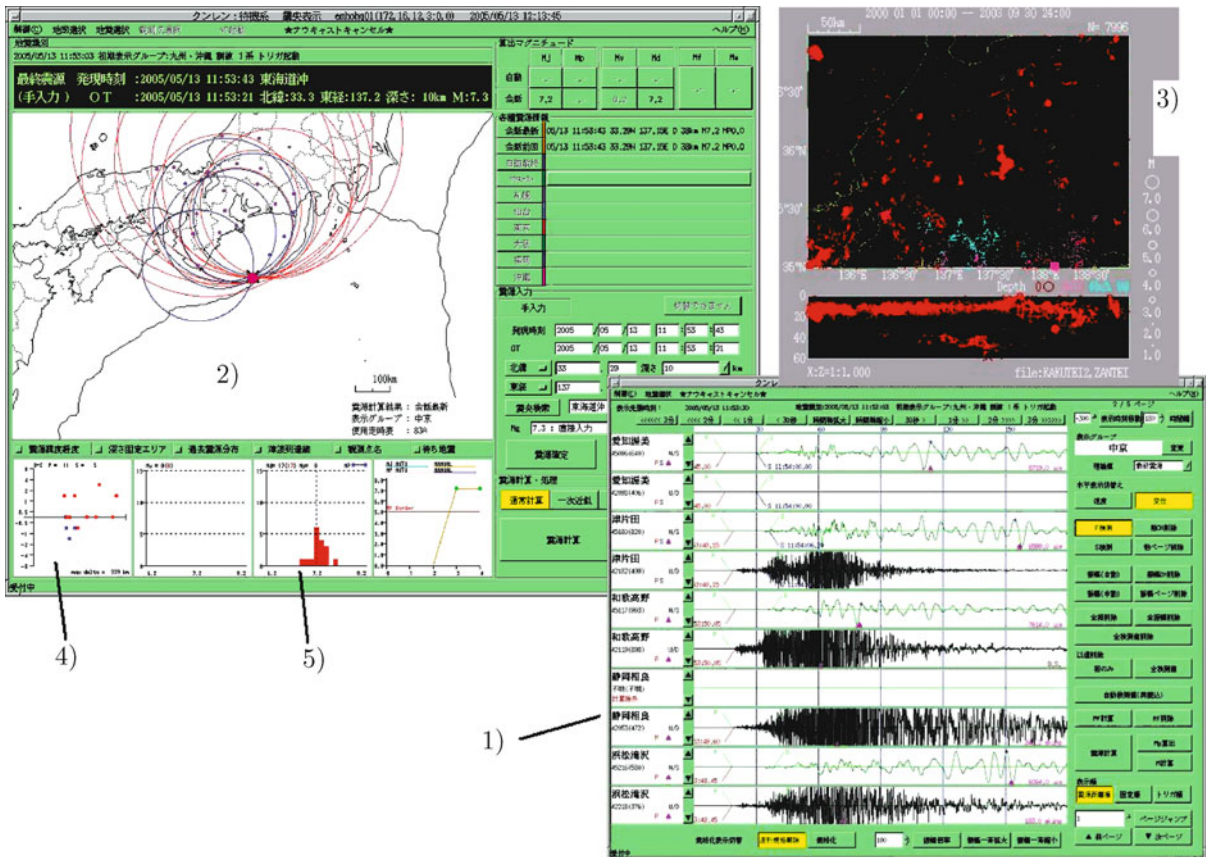
Tsunami Forecasting and Warning, Figure 3  
Relation between  $M_w$  (Global CMT) and  $M_{jma}$ , based on events that occurred after September 2003 in and around Japan. Both magnitudes agree on average well up to around magnitude 8.0

5 seconds. Since unsaturated strong-motion data are used, magnitude calculation can be started promptly after the earthquake occurrence by using data from stations very close to the epicenter. Therefore, such magnitude data are suitable for the tsunami warning purposes of local events.

The comparison between  $M_{jma}$  and the moment magnitude  $M_w$  of the Global CMT Project’s solution is shown in Fig. 3.  $M_{jma}$  values are on average comparable with  $M_w$  for earthquakes with magnitudes between  $M_{6.0}$  and about 8.0.

All procedures mentioned above are run automatically by the computer. Considering the possible impact of a tsunami warning on the potentially affected communities, a human check of the adequacy of the calculated results is indispensable for minimizing false alarms. The following items are checked by the operators, using a man-machine interface:

1. Adequacy of the phase pickings and maximum amplitude readings: The seismic waveforms are plotted together with picked and theoretical arrival times for  $P$  and  $S$ , and the marked position at which the maximum amplitude was measured. This plot also allows them to discriminate noise from seismic signals.
2. Adequacy (mathematical) of the hypocenter location: A hypocenter plot map is used in which the locations of used seismic stations and the circles denoting the calculated epicentral distance from each station are shown. If the circles intersect densely at one point, and if the



Tsunami Forecasting and Warning, Figure 4

Examples of the JMA's man-machine interface screen image in the seismic data processing system. Numbers in the figure are related to the respective numbers of items explained in the main text

- used stations assure a wide azimuthal coverage around the source, then the calculation is judged as fine.
- Adequacy (seismological) of the hypocenter location: The hypocenter plot on the background seismicity map is used together with the vertical cross section. If the hypocenter is located in a region where no background seismicity is found, it should be re-examined carefully.
  - Adequacy of the depth estimation: Travel time residuals are plotted as a function of epicentral distance. If the residuals depend on epicentral distance, the depth estimation should be revised.
  - Adequacy of the magnitude calculation: A histogram of station magnitudes is used. Outliers are checked and excluded in the re-calculation.

Sample images of the man-machine interface are shown in Fig. 4.

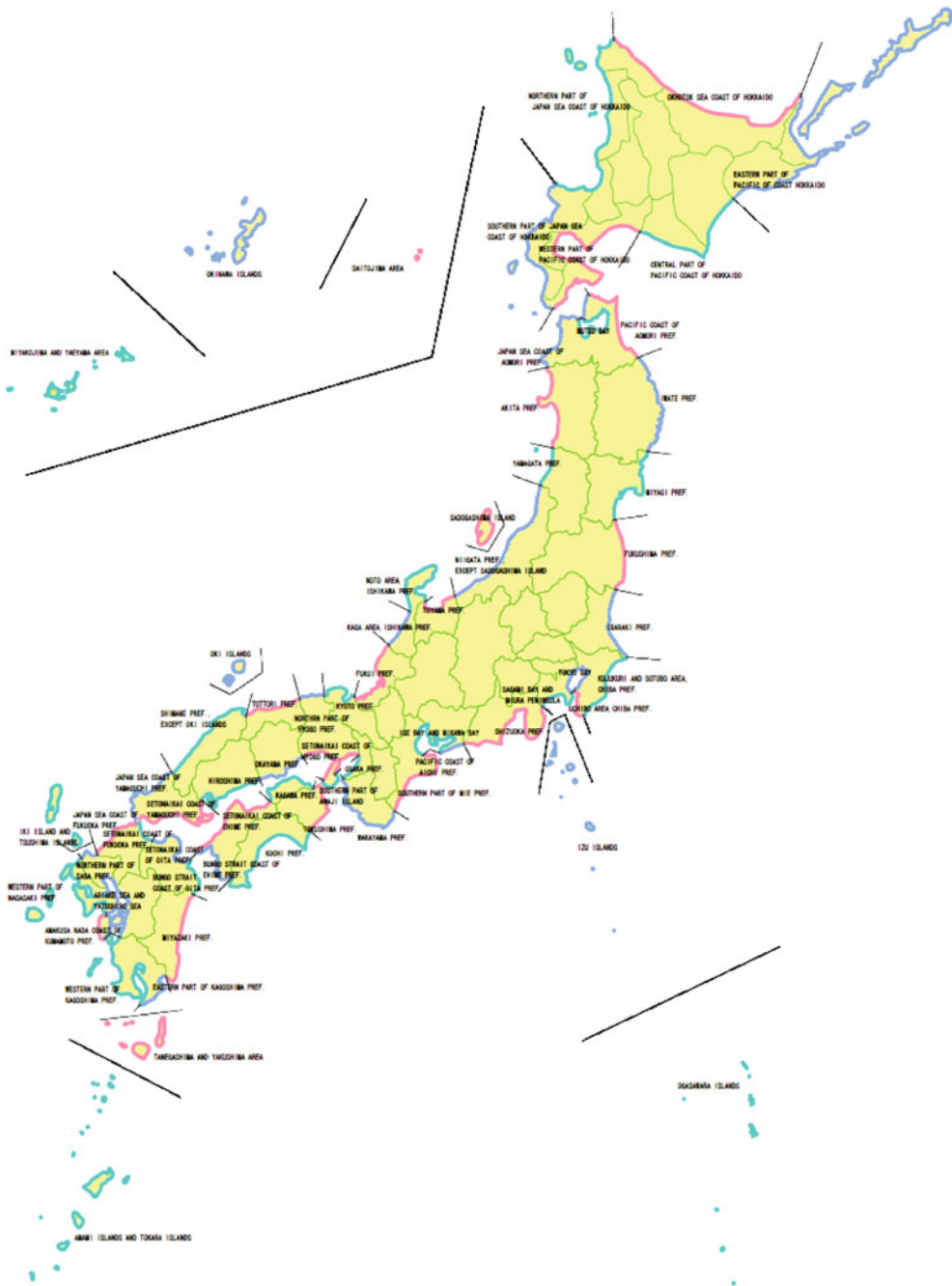
The JMA's man-machine interface is designed so that re-picking and recalculation of hypocenter and magnitude can be done within about 30 seconds.

Calculated hypocenter parameters (latitude, longitude, depth and magnitude) are transferred to the following tsunami forecast system as its input.

### Tsunami Forecast System

**Tsunami Forecast in Japan** For a tsunami forecast to be useful, a mere general description of the tsunamigenic potential estimated from calculated hypocentral parameters is not sufficient. From the issued message, the recipient disaster management organization should be able to clearly grasp the severity of the expected disaster in its own jurisdiction. In Japan, a tsunami forecast is disseminated to each coastal block (the coast line of Japan is separated into certain number of blocks), giving three kinds of grade corresponding to the anticipated severity of the disaster.

**Tsunami Forecast to Coastal Block** The coastal blocks divide the Japanese coasts into 66 regions (Fig. 5). These blocks are defined by taking into account:



Tsunami Forecasting and Warning, Figure 5  
 Coastal block partitioning for tsunami forecast in Japan. *Straight lines* denote the borders between the 66 coastal blocks

- a) The administrative districts of local governments, which are the disaster management units to take actions in emergency situations
- b) The sea areas to which each coast is facing
- c) The uniformity in the behavior of tsunami
- d) The precision of tsunami forecast technique

Basically, one coastal block corresponds to one prefecture, or finer.

Before the introduction of numerical tsunami simulation technique, the total number of coastal blocks was 18, due to a low precision of empirical tsunami amplitude estimation method.

**Tsunami Forecast Criterion and Category** The JMA categorizes tsunami forecast into “Tsunami Warning” and “Tsunami Advisory”. Further, “Tsunami Warning” is divided into two grades, “Tsunami” and “Major Tsunami”.

In Japan, there exist ample number of materials on the relation between tsunami disaster and observed tsunami amplitude along the coast. From the materials, high correlation between these two can be seen. Therefore, estimated maximum tsunami amplitude at the coast is used as a criterion of the tsunami forecast in Japan. The speed of water current in the sea near the coast could be another good index for assessing the tsunami disaster potential. But, up to now, only a few examples of the current observation are available. Therefore, the estimated tsunami amplitude at the coast is still the most appropriate criterion for tsunami forecast.

In Japan, tsunami disaster in the land area occurs when the tsunami amplitude exceeds 1 meter [30], so the warning criterion is set at 1 meter. When tsunami amplitude exceeds 3 meters, the proportion of damaged wooden houses and fishing boats increases remarkably [9,28,29]. Therefore, the JMA issues the highest warning grade “Major Tsunami” when the estimated maximum tsunami amplitude is 3 meters or higher. When the estimated maximum tsunami amplitudes range between 1 meter to 3 meters, the grade name is “Tsunami”. If a “Tsunami Warning” is issued, the head of municipality must order evacuation to the residents in the “Tsunami Evacuation Zone” assigned by each municipality in advance.

Even when the tsunami amplitude is less than 1 meter, it is occasionally the case that bathing people are affected and aquaculture rafts for marine products industry are damaged. Considering safety criteria in sea bathing places (beaches, bays), and the relation between the tsunami amplitude near the coast and damage on aquaculture rafts in the past, the JMA’s criterion for “Tsunami Advisory” is set at 20 centimeters. In the case that a “Tsunami Advisory” is issued, there is no need of evacuation in the land areas, except vulnerable very low-land areas, but people on beaches or swimming should go to higher places.

Table 1 shows the tsunami forecast grades and levels of the estimated maximum tsunami amplitude that are used by the JMA when issuing tsunami information for each coastal block after a tsunami forecast has been made.

When the estimated tsunami amplitude is less than 20 cm, no warnings or advisories are issued, but a message “No threat of tsunami disaster” is promptly provided to the public in order to prevent unnecessary spontaneous evacuations in accordance with the widely spread recommendation in Japan: “When you feel a strong shaking near

Tsunami Forecasting and Warning, Table 1

**Tsunami forecast grades and corresponding levels of expected maximum tsunami amplitudes as used in tsunami information issued by the JMA for each coastal block after a tsunami forecast has been made**

Forecast Grade		Levels of Estimated Tsunami Amplitude
Warning	Major Tsunami	3 m, 4 m, 6 m, 8 m, 10 m or greater
	Tsunami	1 m, 2 m
Advisory		0.5 m

a coast, run to a high place without waiting for tsunami forecasts from the JMA”.

**Tsunami Database Creation** The JMA introduced in 1999 a numerical simulation technique to implement accurate tsunami warning. However, even most-advanced computers require substantial time for the completion of the calculations. Therefore it is impossible to issue prompt tsunami warning based on numerical simulation technique even if the calculations start simultaneously with the occurrence of an earthquake.

Alternatively, the JMA employs a database method to achieve a breakthrough on this issue. Tsunami propagation originating from various locations, fault sizes and likely rupture mechanisms are calculated in advance and the results, namely estimates of maximum tsunami amplitudes and arrival times, are stored in a database together with the associated hypocenter location and magnitude.

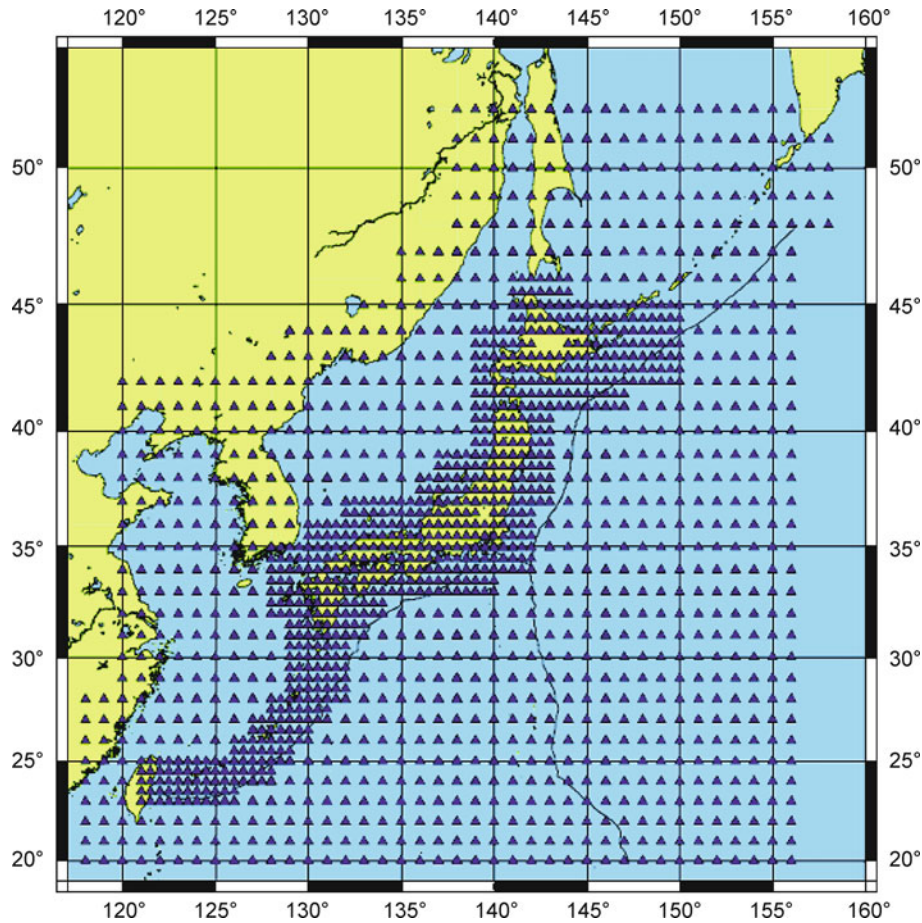
If an earthquake occurs, the most appropriate case for the actual hypocenter location and magnitude is retrieved from the database and the tsunami forecast is issued accordingly.

**Hypocentral Location** The first step to create a tsunami database is to set fault models. ‘Simulation points’ are defined as surface projection locations of the center of possible causing faults, and are placed in the areas where tsunamigenic earthquakes are likely to occur. These points are selected on the basis of background seismicity and the records of the past earthquakes that generated tsunami.

Figure 6 shows the locations of simulation points in Japan and surrounding areas.

The total number of scenarios calculated for these points is about 64,000 (lat.  $\times$  lon.  $\times$  dep.  $\times$  mag.). Earthquakes with hypocenter depth of more than 100 km or magnitude less than 6.2 are not considered, because they do not cause disastrous tsunami [23].

**Fault Parameters Setting** Figure 1 shows fault parameters used in the numerical simulation. Length  $L$ , width  $W$ ,



Tsunami Forecasting and Warning, Figure 6

Distribution of simulation points in and around Japan. *Solid triangles* denote locations of simulation points to which likely hypothetical fault models are attributed. *Spacing* between the simulation points is 0.5 degrees in near coastal areas, and 1.0 degree in off-shore areas

and slip amount  $D$  of the fault are represented in terms of empirical formulas as functions of magnitude. The following are common relations used in Japan [37] based on  $M = M_j$ ma:

$$\begin{aligned} \log L &= 0.5M - 1.8 \quad \text{with } L \text{ in km,} \\ W &= L/2 \quad \text{or } \log W = 0.5M - 2.1 \quad \text{with } W \text{ in km,} \\ \log D &= 0.5M - 3.3 \quad \text{with } D \text{ in m.} \end{aligned}$$

These formulas are consistent with the next two formulas, assuming a common rigidity  $\mu = 2.0 \times 10^{10}$  Newton/m<sup>2</sup> of the rock-material in the source area

$$\begin{aligned} M_0 &= \mu \times L \times W \times D \\ \log M_0 &= 1.5M_w + 9.1 \end{aligned}$$

(in international standard units, i. e.,  $M_0$  in Nm = Newton meter).

The dip ( $\delta$ ), strike ( $\phi$ ) and slip ( $\lambda$ ) angles of the fault at each simulation point are based on average values of past earthquakes at or near these locations. But if they are uncertain, they are assumed to be those of a pure reverse fault ( $\lambda = 90$  degrees) whose strike is parallel to the trench or nearby coast, and with a dip angle of 45 degrees, corresponding to an efficient tsunami generation in view of disaster management.

*Initial Value for Numerical Simulation of Tsunami Propagation* The vertical deformation of sea floor due to fault motion is calculated by the elastic theory [24], using fault parameters as set above. The initial deformation of the sea surface is assumed to be identical to the vertical deformation of the sea floor and used as input value for the numerical simulation, because in most cases, rupture propagation

velocity is much higher than the tsunami propagation velocity, and the ocean bottom deformation can be treated to occur instantaneously [14].

*Numerical Simulation of Tsunami Propagation* Non-linear long-wave approximation equation with advection term and ocean-bottom friction term is adopted as the equation of motion, and solved together with the equation of continuity as shown below by finite difference method with staggered leap frog scheme [27].

$$\frac{\partial V_x}{\partial t} + V_x \frac{\partial V_x}{\partial x} + V_y \frac{\partial V_x}{\partial y}$$

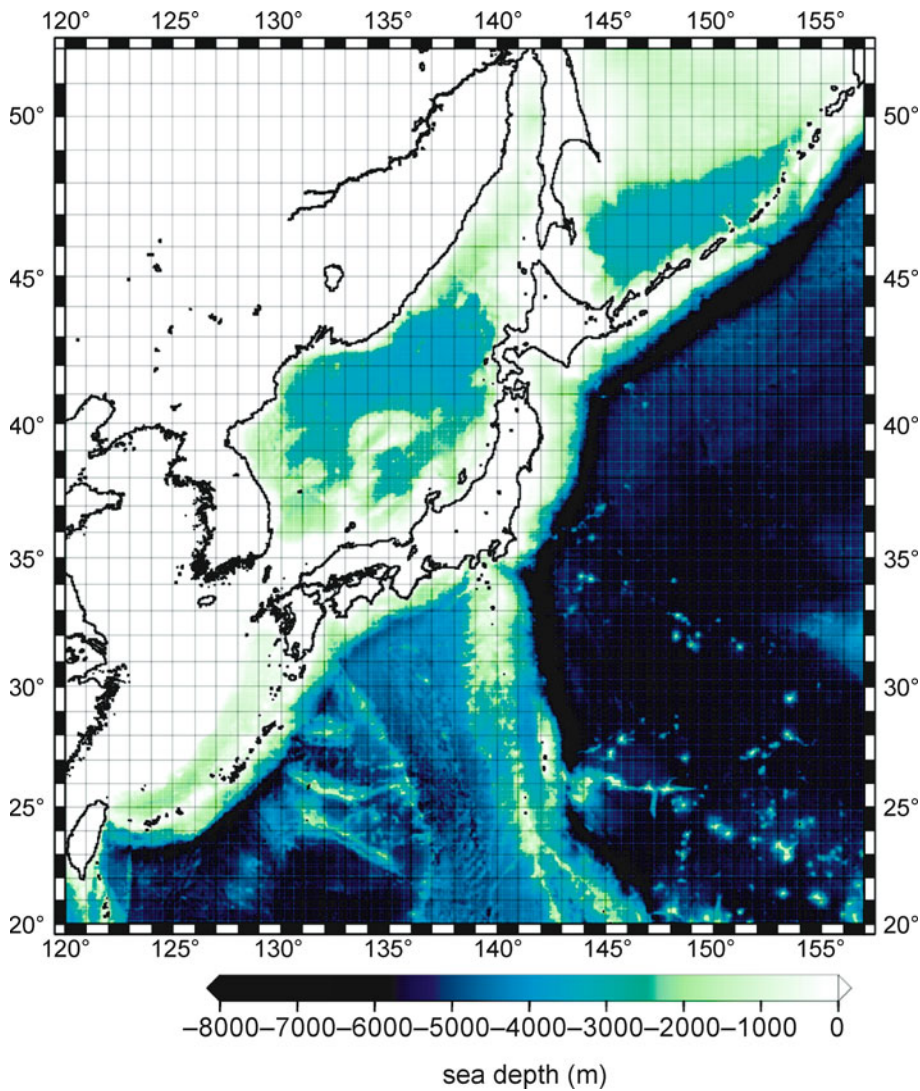
$$= -f V_y - g \frac{\partial h}{\partial x} - C_f \frac{V_x \sqrt{V_x^2 + V_y^2}}{d + h}$$

$$\frac{\partial V_y}{\partial t} + V_x \frac{\partial V_y}{\partial x} + V_y \frac{\partial V_y}{\partial y}$$

$$= f V_x - g \frac{\partial h}{\partial y} - C_f \frac{V_y \sqrt{V_x^2 + V_y^2}}{d + h}$$

$$\frac{\partial h}{\partial t} + \frac{\partial}{\partial x} \{V_x(h + d)\} + \frac{\partial}{\partial y} \{V_y(h + d)\} = 0,$$

where  $V_x$  and  $V_y$  are the  $x$  (east) and  $y$  (south) components of the average water particle motion velocity in the depth direction, and  $h$  and  $d$  are sea surface displacement



Tsunami Forecasting and Warning, Figure 7  
 Computational area for creating the Japanese tsunami simulation database. Bathymetry is given by a color grade scale. One arc minute mesh bathymetry data are used throughout this area for the tsunami simulation



and sea depth, respectively.  $f$  is the Coriolis parameter ( $= 2\Omega \cos \theta$ ,  $\Omega$  is the angular frequency of Earth's self rotation,  $\theta$  is the co-latitude), and  $C_f$  is the sea bottom friction coefficient.

Advection and ocean bottom friction are considered only in sea areas with sea depth less than 100 m. Coriolis force is considered only in the distant tsunami case (see Sect. “**Tsunami Forecast for Distant Event and Northwest Pacific Tsunami Advisory (NWPTA)**”).

Figure 7 shows the computational area. Mesh size is 1 arc minute throughout the computational area. Eight hours of propagation simulation is conducted for all scenarios with a fixed time interval three seconds for the integration. The Coriolis force is not considered for local and regional tsunami simulations. Total reflection boundary condition is used at the land-ocean boundary, and open boundary condition is used outside of the computational area.

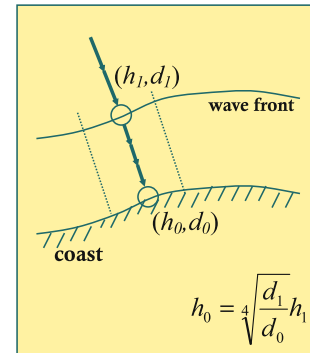
To represent the tsunami waveform correctly in a shallow sea area, very fine bathymetry data mesh is necessary (in a strict sense, 20 or more grid points are necessary within one wave-length [31]), and a vast time is required for the completion of such detailed calculations. To overcome this difficulty, the numerical simulation with the long-wave approximation is applied only to points which are a few to a few ten kilometers seaward from the coast (“forecast points”) where sea depth is about 50 m. Then, tsunami amplitude at the coast is calculated by using Green's law described in the next section.

*Derivation of Tsunami Amplitude at Coast* To estimate tsunami amplitude at the coast, Green's law (= energy conservation law) is applied to the estimated amplitude at the forecast points just seaward of the coast. As shown in Fig. 8, Green's law says that the tsunami amplitude at the coast is represented by fourth root of the ratio of sea depth at the forecast point and at the coast:

$$h_0 = \sqrt[4]{\frac{d_1}{d_0}} h_1 .$$

$h$  and  $d$  denote tsunami amplitude and sea depth, and suffix ‘0’ and ‘1’ denotes the value at the coast and forecast point.

As the forecast points are set close enough to the coast, horizontal convergence and divergence of the rays can be ignored, i. e., the wave front is regarded as almost parallel to the coast due to a refraction towards the direction of the largest sea-depth gradient. As for the “ $d_0$ ” value, the sea depth at the coast, 1 m is assumed, but actually measured value would be more appropriate, if available.

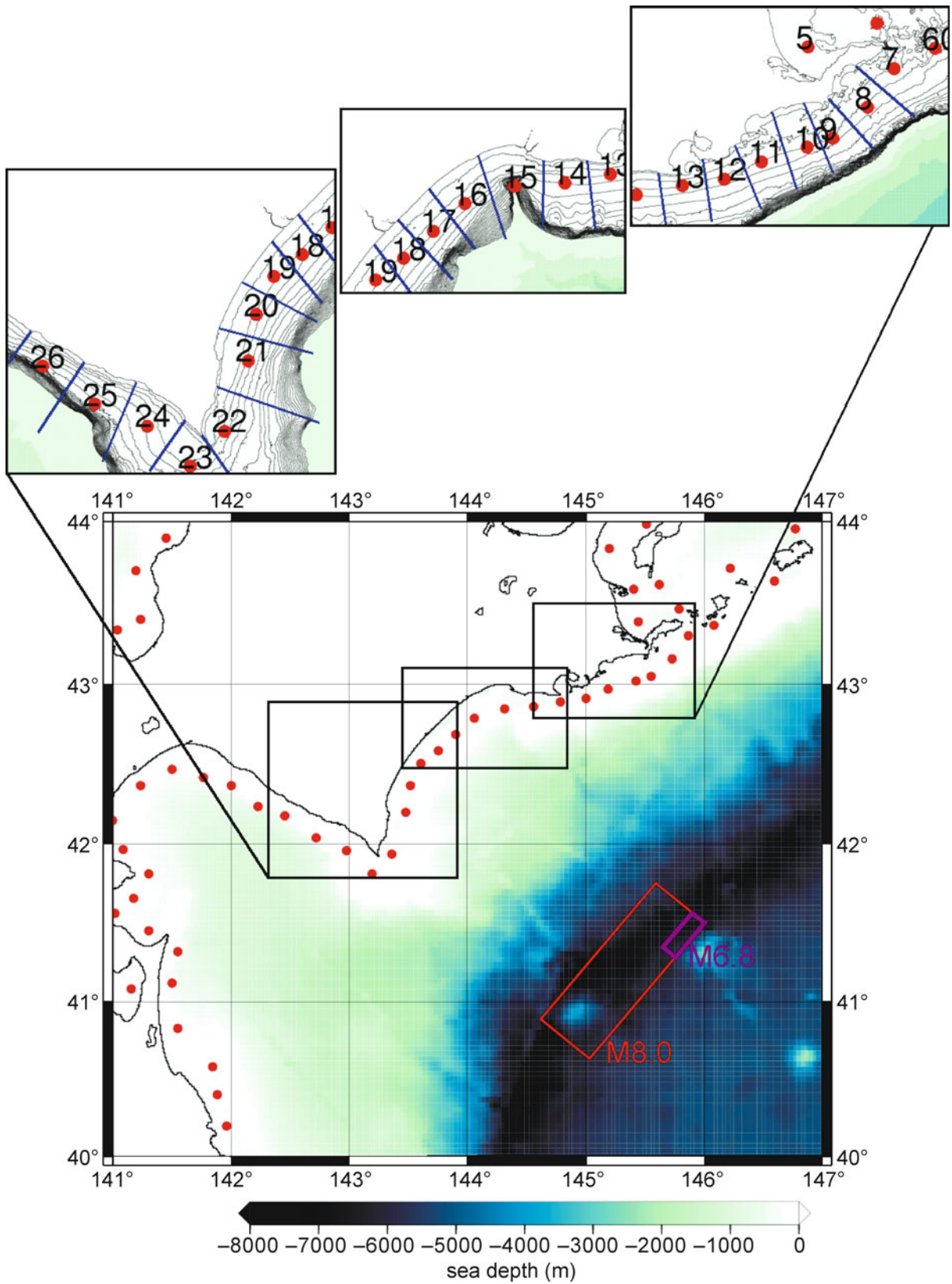


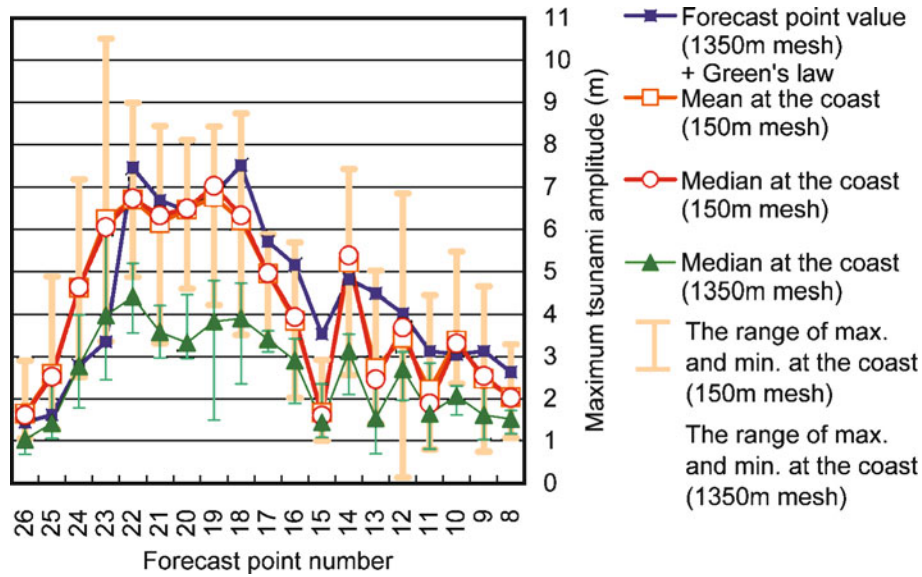
Tsunami Forecasting and Warning, Figure 8  
Schematic sketch to illustrate the application of Green's Law for tsunami amplitude calculation in near coastal areas. Ray convergence and divergence are ignored

*Adequacy of the Green's Law Application* The adequacy of Green's Law application to estimate tsunami amplitude at the coast is examined by using the TSUNAMI-N2 [12] program. It is equivalent to the program used at the JMA [27] to create tsunami simulation database. Up to 150 m mesh at the finest was used by adopting nesting technique, and the tsunami amplitudes estimated at the coastal mesh are considered “true” ones. We compared the case using 1,350 m mesh throughout the computational area (which is equivalent to 1 arc minute mesh) and then estimating the tsunami amplitude at the coast by applying Green's Law to the value at the offshore forecast point, with the expected “true” amplitude at the coastal point using finer mesh. Forecast points actually used for the database creation are shown in Fig. 9. Hypothetical fault models corresponding to two different magnitudes (8.0 and 6.8) have been depicted in this Figure as red and violet rectangles, respectively, south off the Pacific coast of Hokkaido Island. The results are shown in Fig. 10a,b. For

► Tsunami Forecasting and Warning, Figure 9

The map shows the distribution of forecast points (red dots) along the southern and eastern shore of Hokkaido and part of NE Honshu together with the surface projections of two hypothetical faults representing earthquakes with magnitude 8.0 (red rectangle) and 6.8 (violet rectangle), respectively. The depth of the fault top is 1 km from the sea bottom for both cases. Pure reverse fault with dip angle 45 degrees is assumed. In the upper half of the figure the coastal blocks are enlarged, showing the locations of forecast points denoted by numerals and the blue separation lines between sub-sections of the coast. The numbers correspond to the ‘forecast point number’ given on the abscissa of Fig. 10. Also depicted are bathymetry contour lines near the coast at 20 m depth intervals





Tsunami Forecasting and Warning, Figure 10a

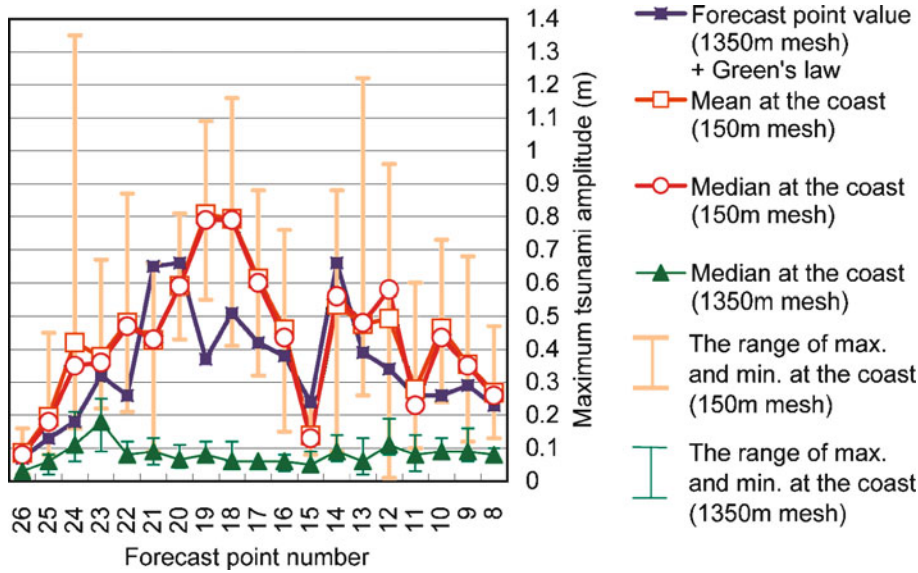
The presented diagram illustrates the adequacy of the Green's law application. Shown are the tsunami forecast results for the hypothetical earthquakes of magnitude 8.0 (above) and 6.8 (below). On the abscissa the numbers of the 'forecast points' depicted in Fig. 9 are given. On the ordinate the maximum tsunami amplitudes in meter are given, as they result from the application of different modeling procedures (fine mesh as well as coarse mesh with and without applying Green's law; see legend). Vertical thick bars denote the range between maximum and minimum values of estimated tsunami amplitudes in each sub-section using an up to 150 m mesh. Vertical thin bars denote the range of directly estimated maximum and minimum tsunami amplitudes at the coast in each sub-section using a 1,350 m mesh

both events the estimated tsunami amplitudes at the coast by using a 1,350 m mesh and applying Green's Law to the forecast points (solid squares) agree well with the means (open squares) and medians (open circles) of tsunami amplitudes calculated by using a finer mesh in each coastal subsection. The coastal subsections are separated by middle points between the forecast points (see upper part of Fig. 9). Thus, it was confirmed that the application of Green's Law to tsunami amplitudes at the offshore forecast points yields amplitude estimates at the coast that are comparable with those calculated with much more time consuming fine mesh simulations. Therefore, we consider the former as representative estimates of the tsunami amplitude in the coastal subsection centered at the coastal projection of the forecast point.

It could also be shown that direct estimates of tsunami amplitudes at the coast based on coarse mesh simulations will underestimate values. This trend is even clearer for smaller event with shorter typical tsunami wavelengths (Fig. 10b). Therefore, it is not proper to estimate the tsunami amplitude directly at the coastal mesh when the mesh size is coarse relative to the typical tsunami wavelength.

However, there will be some issues to be considered in future in this method as follows.

- Strictly speaking, the Green's Law should only be applied to a direct wave. Therefore, in case the maximum amplitude is created by the superposition of multiply reflected waves, then this method can give overestimated tsunami amplitude.
- Scatter of estimated tsunami amplitude in one coastal subsection differs from section to section depending on the complexity of the coast line feature. It would be very difficult to precisely represent tsunami amplitudes for individual coastal points even if very fine bathymetry data are used in a simulation, because a short wavelength tsunami is significantly affected by small scale bathymetry and coastal feature or by a small change in an initial tsunami waveform. Therefore, incorporating a standard deviation of estimated tsunami amplitude to represent a degree of scatter in one coastal subsection, as well as a median or mean, in the tsunami warning/advisory grade determination would be effective for disaster management.



Tsunami Forecasting and Warning, Figure 10b  
(continued)

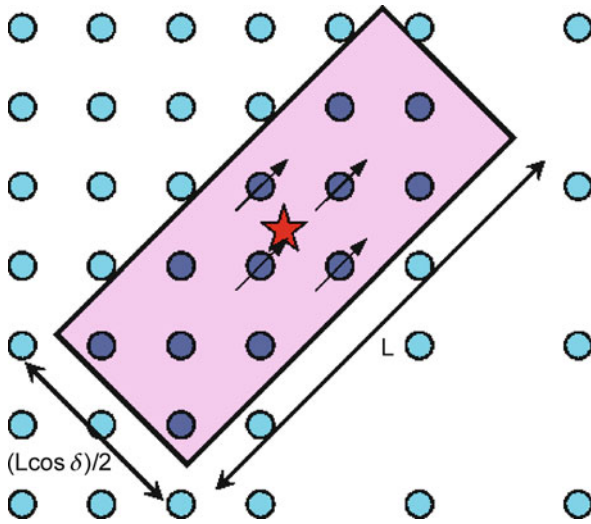
**Retrieval of the most Appropriate Case from the Database** In general, no identical case for actually determined latitude, longitude, hypocenter depth, magnitude exists in the database. Thus, some kind of selection rule is necessary to choose the most appropriate one.

**Hypocenter Depth and Magnitude** The logarithm of the estimated tsunami amplitude changes approximately positively linear with magnitude and negatively linear with hypocenter depth. Therefore, first, one has to calculate the logarithm of the estimated tsunami amplitudes for four cases (i. e., for two different depths and magnitudes each) which include the calculated depth and magnitude in between, then apply a two-dimensional linear interpolation and finally convert the log values back to a tsunami amplitudes.

**Latitude and Longitude** As for the location of the epicenter, either one of the two methods is used depending on the hypocentral area.

**Interpolation method** One selects the nearest four locations of simulation points in the data base which enclose the epicenter determined for the real event and performs a two-dimensional linear interpolation. This method is used far off the coast areas where horizontal shifts of the source location cause relatively smooth changes in the estimated tsunami amplitude at the coast. This is applicable in areas with a simulation point spacing of 1 degree (see Fig. 6).

**Maximum risk method** In the case that the assumed fault locations in the data base differ from the real location, tsunami amplitude estimates near the epicenter may differ substantially from observed ones for events occurring near to the coast. The uncertainty in the relative location of the epicenter on the surface projection of the seismic fault has to be considered at an early stage of the seismic data analysis, and all possible cases should be taken into account. Four cases with the epicenter located on one of the four corners of the fault are the most extreme cases, and the truth lies in between. Therefore, in the case of near coastal sources and with a view to disaster management, cases corresponding to every simulation point inside the rectangular area shown in Fig. 11 (solid circles) have to be considered. This rectangular area is centered at the epicenter (star mark). Its length and width are determined via the empirical formulas given in Subsect. "Fault Parameters Setting". Note, however, that the projected width in Fig. 11 is only  $0.5L \cos \delta$  and depends on the source dip. Also one has to remember that each simulation point is located in the center of the respective scenario fault in the data base. Therefore, the length and width of the rectangle to select all possible 'simulation points' are,  $L$  (not  $2L$ ) and  $0.5L \cos \delta$  respectively, and the strike of the length is that given to the simulation point (denoted by an arrow), which is closest to the determined epicenter. And the worst case is chosen amongst them for each of the forecast points.



Tsunami Forecasting and Warning, Figure 11  
**Maximum Risk Method:** The star denotes the epicenter and the circles the locations of simulation points. The rectangle delimitates the searching area from which the simulation points have to be retrieved. This rectangle is centered at the epicenter, and its length and width are  $L$  and  $0.5L \cos \delta$ , respectively, where  $L$  is given by the scaling law (see Subsect. "Fault Parameters Setting"). The strike of the length is the same as that given to the nearest simulation point to the epicenter. All simulation points inside of the rectangle (solid circles) are selected. Small arrows point into the strike direction assigned to simulation points

When readings of the tsunami arrivals at tidal gauges or tsunami meter become available, improbable simulation points can be excluded by applying inverse refraction diagrams. Thus the uncertainty in the spatial extent of the tsunami source area can be reduced.

If the number of the simulation points inside of the rectangular area is less than 4, the nearest 4 simulation points from the epicenter are retrieved.

**Tsunami Arrival Time Estimation** Forecast points are located not along the coast but offshore. Travel time estimation by adding the travel time from the source area to the forecast point to that from the forecast point to the coastal point may lead to substantial errors, especially when the tsunami source area is close to the coast. Therefore, for the tsunami arrival time estimation at the coastal point, an inverse refraction diagram from the coastal point to the source area is used to reduce such errors. Finite spatial extent of the source area, depending on the calculated magnitude value, is considered, and the earliest arrival time corresponding to the crossing point of the inverse refraction diagram and the outer rim of the source area is used with a view to disaster management.

**Tsunami Forecast Assembling** The tsunami forecast grade for a considered coastal block depends on the maximum of the expected tsunami amplitudes at the forecast points located in that block which are converted to amplitudes at the coast by applying Green's law. Forecast points are placed off-shore parallel to the coast with spacing of about 20 km. The average number of forecast points in one coastal block is 9. As mentioned above, expected tsunami amplitude at a forecast point, including Green's law application when using 1 arc minute bathymetry mesh, agrees reasonably well with expected mean or median tsunami amplitudes on the coast that have been calculated by using very fine bathymetry mesh in a coastal subsection separated in the middle between forecast points. This means that the expected tsunami amplitude at the forecast point can be regarded as being representative for the respective coastal subsection. We think this approach is reasonable with a view to disaster management requirements because the forecast grades for a coastal block are based on the maximum of these representative values in each coastal subsection. This prevents that the forecast grade is affected by abnormally large local maxima as they may re-

Tsunami Forecasting and Warning, Table 2a  
**Tsunami forecast/information examples of the JMA.** Original texts are in Japanese and these are English translations: **a** Tsunami forecast of the first issuance, **b** Tsunami information (estimated tsunami arrival time and amplitude) that follows a, **c** Tsunami information (estimated high tide and tsunami arrival times) that follows a, **d** Tsunami information (Observed results), **e** Tsunami forecast revision, **f** Tsunami forecast cancellation

Tsunami Warning
Issued by the Japan Meteorological Agency (JMA)
Issued at 2029JST 15 Nov 2006
"Tsunami Warning"
Eastern Part of Pacific Coast of Hokkaido
Okhotsk Sea Coast of Hokkaido
"Tsunami Advisory"
Central Part of Pacific Coast of Hokkaido
Western Part of Pacific Coast of Hokkaido
Northern Part of Japan Sea Coast of Hokkaido
Pacific Coast of Aomori Pref.
Iwate Pref.
Miyagi Pref.
Fukushima Pref.
Ibaraki Pref.
Kujukuri and Sotobo Area of Chiba Pref.
Uchibo Area of Chiba Pref.
Izu Islands
Sagami Bay and Miura Peninsula
Shizuoka Pref.

sult from very fine-mesh simulations (see Fig. 10). However, the JMA informs in its public information releases that tsunami amplitude might very locally be much higher than in the forecast.

The content of tsunami warning/advisory is very simple, namely, tsunami grades and corresponding coastal block names. This is to enable recipient organizations to understand the most important information easily, namely the necessity of evacuation. Tsunami information that follows just after the warning/advisory provides then more detailed estimates of tsunami amplitudes and arrival times for each coastal block, as well as hypocentral parameters. Examples of tsunami forecast/information are given in Table 2a–f.

**Tsunami Forecast Dissemination** Like other forecast/information disseminated from the JMA, tsunami

forecast/information is transmitted to relevant organizations by fully utilizing existing online facilities. In case of land line failure, forecast/information is also transmitted through satellite link. Loud speaker/sires operated by municipalities/prefectures are used in ‘the last mile’ to reach the public. Police and fire departments have their own robust transmission routes, which are also used. Additionally, the public broadcasting company plays a very important role because of its wide outreach and timeliness.

### Tsunami Monitoring System

Data from more than 100 tidal observation stations (including those operated by other institutes, like the Japan Coast Guard) are collected on-line at the JMA (Fig. 12). Data sampling and transmission rate are every one second.

Tsunami Forecasting and Warning, Table 2b  
(continued)

Tsunami Information (Estimated Tsunami Arrival Time and Amplitude) Issued by the Japan Meteorological Agency (JMA) Issued at 2030JST 15 Nov 2006		
— Estimated Tsunami Arrival Time and Amplitude —		
Coastal Block	Arrival Time	Amplitude
“Tsunami Warning”		
Eastern Part of Pacific Coast of Hokkaido	2110 15 Nov	1 m
Okhotsk Sea Coast of Hokkaido	2120 15 Nov	2 m
“Tsunami Advisory”		
Central Part of Pacific Coast of Hokkaido	2130 15 Nov	0.5 m
Western Part of Pacific Coast of Hokkaido	2150 15 Nov	0.5 m
Northern Part of Japan Sea Coast of Hokkaido	2250 15 Nov	0.5 m
Pacific Coast of Aomori Pref.	2140 15 Nov	0.5 m
Iwate Pref.	2140 15 Nov	0.5 m
Miyagi Pref.	2140 15 Nov	0.5 m
Fukushima Pref.	2210 15 Nov	0.5 m
Ibaraki Pref.	2210 15 Nov	0.5 m
Kujukuri and Sotobo Area of Chiba Pref.	2210 15 Nov	0.5 m
Uchibo Area of Chiba Pref.	2210 15 Nov	0.5 m
Izu Islands	2210 15 Nov	0.5 m
Sagami Bay and Miura Peninsula	2220 15 Nov	0.5 m
Shizuoka Pref.	2220 15 Nov	0.5 m
Tsunami may be higher than the estimation in some places.		
Sea level may slightly fluctuate in other coastal areas but no danger.		
— Earthquake Information —		
Origin Time: 2015JST 15 Nov 2006		
Epicenter: 46.6 North, 153.6 East		
Depth: 30 km		
Magnitude: 8.1		

Tsunami Forecasting and Warning, Table 2c  
(continued)

Tsunami Information (Estimated High Tide and Tsunami Arrival Time) Issued by the Japan Meteorological Agency (JMA) Issued at 2030JST 15 Nov 2006		
— Estimated High Tide and Tsunami Arrival Time —		
If tsunami arrives at coasts at around high tide time, tsunami becomes higher.		
Coastal Block	High Tide Time	Tsunami Arrival
"Tsunami Warning"		
Eastern Part of Pacific Coast of Hokkaido		2110 15 Nov
Kushiro	2335 15 Nov	2130 15 Nov
Hanasaki	2338 15 Nov	2120 15 Nov
Okhotsk Sea Coast of Hokkaido		2120 15 Nov
Abashiri	2226 15 Nov	2140 15 Nov
Monbetsu	2206 15 Nov	2200 15 Nov
Esashi-ko	2228 15 Nov	2220 15 Nov
"Tsunami Advisory"		
Central Part of Pacific Coast of Hokkaido		2130 15 Nov
Urakawa	2353 15 Nov	2140 15 Nov
Tokachi-ko	2359 15 Nov	2140 15 Nov
Western Part of Pacific Coast of Hokkaido		2150 15 Nov
Muroran	2337 15 Nov	2210 15 Nov
Hakodate	0008 16 Nov	2220 15 Nov
Tomakomai-nishi-ko	0007 16 Nov	2210 15 Nov
Yoshioka	0026 16 Nov	2230 15 Nov
Northern Part of Japan Sea Coast of Hokkaido		2250 15 Nov
Wakkanai	0229 16 Nov	2310 15 Nov
Rumoi	0114 16 Nov	2250 15 Nov
Otaru	0105 16 Nov	2250 15 Nov
Pacific Coast of Aomori Pref.		2140 15 Nov
Hachinohe	0000 16 Nov	2200 15 Nov
Sekinehama	0005 16 Nov	2200 15 Nov
Iwate Pref.		2140 15 Nov
Miyako	0006 16 Nov	2150 15 Nov
Ofunato	0017 16 Nov	2150 15 Nov
Kamaishi	0011 16 Nov	2150 15 Nov
(abbreviated)	(abbreviated)	(abbreviated)
Sea level may slightly fluctuate in other coastal areas but no danger.		
— Earthquake Information —		
Origin Time: 2015JST 15 Nov 2006		
Epicenter: 46.6 North, 153.6 East		
Depth: 30 km		
Magnitude: 8.1		

Arrival time, initial amplitude, polarity, maximum tsunami amplitude and corresponding time are measured on the man-machine interface depicted in Fig. 13, after prior removal of the ocean tidal component. Tsunami observation results are issued in tsunami information.

Sea level data are used for the following: 1) Deciding about the time of tsunami warning/advisory cancellation; 2) Revision of the grade of warning/advisory in the case that estimated and observed tsunami amplitudes differ significantly and 3) Informing the public on the up-to-date status of tsunami observation.

Tsunami Forecasting and Warning, Table 2d  
(continued)

Tsunami Information(Observation Results)				
Issued by the Japan Meteorological Agency (JMA)				
Issued at 2207JST 15 Nov 2006				
— Tsunami Observation at Sea Level Stations as of 2205 15 Nov —				
Higher tsunami may have arrived at some places other than the sea level stations.				
Tsunami may glow higher later.				
Kushiro	First Wave	2143 15 Nov	(+)	0.2 m
	Maximum Wave	2155 15 Nov		0.2 m
Hanasaki	First Wave	2129 15 Nov	(+)	0.4 m
	Maximum Wave	2143 15 Nov		0.4 m
Tokachi-ko	First Wave	2149 15 Nov		unclear
	Maximum Wave	(arriving)		
"Tsunami Warning" has been in effect for the following coastal blocks.				
Eastern Part of Pacific Coast of Hokkaido				
Okhotsk Sea Coast of Hokkaido				
"Tsunami Advisory" has been in effect for the following coastal blocks.				
Central Part of Pacific Coast of Hokkaido				
Western Part of Pacific Coast of Hokkaido				
Northern Part of Japan Sea Coast of Hokkaido				
Pacific Coast of Aomori Pref.				
Iwate Pref.				
Miyagi Pref.				
Fukushima Pref.				
Ibaraki Pref.				
Kujukuri and Sotobo Area of Chiba Pref.				
Uchibo Area of Chiba Pref.				
Izu Islands				
Sagami Bay and Miura Peninsula				
Shizuoka Pref.				
Sea level may slightly fluctuate in other coastal areas but no danger.				
— Earthquake Information —				
Origin Time: 2015JST 15 Nov 2006				
Epicenter: 46.6 North, 153.6 East				
Depth: 30 km				
Magnitude: 8.1				

### Tsunami Forecast for Distant Event and Northwest Pacific Tsunami Advisory (NWPTA)

For a distant event, a similar method to a local event as mentioned above is used, but the seismic waveform data from global network which are available in real-time through the Internet is used for hypocenter and magnitude calculation, instead of those from domestic seismic network. The PTWC's bulletins are also incorporated in the framework of ICG/PTWS. For magnitude estimation, Mwp [34] is used commonly with PTWC and

WC/ATWC. Tsunami simulation database has been created also for distant events, taking the Coriolis force effect into account considering long distance propagation. In case enough time is left until the earliest estimated tsunami arrival time at Japan coast, the JMA monitors sea level data collected via the data collection platform function of geostationary meteorological satellites, and circulated through GTS (global telecommunication network operated by the World Meteorological Organization of UN) circuit in near-real-time. After confirming the generation of tsunami at sea level observation sites near the

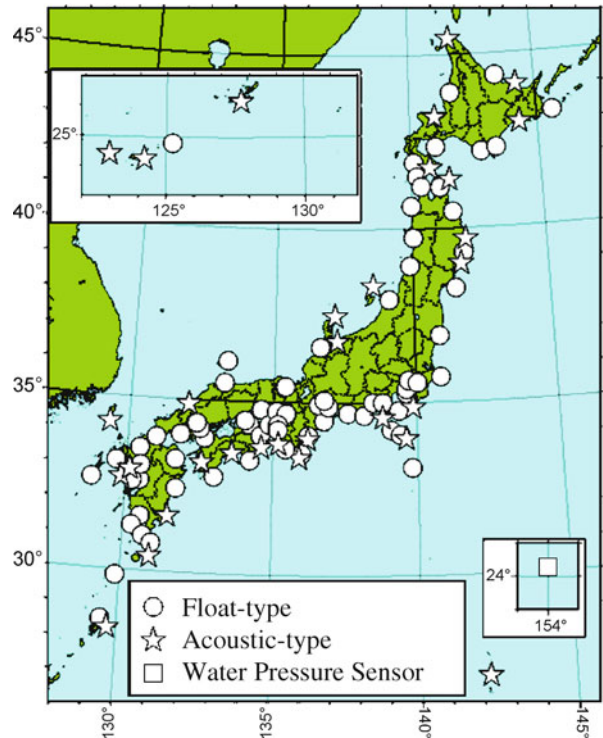


Tsunami Forecasting and Warning, Table 2e  
(continued)

<b>Tsunami Warning (Revision)</b>
Issued by the Japan Meteorological Agency (JMA)
Issued at 2330JST 15 Nov 2006
"Tsunami Warning" has been revised into "Tsunami Advisory" for the following coastal blocks.
Eastern Part of Pacific Coast of Hokkaido
Okhotsk Sea Coast of Hokkaido
"Tsunami Advisory" has been in effect for the following coastal blocks.
Eastern Part of Pacific Coast of Hokkaido
Okhotsk Sea Coast of Hokkaido
Central Part of Pacific Coast of Hokkaido
Western Part of Pacific Coast of Hokkaido
Northern Part of Japan Sea Coast of Hokkaido
Pacific Coast of Aomori Pref.
Iwate Pref.
Miyagi Pref.
Fukushima Pref.
Ibaraki Pref.
Kujukuri and Sotobo Area of Chiba Pref.
Uchibo Area of Chiba Pref.
Izu Islands
Sagami Bay and Miura Peninsula
Shizuoka Pref.
Ogasawara Islands

Tsunami Forecasting and Warning, Table 2f  
(continued)

<b>Cancellation of Tsunami Warning</b>
Issued by the Japan Meteorological Agency (JMA)
Issued at 0130JST 16 Nov 2006
Tsunami warning has been all canceled.
"Tsunami Advisory" has been canceled for the following coastal blocks.
Eastern Part of Pacific Coast of Hokkaido
Central Part of Pacific Coast of Hokkaido
Western Part of Pacific Coast of Hokkaido
Pacific Coast of Aomori Pref.
Iwate Pref.
Miyagi Pref.
Fukushima Pref.
Ibaraki Pref.
Kujukuri and Sotobo Area of Chiba Pref.
Uchibo Area of Chiba Pref.
Izu Islands
Sagami Bay and Miura Peninsula
Ogasawara Islands
However, sea level is expected to fluctuate in above coastal areas.
Be careful in sea bathing and fishing.

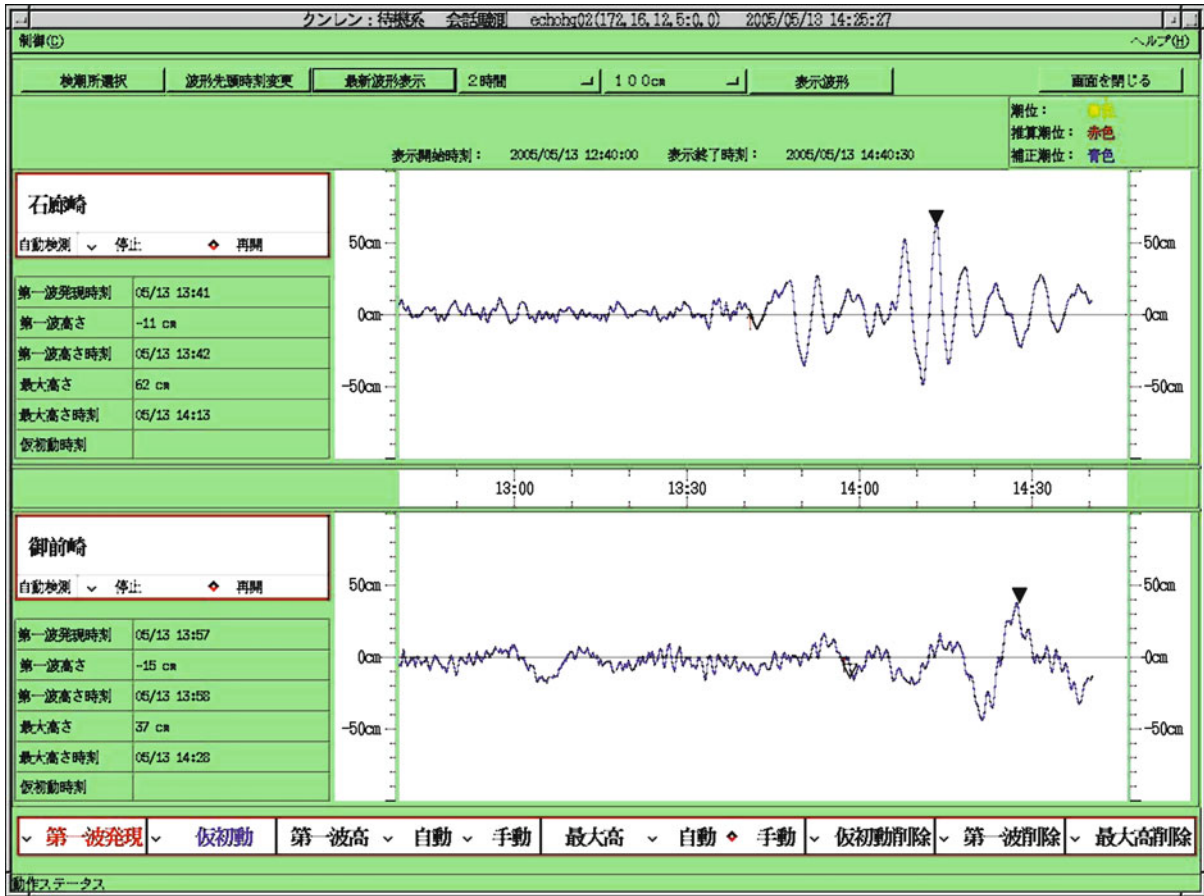
Tsunami Forecasting and Warning, Figure 12  
Tidal stations network used for tsunami monitoring by the JMA. More than 100 tidal stations are monitored in real time. Symbols denote instrumentation type. Circle, star and square denote float-type, acoustic-type and pressure sensor type, respectively

source, tsunami warnings/advisories are disseminated for the same coastal blocks as for local tsunami if necessary. Estimated tsunami amplitudes in the database can be calibrated by the actually observed tsunami amplitudes before the dissemination. Furthermore, if a reliable causing fault model is available, the JMA conducts tsunami propagation simulation for a specific setting of fault location, depth, magnitude and mechanism so that tsunami warnings/advisories can be based on more reliable tsunami estimation.

By using the same method as described above in this section, the JMA, as a regional center of the ICG/PTWS, is providing international tsunami information (NWPTA: Northwest Pacific Tsunami Advisory) to relevant countries when an earthquake with magnitude 6.5 or larger occurs in the northwest Pacific area since March 2005. NWPTA contains estimated tsunami amplitudes and arrival times, as well as estimated hypocenter parameters.

#### Some Lessons Learnt from a Recent Event

At 11:14 (UTC) on Nov. 15, 2006, a large earthquake with Mw8.3 occurred in the Kuril Island Region which gen-



Tsunami Forecasting and Warning, Figure 13

Example of the JMA's man-machine interface screen image in the Sea Level Data Monitoring System. Shown are two sea level recordings by different tidal gauge stations. The ocean tide component has been removed for more precise reading of the relevant tsunami wave parameters. Arrival time, initial wave amplitude, polarity, maximum amplitude and corresponding time are picked. The reversed solid triangles denote the maximum amplitude within the analyzed time window

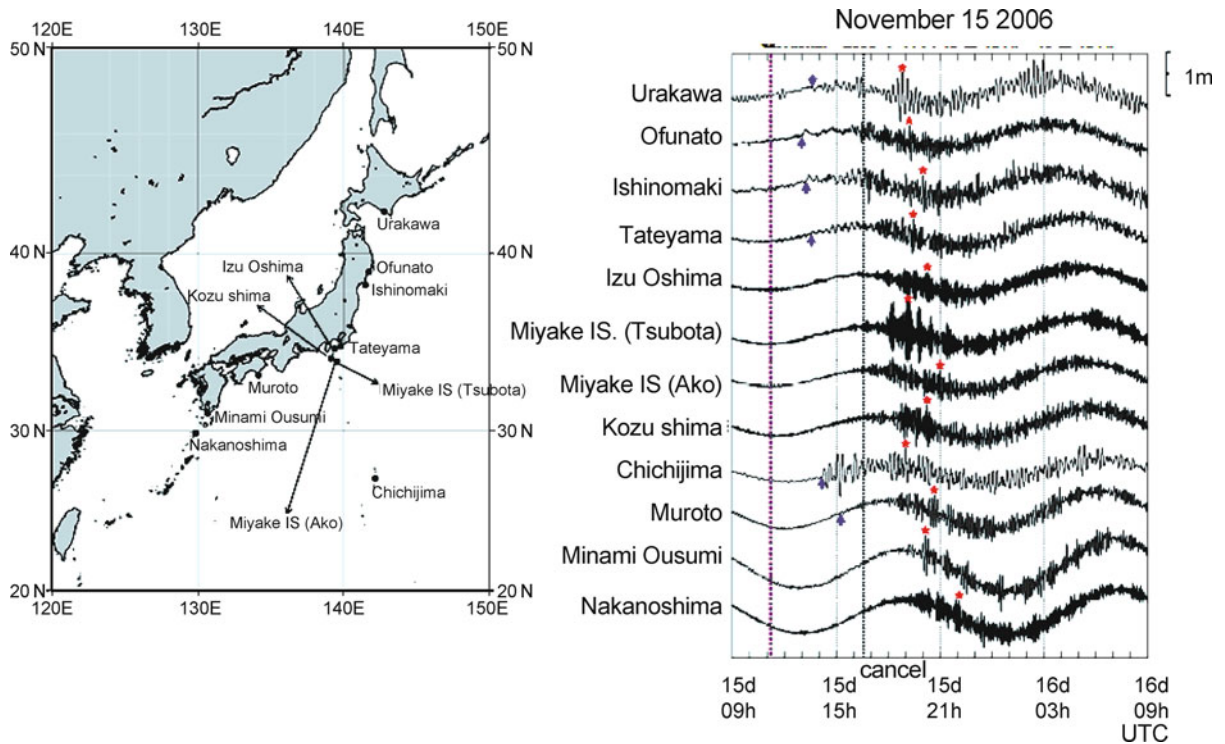
erated a Pacific-wide tsunami. Tsunami records at tidal gauges in Japan are shown in Fig. 14. The JMA disseminated tsunami forecasts at 11:29 (UTC) (corresponding to 20:29 JST in the related case example given in Table 2) based on the estimated tsunami amplitudes in the database. About 4 hours after the first detection of the tsunami on the Japan coast, the JMA cancelled all the forecasts at 16:30 (UTC) judging from the amplitude decay in the sea level record of Chichijima station (cf. Fig. 14) that no higher tsunami will be observed. But, after the cancellation, maximum amplitudes were recorded at many of the tidal gauges along the Pacific coast of Japan up to 5 hours later. The largest tsunami amplitude recorded in Japan for this event was 84 cm at Miyake Is. (Tsubota).

The JMA conducted a close examination of this case and found that these large amplitudes were tsunami waves reflected from the Emperor Sea Mount Chain in the mid-

dle of the Pacific (Fig. 15). Since the computational area for the tsunami simulation database of the JMA was limited around Japan (cf. Fig. 7), such reflected wave could not be represented by the simulation. Therefore, the JMA examines now the scenarios for which the computational area should be expanded.

### Recent Improvements

**Application of Earthquake Early Warning (EEW) Technique to Quicken Tsunami Forecast** In the near-coastal area (i. e., roughly within 100 km from the coast) EEW hypocenter and magnitude estimates [15] are equivalent to those resulting from the common procedures described above. Therefore, EEW results can be used as input data for tsunami forecast. The JMA started this incorporation in October 2006. In the case of the Mjma6.8 Niigataken



Tsunami Forecasting and Warning, Figure 14

Tidal station records of regular ocean tides superimposed by tsunami waves generated by the 15 Nov. 2006 Kuril earthquake ( $M_w = 8.3$ ). *Left*: Locations of tidal gauges whose data are shown in the right figure. *Right*: Tidal records of the stations shown in the left figure. Time is in UTC. The *thick blue dotted line* denotes the origin time of the earthquake and the *blue arrows* denote the arrival time of tsunami at each site. The *thin dotted line* denotes the time of tsunami forecast cancellation. *Red stars* mark the time of maximum amplitude at each site. At all sites maximum amplitudes were recorded about 2 to 5 hours after the forecast had been canceled. The largest tsunami amplitude recorded in Japan for this event was 84 cm at Miyake Is. (Tsubota)

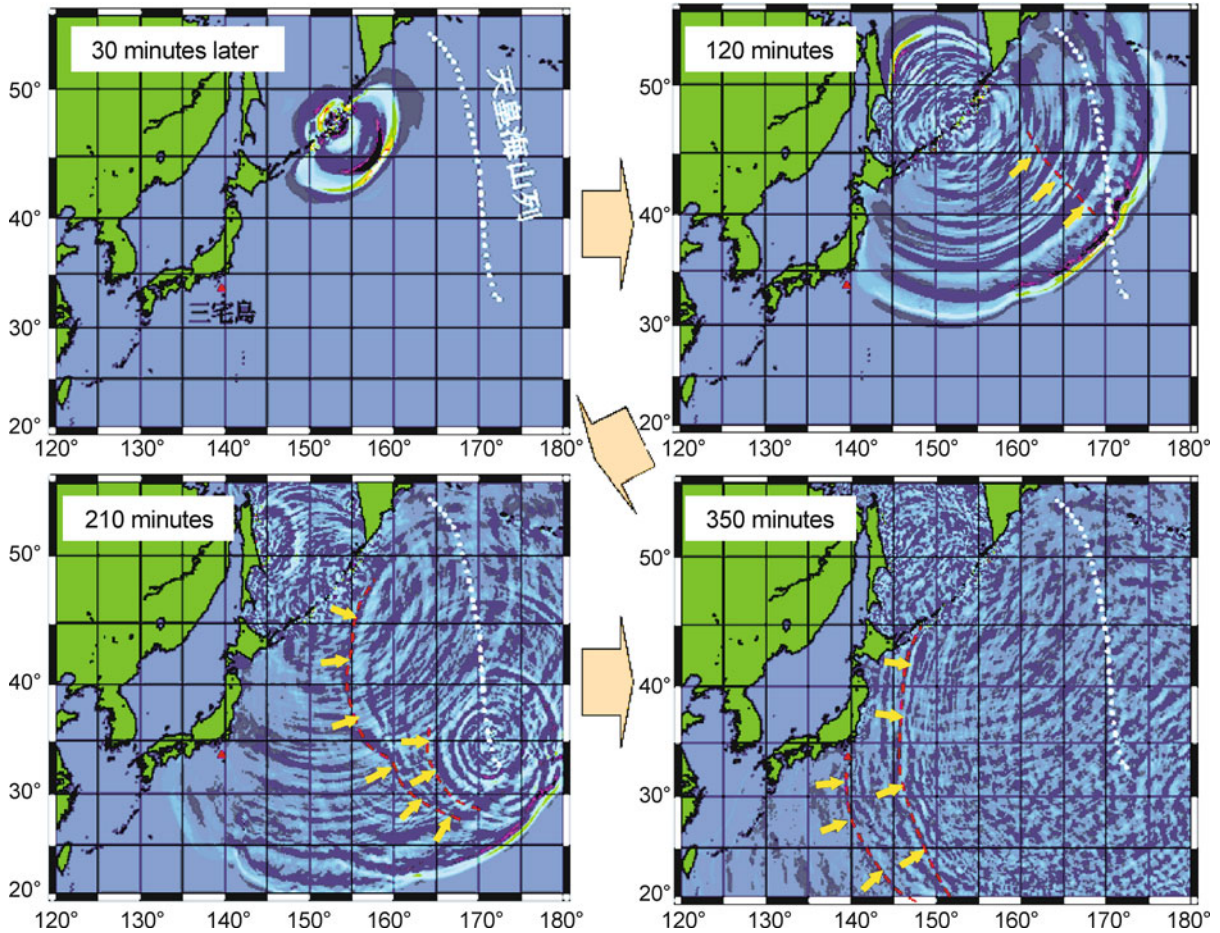
Chuetu-oki earthquake of 16 July 2007, the JMA disseminated tsunami advisory one minute after the first detection of the seismic wave.

**Quicker Revision/Cancellation of Tsunami Forecasts by Utilizing CMT Solutions** In JMA, automatic CMT solutions are available in 10 to 20 minutes after the earthquake occurrence. They yield more realistic magnitudes than Mjma for really great earthquakes ( $M_w > 8$ ) and reliable fault plane solutions. On the other hand, it is not proper to wait for the CMT solution for issuing the first tsunami warning, and it is not realistic neither to prepare the tsunami simulation database for so many different dip, strike and slip angle settings with fine increments and choose the best fitting one, because it requires huge storage and computational time for the conduct of simulations for all parameter settings. Therefore, we use the CMT solution for the revision/cancellation of the first tsunami warning, and we limit the number of different fault parameter set-

tings to 4, namely pure reverse fault with dip angles 10, 45 and 80 degrees and pure strike fault. Normal fault with the opposite slip angle ( $\lambda$ ) direction to the reverse fault with the same values for other fault parameters has almost identical tsunamigenic potential, and the difference is just a polarity of initial tsunami wave. Therefore, normal fault is treated as reverse fault with the same dip angle for this purpose. Strike angles have been fixed to the representative values of the considered regions, or set parallel to the nearby trench axis or coast line. The case with the highest resemblance to the actually calculated CMT solution with respect to its tsunamigenic potential is then retrieved.

The JMA incorporates CMT solutions in the revision/cancellation of tsunami forecast as follows:

- (a) In the case of reverse or normal faults, and if the centroid depth is less than the depth determined by  $P$  and  $S$  arrival times +30 km, and  $M_w$  0.5 or more larger than Mjma, the tsunami forecast is then upgraded in accordance with respective database cases.



Tsunami Forecasting and Warning, Figure 15

Snap shots of numerical tsunami simulation for the 2006 Nov. 15 Kuril event (Mw8.3). Coriolis force has been considered for this case. Four snap shots (30, 120, 210 and 350 minutes after the earthquake occurrence) are shown. The white dotted line marks the position of the Emperor Sea Mount Chain. Broken red curves with arrows denote reflected tsunami waves from the Emperor Sea Mount Chain. The estimated arrival times of the reflected waves at the coast of Japan are consistent with the actual tidal records

(b) In the case of strike faults, taking the database into account, and after confirmation by low or not observed tsunami amplitudes at tide gauges, the tsunami forecast is then downgraded or canceled.

The JMA started this incorporation in July, 2007 for the sea area where additional database creation has been completed.

### Future Outlook

In tsunami forecasting, trade-off exists between promptness and accuracy/reliance. The JMA's tsunami forecasting strategy, especially for local events, is to satisfy both by using state-of-the-art technologies. They assure promptness of the first issuance of tsunami forecast, based on pre-

liminary seismological results including that of EEW, and its subsequent revision, if required, as soon as more accurate and reliable data such as sea level change, and the results of a more detailed complex data analysis have become available. And as for accuracy/reliability there are two ways to be taken. One is quickened the process of reducing the uncertainty of initial tsunami wave distribution assessments. Quick focal process inversion technique, new magnitude definitions applicable for gigantic or tsunami earthquakes and tsunami source inversion techniques using sea level data (also from satellite altimetry) will become effective for this purpose soon. The other way is very detailed numerical simulation, after reduction of uncertainties in the initial tsunami wave distribution, using fine bathymetry mesh data. Along with the development of in-

egrated calculation algorithms, the improvement of the computer performance might solve this problem in future. But even then, due to the stochastic nature of tsunami behavior near the coast, one has carefully to examine on what statistical quantity of simulated results the tsunami forecast criterion has to be based.

### Acknowledgments

We thank Dr. Peter Bormann and Dr. Kenji Satake for reviewing the manuscript, their comments and suggestions greatly improved it.

### Bibliography

#### Primary Literature

- Abe K (1973) Tsunami and mechanism of great earthquakes. *Phys Earth Planet Inter* 7:143–153
- Bormann P, Wylegalla K (2005) Quick estimator of the size of great earthquakes. *Eos* 86(46):464
- Bormann P, Baumbach M, Bock G, Gresser H, Choy GL, Boatwright J (2002) Seismic sources and source parameters. In: Bormann P (ed) IASPEI new manual seismological observatory practice, vol 1, Chap 3. GeoForschungsZentrum Potsdam, Potsdam, pp 1–94
- Geller RJ (1976) Scaling relations for earthquake source parameters and magnitudes. *Bull Seism Soc Am* 66:1501–1523
- Geographical Survey Institute of Japan (2006) Real-time collection and analysis of crustal deformation data, Report on technical development and promotion plan concerning prompt disaster mitigation countermeasures based on disaster information, Chap 2. Ministry of Land, Infrastructure and Transport, Tokyo
- González FI, Bernard EN, Meinig C, Eble M, Mofjeld HO, Stalin S (2005) The NTHMP tsunameter network. *Nat Hazards (Special Issue, US National Tsunami Hazard Mitigation Program)* 35(1):25–39
- Jim Gower J (2005) Jason 1 detects the 26 December 2004 tsunami. *EOS Trans Am Geophys Union* 86(4):37–38
- Hara T (2007) Measurement of duration of high-frequency energy radiation and its application to determination of magnitudes of large shallow earthquakes. *Earth Planets Space* 59:227–231
- Hatori T (1984) On the damage to houses due to tsunamis. *Bull Earthq Res Inst* 59:433–439 (in Japanese)
- Hayashi Y (2008) Extracting the 2004 Indian Ocean tsunami signals from sea surface height data observed by satellite altimetry. *J Geophys Res* 113:C01001
- Ide S, Takeo M, Yoshida Y (1996) Source process of the 1995 Kobe earthquake: Determination of spatio-temporal slip distribution by Bayesian modeling. *Bull Seism Soc Am* 87:547–566
- Imamura F (1997) IUGG/IOC TIME PROJECT Numerical method of tsunami simulation with the leap-frog scheme, part 3 (Programme lists for near field tsunami), vol 35. IOC Manuals and Guides, Paris
- Japan Meteorological Agency (2005) A magnitude estimation using borehole volume strainmeters for earthquake events near the coast of Sumatra, Indonesia. *Rep Coord Comm Earthq Predict* 74:575–577 (in Japanese)
- Kajiura K (1970) Tsunami source, energy and the directivity of wave radiation. *Bull Earthq Res Inst (Univ. of Tokyo)* 48:835–869
- Kamigaichi O (2004) JMA earthquake early warning. *J Japan Assoc Earthq Eng (Special Issue)* 4:134–137
- Kanamori H, Anderson DL (1975) Theoretical basis of some empirical relations in seismology. *Bull Seism Soc Am* 65:1073–1095
- Kanamori H, Rivera L (2007) Speeding up seismic tsunami warning using *W* phase. In: Abstracts of AGU Fall Meeting 2007, S43C-06
- Kasahara M, Sasatani T (1986) Body wave analyses of strain seismograms observed at Erimo, Hokkaido, Japan. *J Fac Sc Hokkaido Univ Ser. VII (Geophysics)* 8:83–108
- Katsumata A (2004) Revision of the JMA displacement magnitude. *Q J Seismol* 67:1–10 (in Japanese)
- Kikuchi M, Kanamori H (1991) Inversion of complex body waves, III. *Bull Seism Soc Am* 81:2335–2350
- Lomax A, Michelini A, Piatanesi A (2007) An energy-duration procedure for rapid determination of earthquake magnitude and tsunamigenic potential. *Geophys J Int* 170:1195–1209
- Matsu'ura M, Hasegawa Y (1987) A maximum likelihood approach to nonlinear inversion under constraints. *Phys Earth Planet Inter* 47:179–187
- Okada M, Tanioka Y (1998) Relation of tsunami generation ratio with earthquake magnitude and hypocentral depth. *Mon Kaiyo (Special issue)* 15:18–22 (in Japanese)
- Okada Y (1985) Surface deformation due to shear and tensile faults in a half-space. *Bull Seism Soc Am* 75:1135–1154
- Ozawa S (1996) Geodetic inversion for the fault model of the 1994 Shikotan Earthquake. *Geophys Res Lett* 23(16):2009–2012
- Satake K (1989) Inversion of tsunami waveforms for the estimation of heterogeneous fault motion of large submarine earthquakes – the 1968 Tokachi-Oki and 1983 Japan Sea earthquakes. *J Geophys Res* 94:5627–5636
- Satake K (1995) Linear and nonlinear computations of the 1992 Nicaragua earthquake tsunami. *PAGEOPH* 144:455–470
- Shuto N (1991) Historical changes in characteristics of tsunami disasters. In: Proc of international symposium on natural disaster reduction and civil engineering. Japan Society of Civil Engineering, Tokyo, pp 77–86
- Shuto N (1992) Tsunami Intensity and damage, Tsunami engineering technical report. Tohoku Univ 9:101–136 (in Japanese)
- Shuto N (1998) Present state of tsunami research and defense works. *Bull Coastal Oceanogr* 35(2):147–157 (in Japanese)
- Shuto N et al (1986) A study of numerical techniques on the tsunami propagation and run-up. *Sci Tsunami Hazard* 4:111–124
- Takanami T, Kitagawa G (eds) (2002) Methods and application of signal processing in seismic network operations. Lecture Notes in Earth Science vol 98. Springer, Berlin
- Titov VV, González FI, Bernard EN, Eble MC, Mofjeld HO, Newman JC, Venturato AJ (2005) Real-time tsunami forecasting: Challenges and solutions. *Nat Hazards (Special Issue, US National Tsunami Hazard Mitigation Program)* 35(1):41–58
- Tsuboi S, Abe K, Takano K, Yamanaka Y (1995) Rapid determination of Mw from broadband *P* waveforms. *Bull Seism Soc Am* 83:606–613

35. Tsushima H, Hino R, Fujimoto H, Tanioka Y (2007) Application of cabled offshore ocean bottom tsunami gauge data for real-time tsunami forecasting. In: Proc symposium on underwater technology 2007/Workshop on scientific use of submarine cables and related technologies 2007. The University of Tokyo, Tokyo, pp 631–639
36. Ueno H, Hatakeyama S, Aketagawa T, Funasaki J, Hamada N (2002) Improvement of hypocenter determination procedures in the Japan Meteorological Agency. *Q J Seism* 65:123–134 (in Japanese)
37. Utsu T, Shima E, Yoshii T, Yamashina K (2001) *Encyclopedia of Earthquakes*, 2nd edn. Asakura, Tokyo, pp 657
38. Weinstein S, Okal E (2005) The mantle magnitude  $M_m$  and the slowness parameter  $\theta$ : Five years of real-time use in the context of tsunami warning. *Bull Seism Soc Am* 85: 779–799
39. Wells DL, Coppersmith KJ (1994) New empirical relationships among magnitude, rupture length, rupture width, rupture area, and surface displacement. *Bull Seism Soc Am* 84(4): 974–1002
40. Yokota T, Zhou S, Mizoue M, Nakamura I (1981) An automatic measurement of arrival time of seismic waves and its application to an on-line processing system. *Bull Earthq Res Inst* 56:449–484 (in Japanese)

#### **Books and Reviews**

- Bormann P (ed) (2002) *IASPEI new manual of seismological observatory practice*, vol 1 and 2. GeoForschungsZentrum Potsdam, Potsdam
- Satake K (2007) *Tsunamis*, chap 4, 17. *Treatise on geophysics*, vol 4. Elsevier, Amsterdam, pp 483–511

## Tsunami Inundation, Modeling of

PATRICK J. LYNETT

Texas A&M University, College Station, USA

### Article Outline

Glossary

Definition of the Subject

Introduction

Brief Review of Tsunami Generation  
and Open Ocean Propagation

Physics of Nearshore Tsunami Evolution

Effects of Bathymetric and Topographical Features  
on Inundation

Hydrodynamic Modeling of Tsunami Evolution

Moving Shoreline Algorithms

Future Directions

Bibliography

### Glossary

**Beach profile** A cross-shore, or normal to the beach, survey of the seafloor and dry ground elevation (bathymetry and topography); a series of spatial location and bottom elevation data pairs.

**Bore** A steep hydraulic front which transitions between areas of different water level. Tsunamis can approach land as a turbulent, breaking bore if the incident tsunami is of sufficiently large height.

**Boussinesq equations** An approximate equation model, used for waves with wave length of at least two times the local water depth; a long-wave-based model, but includes some frequency dispersion

**Dispersion, amplitude** The separation of wave components due to a wave-height related difference in wave speed; all else being equal, a wave with a large height will travel faster than one with a small height.

**Dispersion, frequency** The separation of wave components due to a frequency related difference in wave speed; all else being equal, a wave with a longer period will travel faster than one with a short period.

**Navier–Stokes equations** The full equations of fluid motion, including dissipation through the fluid molecular viscosity only. Other models discussed here, namely the Shallow Water Wave and Boussinesq equations, are approximations to these equations.

**Runup, or runup height** The ground elevation (a vertical measure) at the furthest point of landward inundation.

**Shallow water wave equations** An approximate equation model, used for waves with wave length many times

larger than the water depth; a non-dispersive, long-wave model; there is no frequency dispersion in this model.

**Tsunami inundation** The spatial area flooded as a tsunami rushes inland.

### Definition of the Subject

Tsunami inundation is the one of the final stages of tsunami evolution, when the wave encroaches upon and floods dry land. It is during this stage that a tsunami takes the vast majority of its victims. Depending on the properties of the tsunami (e. g. wave height and period) and the beach profile (e. g. beach slope, roughness), the tsunami may approach as a relatively calm, gradual rise of the ocean surface or as an extremely turbulent and powerful bore – a wall of white water. The characteristics of this approach determine the magnitude and type of damage to coastal infrastructure and, more importantly, the actions required of coastal residents to find a safe retreat or shelter.

To gage the nearshore impact of tsunami inundation, engineers and scientists rely primarily on three different methods: 1) Field survey of past events, 2) Physical experimentation in a laboratory, and 3) Numerical modeling. It is the last of these methods – numerical simulation of tsunami inundation – that will be the focus of this article. With numerical simulation, it is possible to predict the consequence of future tsunamis on existing coastal towns and cities. This information allows for the establishment of optimum evacuation routes, identification of high-risk and/or unprepared areas, re-assessment of building codes to withstand wave impact, and placement of tsunami shelters, for example. It is the hope that, through accurate prediction of tsunami effects, in conjunction with policy makers willing to implement recommended changes and a strong public education program, communities will show resiliency to tsunami impact, with minimal loss of life and damage to critical infrastructure.

### Introduction

On December 26, 2004, the boundary between the Indo–Australian and Eurasian plates off the coast of northern Sumatra ruptured in a great (Mw 9.3) earthquake at 00:58:53 universal time (U.T.). Up to 15m of thrust on the plate interface [31] displaced tens of cubic kilometers of seawater and propagated a tsunami across the Indian Ocean. The earthquake was widely felt throughout South Asia and was locally destructive in Sumatra and the Andaman and Nicobar islands, but it was the tsunami that caused widespread damage to densely populated coastal

communities both nearby and thousands of kilometers away.

Due to the extensive damage left behind by large tsunamis such as the Indian Ocean tsunami, it is difficult if not impossible to put together a complete picture of the event with field observations alone. Additionally, for some parts of the world that have not seen a tsunami in recent times, there are no field observations on which to develop safety procedures and protect residences from future tsunamis. It is for these purposes – understanding the detail of tsunami inundation and to estimate tsunami hazard – that we must rely on modeling of tsunamis. There are two primary modeling approaches - physical and numerical. The physical, or experimental, approach uses scaled-down models to look at a particular aspect of a phenomenon. While this approach is integral to the fundamental understanding of waves, because of the huge wavelengths of tsunamis, experiments are limited. For example, a tsunami might have a wavelength of 100 km in a deep ocean depth of 1 km, with a wave height of 1 m. Note that the above values represent approximate order of magnitudes for a large subduction zone tsunami. Now, to scale this down for the laboratory with a wave tank depth of 1 m – the tank would have to be 100 m long and the created lab-tsunami would have a hardly measurable height of 1 mm. Numerical modeling, while not “real” in the sense that modeling is done on a computer chip with approximated equations of motion rather than in the laboratory, does not suffer from this scaling problem, and can generally accommodate any type of arbitrary wave and ocean depth profile.

Numerical simulations of tsunami propagation have been greatly improved in the last 30 years. In the United States, several computational models are being used in the National Tsunami Hazard Mitigation Program, sponsored by the National Oceanic and Atmospheric Administration (NOAA), to produce tsunami inundation maps and predict tsunami runup in real-time for the warning system. In addition, there are numerous other models used by researchers and engineering companies in an attempt to better understand tsunami impact. In this article, an overview of these models, as well as how they are validated and utilized, is provided.

### **Brief Review of Tsunami Generation and Open Ocean Propagation**

Before introducing the physics behind propagating a tsunami across oceans and overland, we must first discuss how a tsunami is created. For earthquake generated mega-tsunamis, such as the Indian Ocean event, a huge

undersea earthquake along a great fault length of a subduction zone must occur. These earthquakes create large vertical motions of the seafloor. This vertical motion of the seafloor pushes the water above it, essentially creating a small displacement of water above the earthquake. This displacement of water will immediately try to spread out and reach a gravitational equilibrium, and it does so as waves propagating away from the earthquake zone – this is the tsunami.

To represent the tsunami in numerical models, we use an initial condition. Simply put, there is placed some irregular ocean surface profile at the instant after the earthquake, when the numerical simulations will start. Then, based on physics – Newton’s Laws written for fluid – the initial condition evolves and transits oceans. As a tsunami travels unhindered across ocean basins, it does so quickly and with little noticeable change. In the deep ocean, even the largest tsunamis have heights only near 1 m and currents of 10 cm/s, and are not likely to be identified by ships or surface buoys in the presence of wind waves.

### **Physics of Nearshore Tsunami Evolution**

A tsunami in the deep ocean is long and travels extremely fast. As the wave reaches shallow water, near the coastline, the tsunami begins the shoaling process. The speed at which long wave such as a tsunami moves, or celerity, is a function of the local water depth. The less the depth, the slower the wave moves. A tsunami, with its very long length, experiences different water depths at any given instant as it travels up a slope; the depth at the front of the wave, the portion of the tsunami closest to the shoreline, will generally be in the shallowest water and thus is moving the slowest. The back of the tsunami, on the other hand, will be in deeper water and will be moving faster than the front. This leads to a situation where the back part of the wave is moving faster than the front, causing the wavelength to shorten. With a shortening tsunami, the wave energy is in essence squeezed into a smaller region, forcing the height to grow. It is for this reason that, despite having a height of only a meter in the deep ocean, the tsunami elevation over land can easily exceed 10 m. With this great increase in wave height comes a more dynamic and complex phenomenon.

Presented in a more technical manner, a tsunami in the open ocean is generally a linear, non-dispersive wave. First, what is meant by a non-dispersive tsunami will be discussed. Also, the discussion here will be in terms of a large earthquake generation tsunami, such as the 2004 Indian Ocean event. Other impulsive waves, such as landslide or asteroid impact generated waves, are more difficult



to generalize and will be introduced separately at the end of this section.

Any wave condition, whether it is a tsunami or a typical wind wave in the ocean, can be mathematically described as a superposition, or summation, of a series of separate sine (or cosine) waves, each with independent amplitude and speed. For example, with the right choice of individual sine waves, it is possible to construct even the idealized tsunami: a single soliton. If a wave is considered a dispersive wave, then the various sine wave components will have different wave speeds, and the wave will disperse as the faster moving components move away from the slower ones. If a wave is non-dispersive, then all the components move at the same speed, and there is no lengthwise dispersal, or spreading, of the tsunami wave energy. It is for this reason that tsunamis can be devastating across such a large spatial region; the tsunami wave energy will not disperse but will remain in a focused pulse.

The dispersion described above is generally what scientists are referring to during a discussion of dispersive vs non-dispersive waves. However, it is more precisely called “frequency dispersion” as it is dominantly dependent on the period of the component. There is another type of dispersion, called “amplitude dispersion.” This second type of dispersion is a function of the nonlinearity of the wave, and is usually discussed under the framework of linear vs nonlinear waves. For tsunamis, the nonlinearity of the waves is given by the ratio of the tsunami height to the water depth. When this ratio is small, such as in the open ocean, the wave is linear; on the other hand, in shallow water the ratio is order unity and thus the wave is no longer linear. The linear/nonlinear nomenclature is not an intuitive physical description of the waves, but comes from the equations describing the tsunami motion, described later in this section. When this nonlinear effect is taken into account, it is found that the wave speed is no longer just a function of the local depth, but of the wave height as well. More specifically, looking at two components of the same period but with different amplitudes, the component with the larger amplitude will have a slightly larger wave speed. Except for the interesting cases of wave fission, discussed later in this section, the nonlinear effect of amplitude dispersion does not spread tsunami energy with an end result of lessening nearshore impact; in fact it will act to focus wave energy at the front, often leading to a powerful breaking bore.

Thus, open ocean propagation of a conventional tsunami is a relatively uncomplicated process which translates wave energy across basins, subject to wave speed changes that are a function of the local depth. As a tsunami enters the nearshore region, roughly characterized by wa-

ter depths of 100m and less, the wave can undergo a major physical transformation. The properties of this transformation depend heavily on the characteristics of the beach profile and the wave itself. In the simplest inundation case, the beach profile is relatively steep (footnote: here “steep” should be thought of in terms of the tsunami wavelength. If the horizontal distance along the slope connecting deep water to the shoreline is small compared to the tsunami wavelength, the beach would be considered steep) and the tsunami wave height is small, then the runup process closely resembles that of a wave hitting a vertical wall, and the runup height will be approximately twice the offshore tsunami height. In these special cases, a breaking bore front would not be expected; in fact horizontal fluid velocities near the shoreline would be very small. Here, the tsunami inundation would closely resemble that of a quickly rising tide with only very minor turbulent, dynamic impacts. However, even in these cases, overland flow constrictions and other features can create localized energetic inundation.

If the beach profile slope is mild, typical of continental margins, and/or the tsunami wave height is large, then the shallow water evolution process becomes highly nonlinear. However, while the nonlinear effect becomes very important, in the large majority of cases, frequency dispersion is still very small and can be neglected. Nearshore nonlinear evolution is characterized by a strong steepening, and possible breaking, of the wave front with associated large horizontal velocities. In these cases, turbulent dissipation can play a major role.

While it may be intuitive to postulate that wave breaking dissipation at the tsunami front plays a significant role in the tsunami inundation, this may not be altogether correct. This breaking dissipation, while extraordinarily intense, is fairly localized at the front which, both spatial and temporal, often represents only a small fraction of the tsunami. So, for tsunamis such as the 2004 Indian Ocean event, the related dissipation likely had only minor impact on leading-importance quantities such as the maximum runup and inland (off-beachfront) flow velocities. However, the properties of breaking are of great importance to other aspects of tsunami inundation. The maximum forces on beachfront infrastructure, such as ports, terminals, piers, boardwalks, and houses, should include the bore impact force as well as the drag force associated with the following quasi-steady flow [54,68]. If one was interested in understanding how bottom sediments are suspended, transported, and deposited by a tsunami, the bore turbulence again may play an important role. Thus, understanding the dynamics of a breaking tsunami front is not of particular importance for near real-time or operational

tsunami forecast models. This information is of great use for engineers and planners, who can utilize it to design tsunami-resistant structures, for example.

A second energy dissipation mechanism, one that does play a major role in determining maximum runup, is bottom friction. On a fundamental level, this dissipation is caused by the flow interaction with the bottom, where bottom irregularities lead to flow separations and the resulting turbulence. All natural bottoms result in some bottom friction; a smooth, sandy beach may generate only minor dissipation, while a coral reef or a mangrove forest can play a huge role in reducing tsunami energy [12]. Such features will be discussed in additional detail in the next section. Other means of energy dissipation will be largely local, and may include enhanced mixing due to sediment or debris entrainment, large shallow-flow vortex generation by headlands or other natural or artificial obstacles and the resulting dissipation, and flow through/around buildings and other infrastructure, sometimes termed macro-roughness and grouped with bottom friction.

Up to this point, we have only discussed the “typical” nearshore tsunami evolution which is portrayed as a wave without frequency dispersion, and may be called a linear or nonlinear tsunami, depending on a number of physical properties. The rest of this section will be devoted to those situations where the above characterization may no longer be adequate. Looking first to the tsunami source, waves that are generated by underwater landslides, underwater explosions, or asteroid impacts will often not behave as non-dispersive waves in the open ocean [44,72]. These source regions tend to be at least an order of magnitude smaller in spatial extent compared to their subduction zone counterparts. Physically, this implies that the generated waves will be of shorter wavelength. As a rule of thumb, if these generated waves have length scales of less than 10 times the local depth, then it should be anticipated that frequency dispersion will play a role [42]. Under this constraint, individual component wave speeds near the dominant period become frequency dependent.

Understanding that an impulsively generated wave can be dispersive has serious implications. Take, for example, a hypothetical landslide located in the Atlantic Ocean which generates a dispersive tsunami (e. g. Ward and Day, 2001). As this tsunami travels across the Atlantic, to either the USA east coast or the European west coast, frequency dispersion effects will spread the wave energy in the direction of propagation. This will convert the initial short-period pulse into a long train of waves. By spreading this energy out, the inundation impact will be greatly reduced. First, by taking a high-density energy pulse and stretching it into a longer, lower-density train, the max-

imum energy flux, and thus intensity, hitting the shoreline will decrease. Second, by increasing the duration of the time series, and creating many individual crests, energy dissipation can play a bigger role. Using simple energy arguments, it can be shown that, comparing a high-density, short-period pulse to a low-density, long-period train, more energy will be removed through bottom friction and breaking. This increase will be related to the ratio of the period of the entire dispersive wave train to the period of the pulse. Numerically studies have shown that for such cases, the individual wave crests are largely dissipated, and runup is dominated by the carrier wave, or in other terms it becomes a time-dependent, wave setup problem [26].

While a topic of current research in the tsunami community, frequency dispersion may occasionally play a non-negligible role in even the long wavelength, subduction zone tsunamis. To date, there have been two categories of argument that dispersion is important for these tsunamis: 1) short-period energy generated at the source is significant and leads to different patterns of runup if included (e. g. [21,28]), and 2) shallow-water nonlinear interactions can generate short-period components which can become decoupled (or un-locked) from the primary wave, and will change the incident tsunami properties [47].

Thinking of an arbitrary and complex initial free surface displacement generated by an undersea earthquake, there does exist the possibility that dispersive wave energy can be initially generated here. This irregular wave condition can be constructed as a continuous wave energy spectrum, and by definition there will be finite (albeit small) energy at all frequencies. The obvious question in this case is: what length scale characteristics of the initial free surface displacement, or the preceding earthquake, will lead to a significant measure of dispersive wave energy? To provide an answer to this question, a simple order-of-magnitude scaling argument is presented here; see Hammack and Segur [19], for example, for a mathematically rigorous attempt at insight. Let us define a characteristic change in vertical free surface elevation,  $\Delta\eta$ , and a horizontal length scale across which this vertical change occurs,  $\Delta L$ . Reducing to the simplest case, a regular wave with wave height equal to  $\Delta\eta$ , then the wavelength would be  $2\Delta L$ . Following this analogy, which will hold in a proportional sense for a Fourier series of wave components, for any  $\Delta\eta$  measured along a tsunami initial condition, there exists a wave component with wave length equal to  $2\Delta L$ . For that component to be significant to the tsunami evolution, the local vertical change,  $\Delta\eta$ , must be some non-negligible fraction of the maximum tsunami height,  $H$ . Note that the  $H$  dis-

cussed here is a global property of the entire initial tsunami wave condition. For the individual wave component, with length  $2\Delta L$ , to be dispersive, its length should be less than roughly 10 times the local water depth,  $h_0$ . Additionally, the difference between dispersive wave propagation and a non-dispersive propagation is cumulative. For example, if the full linear wave theory predicts, or a specific wave component, a wave speed that is 10% less than the long wave speed, then the predicted arrival time difference will grow by 0.1 times the period for each wavelength of propagation. Thus, the impact of dispersion is related to the distance of propagation, and is proportional to  $D/\lambda$ , where  $D$  is the total distance traveled by the wave, and  $\lambda$  is the average wavelength of the wave across  $D$  which can be expected to be proportional to  $2\Delta L$ . Assuming that  $\lambda$  is approximately equal to  $2\Delta L$ , it can be said that in order for frequency dispersion effects to play a role in tsunami evolution,

$$\max \left| \frac{h_0 \Delta \eta}{\Delta L^2} \right| \frac{D}{H} > \delta, \quad (1)$$

where  $\delta$  is some minimum threshold for importance. What this value should be is an open question, although it is likely to be near 0.1. From this term, it is clear the impact of dispersion is a function of a number of the properties of the initial tsunami condition, and should be taken into consideration when creating tsunami initial conditions. For example, use of discontinuous block-type segments (e. g. [30]), with sharp edges (very small  $\Delta L$ ) may lead to the conclusion that frequency dispersion is important, while it could be a direct result of a coarsely approximated initial tsunami condition. Also note that this exercise does not include the effects of radial spreading, which could very likely be important for small-scale irregularities in the initial condition. Wave height decrease by radial spreading is proportional to the horizontal curvature of the initial condition and to  $(\lambda/D)^n$ , and decreases faster for dispersive waves ( $n \sim 1$ ) as compared to non-dispersive waves ( $n = 0.5$ ) (e. g. [72]). Thus, for this case of significant radial spreading, it would be very difficult for source-based dispersion effects to play a meaningful role in the far field.

Under certain conditions, namely a nonlinear tsunami propagating across a wide shallow shelf, a process called fission may occur. Wave fission is a separation process where wave energy, initially part of a primary wave or pulse, attains certain properties, such as higher or lower phase speed, that allow it to disconnect from the primary wave and propagate as an independent wave. In the context of nearshore tsunami evolution, there is a standard mechanism which is the cause of this fission. First, it is nec-

essary to describe what a nonlinear, phase-locked wave is. To do this, we will examine the acceleration terms of the 1D conservation of momentum equation for the velocity component  $u$ :

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = -\frac{\partial p}{\partial x} + \mu \frac{\partial^2 u}{\partial x^2}. \quad (2)$$

Now assume that there is a single wave component, under which the velocity oscillates as

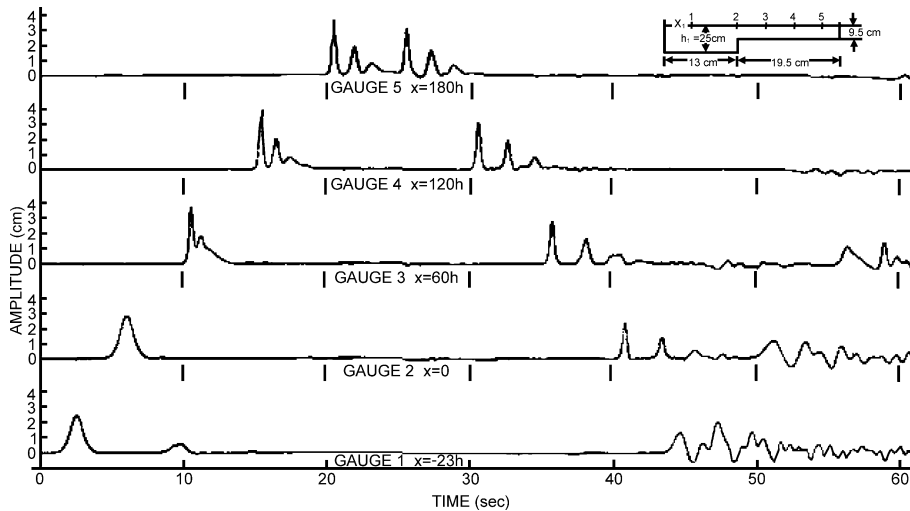
$$\cos(kx - \omega t) = \cos \theta, \quad (3)$$

where  $k$  is the wavenumber,  $\omega$  the frequency, and the speed of the wave is given by  $\omega/k$ . If the wave is nonlinear, which is to say that the convective acceleration term in the above momentum equation is not negligible, the convective term will include the product of

$$\cos \theta * \cos \theta = \cos 2\theta = \cos(2kx - 2\omega t). \quad (4)$$

Thus, through this nonlinear term, a new wave component, with twice the wavenumber and frequency (or half the wavelength and period) has been generated. From linear wave theory, it is expected that this new wave, with a shorter period, will have a different wave speed than the original, primary wave. However, from the phase function of this new wave, there is the speed  $2\omega/2k = \omega/k$ , which is the identical speed of the primary wave. Thus this new wave is *locked* to the *phase* of the original wave. This connection can be rather delicate, and any disruptions to the primary wave, such as a varying seafloor, dissipation, or interactions with other free waves in the train or wave pulse, can cause the new waves to become unlocked. When this occurs, the now free waves retain their frequency  $2\omega$ , but take a wavenumber as given by linear wave theory. Since these freed waves will be of a shorter period than the primary wave, they will travel at a slower speed and generally trail the main wave front.

Long wave fission is most commonly discussed in the literature via a solitary wave propagating over an abrupt change in depth, such as a step (e. g. [17,24,33,37,45,55]). In these cases, there is a deep water segment of the seafloor profile, where a solitary wave initially exists. In this depth, the solitary wave is of permanent form. As the solitary wave passes over the change in depth, into shallower water, the leading wave energy will try to re-discover a balance between nonlinearity and dispersion; the solitary wave. Since this new solitary wave will be a different shape and contain a lower level of mass, by conservation there must be some trailing disturbance to account for the deficient. This trailing disturbance will take the form of a rank-ordered train of solitons. Figure 1 depicts this process. The solitons in the trailing train, while smaller in height than



Tsunami Inundation, Modeling of, Figure 1

Example of experimental data looking at solitary wave fission by propagation onto a shelf from Goring and Raichlen [17]. Note that the flume layout and measurement location is given up in the upper right. The initial solitary wave undergoes the fission process and results in three distinct solitary waves

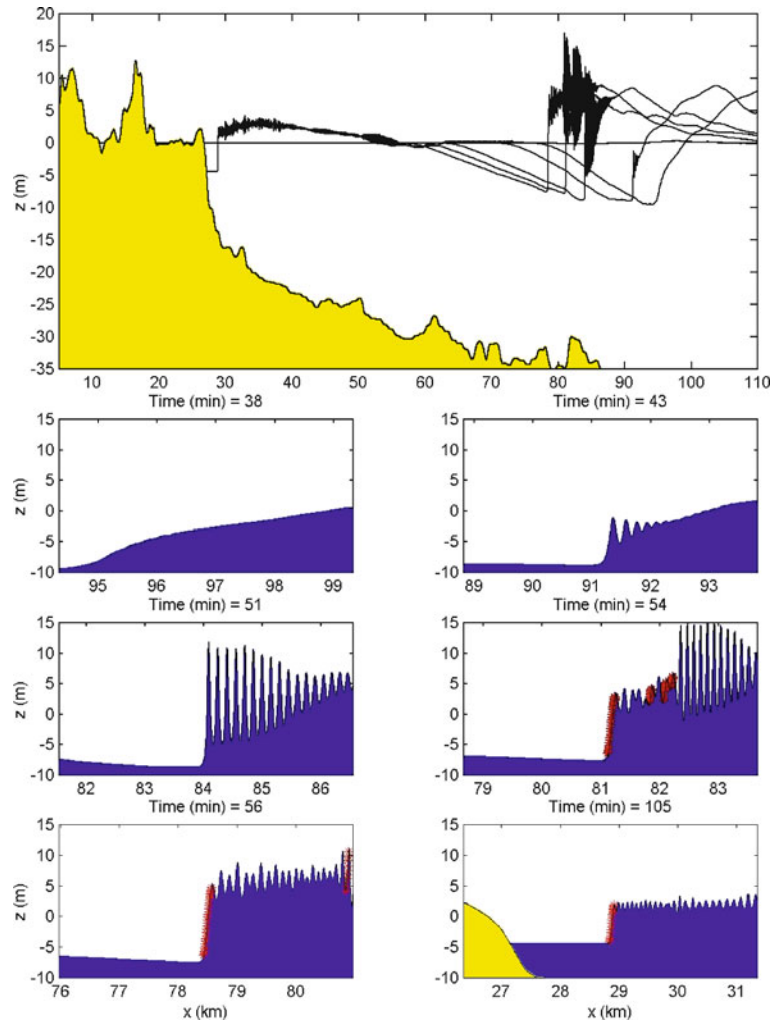
the leading solitary wave, tend to have a similar wavelength; this has been shown both analytically, numerically, and experimentally. Note, however, that discussion of fission in this sense is not particularly relevant to “real” tsunami modeling, where the offshore wave approaching the shelf break rarely resembles a solitary wave solution [62]. However, the offshore wave does not need to specifically be a solitary wave for this process to occur.

In numerous eyewitness accounts and videos recorded of the 2004 Indian Ocean tsunami, there is evidence of the tsunami approaching the coastline as a series of short period (on the order of 1 min and less) breaking fronts, or bores (e.g. [30]). These short period waves may be the result of fission processes of a steep tsunami front propagating across a wide shelf of shallow depth. Along the steep front of a very long period wave, nonlinearity will be very important. There will be a large amount of energy in high-frequency components with wavelengths similar the horizontal length of the tsunami front (on the order of 1 km). As the wave continues to shoal, the high-frequency locked waves may eventually become free waves, and will take the form of very short waves “riding” the main wave pulse. This situation is akin to an undular bore in a moving reference frame. This process is, in fact, identical to that described in the above paragraph, it simply takes place over a much longer distance. The newly freed waves, in the nonlinear and shallow environment, will attempt to reach an equilibrium state, where frequency dispersion and nonlinearity are balanced. Thus, the fission waves will appear

as solitary waves, or more generally, cnoidal waves. This fact provides some guidance as to the wavelength of these fission waves; they can be approximately calculated via solitary wave theory using the tsunami height and depth of the shelf. For example, on a shelf with depth of 30 m and an incident tsunami height of 5 m, fission waves with a wavelength of approximately 240 m and period of 13 s would be generated. In recent work looking at tsunamis along the eastern USA coast, where there exists a wide shallow shelf, this fission process has been investigated [14]. Figure 2 gives a few numerical simulation snapshots, and shows where the fission occurs, and the eventually impact on the waveform. This simulation, run with the dispersive equations, generated fission waves with lengths in the range of 100–200 m, and required a grid size of 5 m to attain numerically convergent results. In this example, the steep fission waves break offshore, and have little impact on the maximum runup. A conclusion of this fission issue is that, if one attempts to simulate tsunami propagation with dispersive equations, and if the grid is not chosen to be fine enough to resolve the short fission waves, the justification to use the dispersive model is greatly degraded.

### Effects of Bathymetric and Topographical Features on Inundation

It is well established that large-scale coastal features, such as small islands, large shoals, canyons, and shelves, can play an important role in tsunami inundation due to con-



Tsunami Inundation, Modeling of, Figure 2

Example of tsunami fission. Simulation results are from Geist et al. [14] for a landslide-generated tsunami off the east coast of the USA. The *top plot* shows the beach profile and six free surface profiles at different times. The *lower subplots* are zoom-in's of those six profiles, with the times given in the individual plot titles. The *red marks* visible in the *lowest plots* indicate regions where the wave is breaking

ventional shallow water effects such as shoaling and refraction (e. g. [4,5,16,36,70]). On the other hand, understanding of the impact of smaller scale features is just now being developed. This work was largely initiated by field observations. Synolakis et al. [60], surveying the coast of Nicaragua for information about the 1992 tsunami in the region, noted that the highest levels of damage along a particular stretch of beach were located directly landward of a reef opening used for boat traffic. It was postulated that the reef gap acted as a lower resistance conduit for tsunami energy, behaving like a funnel and focusing the tsunami. Along neighboring beaches with intact reefs, the

tsunami did not have the intensity to remove even beach umbrellas. Investigating impacts from the same tsunami, Borrero et al. [3], discussed how small scale bathymetry variations affected coastal inundation. One of the conclusions of this work was that bathymetry features with length scales 50 m and less had leading order impact on the runup. Looking to the recent Indian Ocean tsunami, a survey team in Sri Lanka inferred from observations that reef and dune breaks lead to locally increased tsunami impact [38,39]. Also in Sri Lanka, Fernando et al. [12] performed a more thorough survey along the southeastern coastline, and concluded that there was a compelling cor-

relation between coral mining and locally severe tsunami damage. While additional research is needed to quantify the effects of small scale features, the observations hint that defense measures such as seawalls, once thought to be inconsequential to tsunami inundation, may provide some protection.

Onshore, tsunami propagation is effected by the general topography (ground slope), ground roughness, and obstacles (e.g. [41,59,61,67]). The composition of the ground, be it sand, grass, mangroves, or pavement, controls the roughness and the subsequent bottom friction damping. To predict tsunami inundation with high confidence, the ground type must be well mapped and the hydrodynamic interaction with that type must be well understood. If the tsunami approaches the shoreline as a bore, the process of “bore collapse”, or the conversion of potential to kinetic energy, will cause the fluid to rapidly accelerate [56,69]. This fast flow equates to high fluid forces on obstacles such as buildings. Tsunami interaction with these obstacles can lead to a highly variable local flow pattern (e.g. [10,67]). As the flow accelerates around the

corners of a building, for example, the scour potential of that flow increases greatly, and foundation undermining is a concern (e.g. see Fig. 3). As with any fluid flow past an obstacle, the backface of the obstacle is characterized by a low-pressure wake. Combined with the interior flooding of a building, this low pressure wake may lead to an outward “pull” force on the back wall, causing it to fail by falling away from the center of the building. Such failures were observed during field surveys of the 2004 event, as shown in Fig. 4. Increasing the topographical complexity, in built coastal environments, structures are located within close enough proximity to each other such that their disturbances to the flow may interact. This can lead to irregular and unexpected loadings, where for example a 2nd row building experiences a larger force than beach front buildings due to a funneling effect. These types of interactions are very poorly understood, and require additional research.



Tsunami Inundation, Modeling of, Figure 3  
Example of foundation scour. This image was taken by the International Tsunami Survey Team to Sri Lanka in January 2005



Tsunami Inundation, Modeling of, Figure 4  
Example of damage to the backside of a coastal residence. These images were taken by the International Tsunami Survey Team to Sri Lanka in January 2005. The *top* photograph shows the front side of the structure, facing the ocean; there is damage but the main structure is intact. The *lower* photo shows the backside of the same building, showing the walls blown out, away from the center of the structure

## Hydrodynamic Modeling of Tsunami Evolution

Numerical simulations of tsunami propagation have made great progress in the last thirty years. Several tsunami computational models are currently used in the National Tsunami Hazard Mitigation Program, sponsored by the National Oceanic and Atmospheric Administration, to produce tsunami inundation and evacuation maps for the states of Alaska, California, Hawaii, Oregon, and Washington. The computational models include MOST (Method Of Splitting Tsunami), developed originally by researchers at the University of Southern California [66]; COMCOT (Cornell Multi-grid Coupled Tsunami Model), developed at Cornell University [35]; and TUNAMI-N2, developed at Tohoku University in Japan [29]. All three models solve the same depth-integrated and 2D horizontal (2DH) nonlinear shallow-water (NSW) equations with different finite-difference algorithms. There are a number of other tsunami models as well, including the finite element model ADCIRC (ADvanced CIRCulation Model For Oceanic, Coastal And Estuarine Waters; e.g., [53]). For a given source region condition, existing models can simulate propagation of a tsunami over a long distance with sufficient accuracy, provided that accurate bathymetry data exist.

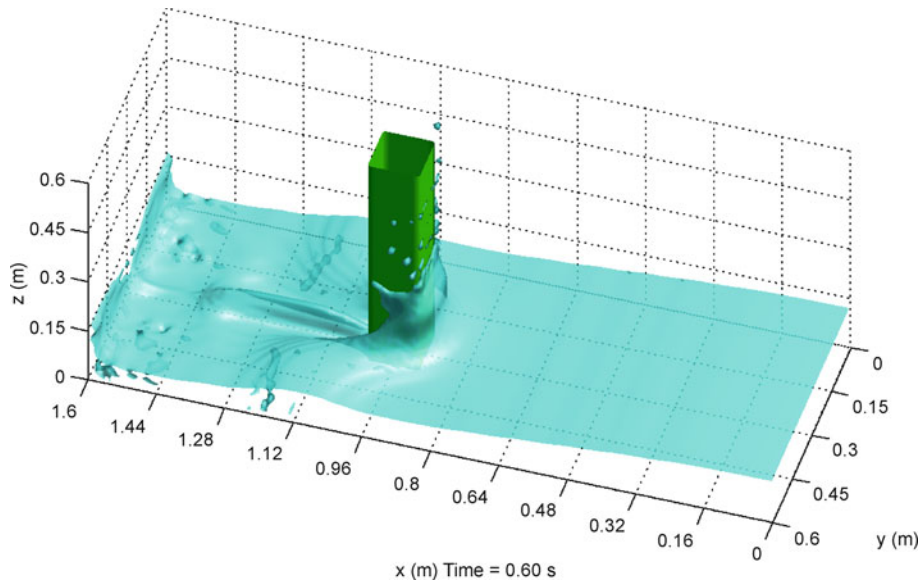
The shallow-water equation models commonly lack the capability of simulating dispersive waves, which, however, could well be the dominating features in landslide-generated tsunamis and for the fission processes described previously. Several high-order depth-integrated wave hydrodynamics models (Boussinesq models) are now available for simulating nonlinear and weakly dispersive waves, such as COULWAVE (Cornell University Long and Intermediate Wave Modeling Package; [42]) and FUNWAVE [25]. The major difference between the two is their treatment of moving shoreline boundaries. Lynett et al. [44] applied COULWAVE to the 1998 PNG tsunami with the landslide source; the results agreed with field survey data well. Recently, several finite element models have also been developed based on Boussinesq-type equations (e.g., [73]). Boussinesq models require higher spatial and temporal resolutions, and therefore are more computationally intensive. Moreover, most of model validation work was performed for open-ocean or open-coast problems. In other words, the models have not been carefully tested for wave propagation and oscillations in semi-enclosed regions – such as a harbor or bay – especially under resonant conditions.

Being depth-integrated and horizontally 2D, NSW and Boussinesq models lack the capability of simulating the details of many coastal effects, such as wave overturning

and the interaction between tsunamis and coastal structures, which could be either stationary or movable. At present, stationary coastal structures are parametrized as bottom roughness and contribute to frictional effects in these 2DH models. Although by adjusting the roughness and friction parameter satisfactory results can be achieved for maximum runup and delineation of the inundation zone (e.g., [35]), these models cannot provide adequate information for wave forces acting on coastal structures.

As a tsunami propagates into the nearshore region, the wave front undergoes a nonlinear transformation while it steepens through shoaling. It is in this nearshore region that dissipative effects can be important. Bottom friction can play a major role in the maximum runup and area of inundation (e.g. [67]). In depth-integrated models, bottom friction is typically approximated through a quadratic (drag) friction term, where the friction factor is calculated often through a Manning's coefficient or a Darcy–Wiesbach type friction factor (e.g. [25,35]). The validity of these steady-flow based coefficients has yet to be rigorously validated for use with tsunamis. If the tsunami is large enough, it can break at some offshore depth and approach land as a bore – the white wall of water commonly referenced by survivors of the Indian Ocean tsunami. Wave breaking in traditional NSW tsunami models has not been handled in a satisfactory manner. Numerical dissipation is commonly used to mimic breaking (e.g. [36]), and thus results become grid dependant. In Boussinesq models, this breaking is still handled in an approximate manner due to the fact that the depth-integrated derivation does not allow for an overturning wave; however these breaking schemes have been validated for a wide range of nearshore conditions (e.g. [40]).

Being depth-integrated, NSW and Boussinesq models lack the capability of simulating the vertical details of many coastal effects, such as strong wave breaking/overturning and the interaction between tsunamis and irregularly shaped coastal structures. To address this deficiency, several 2D and 3D computational models based on Navier–Stokes equations have been developed, with varying degrees of success. An example is COBRAS (Cornell Breaking waves and Structures model Lin and Liu 1998a,b, Lin et al. 1999), which is capable of describing the interactions between breaking waves and structures that are either surface piercing or submerged [6,22]. COBRAS adopted the Volume of Fluid (VOF) method to track free surface movement along with a Large Eddy Simulation (LES) turbulence closure model; several other computational models using different free surface tracking methods are also in use, such as the micro surface cell technique developed by Johnson et al. [23]. This 3D Navier–Stokes



Tsunami Inundation, Modeling of, Figure 5

An simulation snapshot taken from a 3D Navier–Stokes solver with an LES turbulence closure. This setup is looking at a bore impacting a column; the wake and vertical splash are clearly visible (Image provided by Philip L.-F. Liu, Cornell University)

equation model has been tested by two tsunami related experiments. The first is 3D landslide experiments [38,39], while the second involves measurements of solitary wave forces on vertical cylinders. Both experiments were conducted in the NEES tsunami basin at Oregon State. An example of a LES numerical solution of a solitary wave impinging on a circular is shown in Fig. 5.

Due to their high computational costs, full 3D models would best be used in conjunction with a depth-integrated 2DH model (i. e., NSW or Boussinesq). While the 2DH model provides incident far-field tsunami information, the 3D model computes local wave-structure interactions. The results from 3D models could also provide a better parametrization of small-scale features (3D), which could then be embedded in a large-scale 2DH model. One-way coupling (e. g. using a NSW-generated time series to drive a 3D model, but not permitting feedback from the 3D model back into the NSW) is fairly straightforward to construct (e. g. [18]). Two-way coupling, however, is difficult and requires consistent matching of physics and numerical schemes across model interfaces. Previous work in this area of two-way coupling of hydrodynamic models is limited. Fujima et al. [13] two-way coupled a NLSW model with a fully 3D model. While the results appear promising, the approach used by Fujima et al. requires ad-hoc and unphysical boundary conditions at the model matching locations, in the form of spatial gradients forced to zero, to ensure numerical stability. Even with these ad-hoc treat-

ments, their hybrid model compares very well with the completely-3D-domain simulation, requiring roughly 1/5 of the total 3D CPU time to achieve similar levels of accuracy. Sittanggang et al. [58] presented work on two-way coupling of a Boussinesq model and 2D Navier–Stokes model. These results indicate that there is large potential for hybrid modeling, in terms of more rapid simulation as well as the ability to approach a new class of problems.

### Moving Shoreline Algorithms

In order to simulate the flooding of dry land by a tsunami, a numerical model must be capable of allowing the shoreline to move in time. Here, the shoreline is defined as the spatial location where the solid bottom transitions from submerged to dry, and is a function of the two horizontal spatial coordinates and time. Numerical models generally require some type of special consideration and treatment to accurately include these moving boundaries; the logic and implementation behind this treatment is called a moving shoreline, or runup, algorithm.

For typical tsunami propagation models, it is possible to divide runup algorithms into two main approaches: those on a fixed grid and those on a Lagrangian or transformed domain. Both approaches have their advantages and disadvantages; currently fixed grid methods are found more commonly in operational-level models (e. g. [66]),



likely due in large part to their conceptual simplicity. A review of these two classes of models will be given in this section, followed by a review of the standard analytical, experimental, and field benchmarks used to validate the runup models. For additional information, the reader is directed to the comprehensive review given in Pedersen [49].

With a fixed grid method, the spatial locations of the numerical grid points or control volumes are determined at the start of a simulation, and do not change shape or location throughout the simulation duration. These methods can be classified into extrapolation, stairstep, auxiliary shoreline point, and permeable beach techniques. The extrapolation method has its roots in Sielecki and Wurtele [57], with extensions by Hibberd and Peregrine [20], Kowalik and Murty [27], and Lynett et al. [43]. The basic idea behind this method is that the shoreline location can be extrapolated using the nearest wet points, such that its position is not required to be locked onto a fixed grid point; it can move freely to any location. Theoretically, the extrapolation can be of any order; however, from stability constraints a linear extrapolation is generally found. Hidden in the extrapolation, the method is roughly equivalent to the use of low-order, diffusive directional differences taken from the last wet point into the fluid domain [43]. Additionally, there are no explicit conservation constraints or physical boundary conditions prescribed at the shoreline, indicating that large local errors may result if the flow in the extrapolated region cannot be approximately as linear in slope. The extrapolation approach can be found in both NLSW and Boussinesq models with finite difference, finite volume, and finite element solution schemes, and has shown to be accurate for a wide range of non-breaking, breaking, two horizontal dimension, and irregular topography problems (e. g. [8,9,26,44,51]).

Stairstep moving shoreline methods, one of the more common approaches found in tsunami models (e. g. [35]), reconstruct the naturally continuous beach profile into a series of constant elevation segments connected through vertical transitions. In essence, across a single cell width, the bottom elevation is taken as the average value. A cell transitions from a dry cell to a wet cell when the water elevation in a neighboring cell exceeds the bottom elevation, and transitions from wet to dry when the local total water depth falls below some small threshold value. These methods are particularly useful in finite volume and *C*-grid [1] type approaches (e. g. [32,36]), but can be difficult to implement in centered difference models, particularly high-order models or those sensitive to fluid discontinuities, where the “shock” of opening and closing entire cells can lead to numerical noise.

Auxiliary shoreline point methods require dynamic re-gridding very near the shoreline, such that the last wet point is always located immediately at the shoreline. Obviously, this method requires a numerical scheme that can readily accommodate non-uniform and changing node locations. There is some relation to the extrapolation methods discussed above; the moving shoreline point must be assigned some velocity, and it is extrapolated from the neighboring wet points. However, it is fundamentally different in that the shoreline point is explicitly included in the fluid domain. Thus, it would be expected that the governing conservation equations near the shoreline are more precisely satisfied here, although still dependent on the appropriateness of the extrapolation. One such method can be found in Titov and Synolakis [65], and has been successfully applied in NLSW equation models.

Another fixed grid treatment of moving boundary problems is employing a slot or permeable-seabed technique [63,64]. Conceptually, this method creates porous slots, or conduits, through the dry beach, such that there is always some fluid in a “dry” beach cell, although it may exist below the dry beach surface. These porous, “dry” nodes use a modified form of the NLSW; it is noted here that although in concept this approach is modeling a porous beach, it is not attempting to simulate the groundwater flow under a real, sandy beach, for example. The equations governing the “dry” domain contain a number of empirical parameters that are tuned to provide reasonable runup agreement with benchmark datasets. The advantage of this approach is that it allows the entire domain; including the fluid and “dry” nodes, to be determined via a somewhat consistent set of governing equations, without requiring a direct search routine to determine the shoreline location. The method has gained some popularity in wind wave models (e. g. [25,46]) when a highly accurate estimate of the shoreline location is not the highest priority. However, the approach has been used with some success in tsunami studies (e. g. [30]) despite the fact that the empirical coefficients that govern the model accuracy cannot be universally determined for a wide range of problems [7].

Alternative to fixed grid methods is the Lagrangian approach. Here, the fluid domain is discretized into particles, or columns of fluid in depth-integrated models, that are transported following the total fluid derivative. There are no fixed spatial grid locations; the columns move freely in space and time and thus these techniques require numerical flexibility, in terms of utilizing constantly changing space and time steps. The Lagrangian approach can be described as both the more physically consistent and mathematical elegant method of describing shoreline mo-

tion. The shoreline “particle” is included in the physical formulation just as any other point in the domain (i. e. no extrapolations are necessary), and thus the shoreline position accuracy will be compromised only by the overarching physical approximation (e. g. long wave approximation) and the numerical solution scheme (e. g. second-order time integration). The cost for this accuracy is a mathematical system that can be more difficult and tedious to solve numerically, typically requiring domain transformations, mappings, and/or re-griddings. Lagrangian methods have been used successfully in finite difference and finite element nonlinear shallow water (NLSW) and Boussinesq equation models (e. g., [2,15,48,50,52,74]).

### Future Directions

Towards a more robust simulation of tsunami inundation, there are two major issues which require additional fundamental investigations: dissipation mechanisms and interaction with infrastructure. Bottom friction, known to play an important role in inundation, needs to be re-examined starting from its basic formulation. Can a steady-flow based Mannings-type expression for bottom friction be used for tsunami? Does the unsteady nature of the tsunami flow make use of these existing formulations invalid? The answer to these questions may be different depending on what part of the wave is investigated (e. g. front). In addition, the hydrodynamic effect of common coastal vegetation, such as mangroves, needs to be quantified. There is current discussion of the use of such natural roughness as a tsunami defense (e. g. [11]); confidence cannot be put in such measures until it is understood how they behave. In addition to bottom friction, which exists at all locations and times under an inundating tsunami, wave breaking can increase the total energy dissipation. While breaking is generally confined to the leading front of a tsunami, the characteristics of this front are important for hydrodynamic loadings on beachfront structures, and may be significant to the net sediment and debris transport of a tsunami. Three-dimensional tsunami breaking is poorly understood and has received little attention.

Wave loadings and interactions with infrastructure are not well understood. To tackle this problem, tsunami hydrodynamic models need to be coupled with structural and geotechnical models. Ideally, these models should all be two-way coupled, such that the displacement of a structure, be it a single collapsed wall, will change the flow pattern, and scour underneath the foundation will change the structure stability. Additionally, impacts of flow-transported debris (e. g. cars) should be included in this frame-

work. If such a modeling capacity existed, engineering design of coastal structures could be undertaken in a very efficient manner.

### Bibliography

1. Arakawa A, Lamb VR (1977) Computational design of the basic dynamical processes of the UCLA general circulation model. In: Chang J (ed) *Methods in computational physics*. Academic Press, New York, pp 174–267
2. Birknes J, Pedersen G (2006) A particle finite element method applied to long wave run-up. *Int J Numer Methods Fluids* 52(3):237–261
3. Borrero JC, Bourgeois J, Harkins G, Synolakis CE (1997) How small-scale bathymetry affected coastal inundation in the 1992 Nicaraguan tsunami. Fall AGU Meeting, San Francisco
4. Briggs MJ, Synolakis CE, Harkins GS, Green D (1994) Laboratory experiments of tsunami runup on a circular island. *PAGEOPH* 144(3/4):569–593
5. Carrier GF (1966) Gravity waves on water of variable depth. *Fluid J Mech* 24:641–659
6. Chang K-A, Hsu T-J, Liu PL-F (2001) Vortex generation and evolution in water waves propagating over a submerged rectangular obstacle. Part I solitary waves. *Coast Eng* 44:13–36
7. Chen Q, Kirby JT, Dalrymple RA, Kennedy AB, Chawla A (2000) Boussinesq modeling of wave transformation, breaking, and runup: Part I 2D. *J Waterw Port Coast Ocean Eng* 126(1): 57–62
8. Cheung K, Phadke A, Wei Y, Rojas R, Douyere Y, Martino C, Houston S, Liu PL-F, Lynett P, Dodd N, Liao S, Nakazaki E (2003) Modeling of storm-induced coastal flooding for emergency management. *Ocean Eng* 30:1353–1386
9. Cienfuegos R, Barthelemy E, Bonneton P (2007) A fourth-order compact finite volume scheme for fully nonlinear and weakly dispersive Boussinesq-type equations. Part II: Boundary conditions and model validation. *Int J Numer Meth Fluids* 53(9):1423–1455
10. Cross RH (1967) Tsunami surge forces, ASCE JI Waterways & Harbors Division WW4:201–231
11. Danielsen F, Sørensen MK, Olwig MF, Selvam V et al (2005) The Asian tsunami: A protective role for coastal vegetation. *Science* 310:643
12. Fernando HJS, McCulley JL, Mendis SG, Perera K (2005) Coral poaching worsens tsunami destruction in Sri Lanka. *Eos Trans AGU* 86(33):301
13. Fujima K, Masamura K, Goto C (2002) Development of the 2d/3d hybrid model for tsunami numerical simulation. *Coastal Eng J* 44(4):373–397
14. Geist E, Lynett P, Chaytor J (2008) Hydrodynamic modeling of tsunamis from the currituck landslide, *Marine Geology* (in press)
15. Gopalakrishnan TC, Tung CC (1983) Numerical analysis of a moving boundary problem in coastal hydrodynamics. *Intl J Numer Meth Fluids* 3:179–200
16. González FI, Satake K, Boss EF, Mofjeld HO (1995) Edge wave and non-trapped modes of the 25 April 1992 Cape Mendocino tsunami. *Pure Appl Geophys PAGEOPH* 144(3–4): 409–426
17. Goring DG, Raichlen F (1992) Propagation of long waves onto shelf. *J Waterw Port Coast Ocean Eng* 118(1):43–61

18. Guignard S, Grilli ST, Marcer R, Rey V (1999) Computation of shoaling and breaking waves in nearshore areas by the coupling of BEM and VOF methods. Proc. 9th Offshore and Polar Engng. Conf., vol III, ISOPE99, Brest, France, pp 304–309
19. Hammack J, Segur H (1978) Modelling criteria for long water waves. *J Fluid Mech* 84(2):359–373
20. Hibberd S, Peregrine DH (1979) Surf and run-up on a beach. *J Fluid Mech* 95:323–345
21. Horrillo, Juan, Kowalik, Zygmunt, Shigihara, Yoshinori (2006) Wave dispersion study in the Indian Ocean–Tsunami of December 26, 2004. *Marine Geodesy* 29(3):149–166(18)
22. Hsu T-J, Sakakiyama T, Liu PL-F (2002) A numerical model for waves and turbulence flow in front of a composite breakwater. *Coast Eng* 46:25–50
23. Johnson DB, Raad PE, Chen S (1994) Simulation of impacts of fluid free surfaces with solid boundaries. *Int J Num Methods Fluids* 19:153–176
24. Johson RS (1972) Some numerical solutions of a variable-coefficient Korteweg–de Vries equation (with application to solitary wave development on a shelf). *J Fluid Mech* 54:81
25. Kennedy AB, Chen Q, Kirby JT, Dalrymple RA (2000) Boussinesq modeling of wave transformation, breaking, and runup. 1: 1D. *J Waterway Port Coastal Ocean Eng ASCE* 126:39–47
26. Korycansky DG, Lynett P (2007) Runup from impact tsunami. *Geophys J Int* 170:1076–1088
27. Kowalik Z, Murty TS (1993) Numerical simulation of two-dimensional tsunami runup. *Marine Geodesy* 16:87–100
28. Kulikov E (2005) Dispersion of the Sumatra tsunami waves in the Indian Ocean detected by satellite altimetry. Report from P.P. Shirshov Institute of Oceanology, Russian Academy of Sciences, Moscow
29. Imamura F (1995) Review of tsunami simulation with a finite difference method, long-wave runup models. *World Scientific*, Singapore, pp 25–42, [http://en.wikipedia.org/wiki/2004\\_Indian\\_Ocean\\_earthquake](http://en.wikipedia.org/wiki/2004_Indian_Ocean_earthquake)
30. Ioualalen M, Asavanant JA, Kaewbanjak N, Grilli ST, Kirby JT, Watts P (2007) Modeling of the 26th December 2004 Indian Ocean tsunami: Case study of impact in Thailand. *J Geophys Res* 112:C07024, doi:10.1029/2006JC003850
31. Lay T, Kanamori H, Ammon C, Nettles M, Ward S, Aster R, Beck S, Bilek S, Brudzinski M, Butler R, DeShon H, Ekström G, Satake K, Sipkin S (2005) The great Sumatra–Andaman earthquake of December 26, 2004. *Science* 308:1127–1133, doi:10.1126/science.1112250
32. LeVeque RJ, George DL (2004) High-resolution finite volume methods for the shallow water equations with bathymetry and dry states. In: Liu PL, Yeh H, Synolakis C (eds) *Advanced numerical models for simulating tsunami waves and runup*. vol 10 of *Advances in Coastal and Ocean Engineering*. World Scientific, Singapore
33. Losada MA, Vidal V, Medina R (1989) Experimental study of the evolution of a solitary wave at an abrupt junction. *J Geophys Res* 94:14557
34. Liu PL-F, Synolakis CE, Yeh H (1991) Report on the international workshop on long-wave run-up. *J Fluid Mech* 229:675–688
35. Liu PL-F, Cho Y-S, Yoon SB, Seo SN (1994) Numerical simulations of the 1960 Chilean tsunami propagation and inundation at Hilo, Hawaii. in: El-Sabh MI (ed) *Recent development in tsunami research*. Kluwer, Dordrecht, pp 99–115
36. Liu PL-F, Cho Y-S, Briggs MJ, Kanoglu U, Synolakis CE (1995) Runup of solitary waves on a circular island. *J Fluid Mech* 320:259–285
37. Liu PL-F, Cheng Y (2001) A numerical study of the evolution of a solitary wave over a shelf. *Phys Fluids* 13(6):1660–1666
38. Liu PL-F, Lynett P, Fernando H, Jaffe BE, Fritz H, Higman B, Morton R, Goff J, Synolakis C (2005) Observations by the International Tsunami Survey Team in Sri Lanka. *Science* 308:1595
39. Liu PL-F, Wu T-R, Raichlen F, Synolakis CE, Borrero JC (2005) Runup and rundown generated by three-dimensional sliding masses. *J Fluid Mech* 536:107–144
40. Lynett P (2006) Nearshore modeling using high-order Boussinesq equations. *J Waterway Port Coastal Ocean Eng (ASCE)* 132(5):348–357
41. Lynett P (2007) The effect of a shallow water obstruction on long wave runup and overland flow velocity. *J Waterway Port Coastal Ocean Eng (ASCE)* 133(6):455–462
42. Lynett P, Liu PL-F (2002) A numerical study of submarine landslide generated waves and runup. *Proc R Soc London A* 458:2885–2910
43. Lynett P, Wu T-R, Liu PL-F (2002) Modeling wave runup with depth-integrated equations. *Coast Eng* 46(2):89–107
44. Lynett P, Borrero J, Liu PL-F, Synolakis CE (2003) Field survey and numerical simulations: A review of the 1998 Papua New Guinea tsunami. *Pure Appl Geophys* 160:2119–2146
45. Madsen OS, Mei CC (1969) The transformation of a solitary wave over an uneven bottom. *J Fluid Mech* 39:781
46. Madsen PA, Sorensen OR, Schaffer HA (1997) Surf zone dynamics simulated by a Boussinesq-type model: Part I. Model description and cross-shore motion of regular waves. *Coast Eng* 32:255–287
47. Matsuyama M, Ikeno M, Sakakiyama T, Takeda T (2007) A study on tsunami wave fission in an undistorted experiment. *Pure Appl Geophys* 164:617–631
48. Özkan-Haller HT, Kirby JTA (1997) Fourier-Chebyshev collocation method for the shallow water equations including shoreline run-up. *Appl Ocean Res* 19:21–34
49. Pedersen G (2006) On long wave runup models. In: *Proceedings of the 3rd International Workshop on Long-Wave Runup Models*, in Catalina Island, California
50. Pedersen G, Gjevik B (1983) Runup of solitary waves. *J Fluid Mech* 142:283–299
51. Pedrozo-Acuna A, Simmonds DJ, Otta AK, Chadwick AJ (2006) On the cross-shore profile change of gravel beaches. *Coast Eng* 53(4):335–347
52. Petera J, Nassehi V (1996) A new two-dimensional finite element model for the shallow water equations using a Lagrangian framework constructed along fluid particle trajectories. *Int J Num Methods Eng* 39:4159–4182
53. Priest GR et al (1997) Cascadia subduction zone tsunamis: Hazard mapping at Yaquina Bay, Oregon. Final Technical Report to the National Earthquake Hazard Reduction Program DOGAMI Open File Report 0-97-34, p 143
54. Ramsden JD (1996) Forces on a vertical wall due to long waves, bores, and dry-bed surges. *J Waterway Port Coast Ocean Eng* 122(3):134–141
55. Seabra-Santos FJ, Renouard DP, Temperville AM (1987) Numerical and experimental study of the transformation of a solitary wave over a shelf or isolated obstacles. *J Fluid Mech* 176:117

56. Shen M, Meyer R (1963) Climb of a bore on a beach Part 3. Runup. *J Fluid Mech* 16(1):113–125
57. Sielecki A, Wurtele MG (1970) The numerical integration of the nonlinear shallow-water equations with sloping boundaries. *J Comput Phys* 6:219–236
58. Sittanggang K, Lynett P, Liu P (2006) Development of a Boussinesq-RANSVOF Hybrid Wave Model. in Proceedings of 30th ICCE, San Diego, pp 24–25
59. Synolakis CE (1987) The runup of solitary waves. *J Fluid Mech* 185:523–545
60. Synolakis CE, Imamura F, Tsuji Y, Matsutomi S, Tinti B, Cook B, Ushman M (1995) Damage, Conditions of East Java tsunami of 1994 analyzed. *EOS, Transactions, American Geophysical Union* 76(26):257 and 261–262
61. Tadepalli S, Synolakis CE (1994) The runup of N-Waves on sloping beaches, *Proc R Soc London A* 445:99–112
62. Tadepalli S, Synolakis CE (1996) Model for the leading waves of tsunamis. *Phys Rev Lett* 77:2141–2144
63. Tao J (1983) Computation of wave runup and wave breaking. Internal Report. Danish Hydraulics Institute, Denmark
64. Tao J (1984) Numerical modeling of wave runup and breaking on the beach. *Acta Oceanol Sin* 6(5):692–700 (in chinese)
65. Titov VV, Synolakis CE (1995) Modeling of breaking and non-breaking long wave evolution and runup using VTCS-2. *J Harbors Waterways Port Coast Ocean Eng* 121(6):308–316
66. Titov VV, Synolakis CE (1998) Numerical modeling of tidal wave runup. *J Waterway Port Coast Ocean Eng ASCE* 124(4):157–171
67. Tomita T, Honda K (2007) Tsunami estimation including effect of coastal structures and buildings by 3D model. *Coastal Structures '07, Venice*
68. Yeh H (2006) Maximum fluid forces in the tsunami runup zone. *J Waterway Port Coast Ocean Eng* 132(6):496–500
69. Yeh H, Ghazali A, Marton I (1989) Experimental study of bore runup. *J Fluid Mech* 206:563–578
70. Yeh H, Liu P, Briggs M, Synolakis C (1994) Propagation and amplification of tsunamis at coastal boundaries. *Nature* 372:353–355
71. Ward SN, Day S (2001) Cumbre Vieja Volcano – Potential collapse and tsunami at La Palma, Canary Islands. *Geophys Res Lett* 28(17):3397–3400
72. Weiss R, Wunnemann K, Bahlburg H (2006) Numerical modelling of generation, propagation, and run-up of tsunamis caused by ocean impacts: model strategy and technical solutions. *Geophys J Int* 67:77–88
73. Woo S-B, Liu PL-F (2004) A finite element model for modified Boussinesq equations. Part I: Model development. *J Waterway Port Coast Ocean Eng* 130(1):1–16
74. Zelt JA (1991) The run-up of nonbreaking and breaking solitary waves. *Coast Eng* 15(3):205–246

## Tsunamis, Inverse Problem of

KENJI SATAKE  
Earthquake Research Institute, University of Tokyo,  
Tokyo, Japan

### Article Outline

Glossary  
Definition of the Subject  
Introduction  
Tsunami Generation by Earthquakes  
Tsunami Propagation  
Tsunami Observations  
Estimation of Tsunami Source  
Estimation of Earthquake Fault Parameters  
Future Directions  
Bibliography

### Glossary

**Inverse problem** Unlike a forward problem which starts from a tsunami source then computes propagation in the ocean and predicts travel times and/or water heights on coasts, an inverse problem starts from tsunami observations to study the generation process. While forward modeling is useful for tsunami warning or hazard assessments, inverse modeling is a typical approach for geophysical problems.

**Shallow water (long) waves** In hydrodynamics, water waves can be treated as shallow water, or long, waves when the wavelength is much larger than the water depth. In such a case, the entire water mass from water bottom to surface moves horizontally and the wave propagation speed is given as a square root of the product of the gravitational acceleration and the water depth.

**The 2004 Indian Ocean tsunami** On December 26, 2004, a gigantic earthquake, the largest in the last half century in the world, occurred off the west coast of Sumatra Island, Indonesia. With the source extending more than 1,000 km through Nicobar and Andaman Islands, the earthquake generated tsunami which attacked the coasts of Indian Ocean and caused the worst tsunami disaster in history. The total casualties were about 230,000 in many countries as far away as Africa.

**Fault parameters** Earthquake source is modeled as a fault motion, which can be described by nine static parameters. Once these fault parameters are specified, the seafloor deformation due to faulting, or initial condi-

tion of tsunamis, can be calculated by using the elastic dislocation theory.

**Refraction and inverse refraction diagrams (travel time map)** Refraction diagram is a map showing isochrons or lines of equal tsunami travel times calculated from the source toward coasts. Inverse refraction diagram is a map showing arcs calculated backwards from observation points. The tsunami source can be estimated from the arcs corresponding to tsunami travel times.

### Definition of the Subject

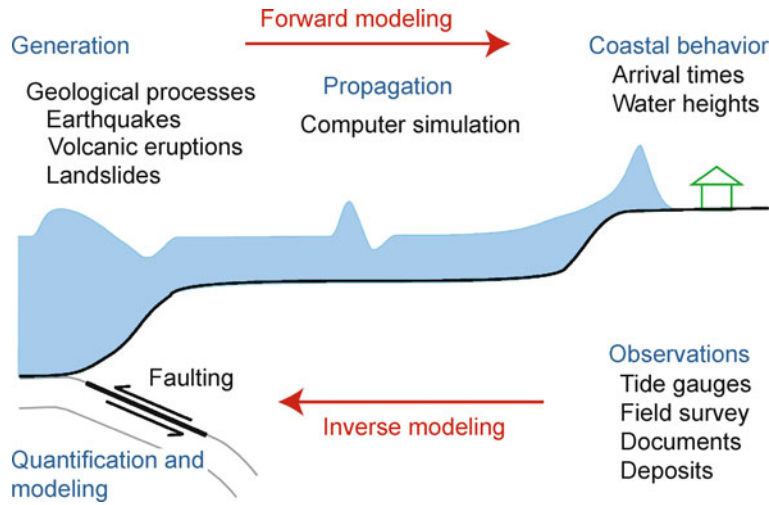
Forward modeling of tsunami starts from given initial condition, computes its propagation in the ocean, and calculates tsunami arrival times and/or water heights on coasts. Once the initial condition is provided, the propagation and coastal behavior can be numerically computed on actual bathymetry (Fig. 1).

Recent technological developments make it possible to carry out tsunami forward modeling with speed and accuracy usable for the early tsunami warning or detailed hazard assessments. However, the initial condition, or the tsunami generation process, is still poorly known, because large tsunamis are rare and the tsunami generation in the ocean is not directly observable. Indirect estimation of tsunami source, mostly on the basis of seismological analyzes, is used as the initial condition of tsunami forward modeling. More direct estimation of tsunami source is essential to better understand the tsunami generation process and to more accurately forecast the tsunami on coasts.

Inverse modeling of tsunami starts from observed tsunami data, to study the tsunami source. The propagation process can be evaluated by using numerical simulation, as in the forward modeling. As the observed tsunami data, tsunami arrival times, heights or waveforms recorded on instruments are used. For historical tsunamis, tsunami heights can be estimated from description of damage on historical documents. For prehistoric tsunamis, geological studies of tsunami deposits can be used to estimate the coastal tsunami heights or flooding areas.

### Introduction

Tsunamis are oceanic gravity waves generated by seafloor deformation due to submarine earthquakes or other submarine geological processes such as volcanic eruptions, landslides, or asteroid impacts. While earthquake tsunamis, such as the 2004 Indian Ocean tsunami caused by the Sumatra–Andaman earthquake, are most frequent, large volcanic eruptions such as the 1883 Krakatau eruption off Sumatra Island also cause ocean-wide tsunamis. Landslides, often triggered by earthquakes, cause locally



Tsunamis, Inverse Problem of, Figure 1

Schematic diagram showing tsunami generation, propagation and coastal behavior. Forward modeling starts from tsunami source and forecasts the coastal behavior, while inverse modeling starts from observed data to estimate the tsunami source

large tsunamis, but the effects are usually limited to the area around the source.

Most tsunamigenic geological processes produce seafloor deformation. When horizontal scale, or wavelength, of the seafloor deformation is much larger than the water depth, a similar disturbance appears on the water surface and becomes the source of tsunami. This is called shallow water, or long-wave, approximation. For large earthquakes, wavelength of seafloor deformation is an order of several tens to hundreds of km, while the ocean depth is up to several km, hence the long-wave approximation is valid. For small scale disturbance relative to water depth, such as submarine landslides or volcanic eruptions in deep seas, the shallow water approximation may not be valid.

This paper reviews inverse methods to study tsunami sources from the observations. Section “**Tsunami Generation by Earthquakes**” describes the tsunami generation by earthquakes, with emphasis on the fault parameters and their effects on tsunamis. Section “**Tsunami Propagation**” describes tsunami propagation: shallow water theory and numerical computation. Section “**Tsunami Observations**” summarizes the tsunami observation: instrumental sea-level data, runup height estimates for modern, historical and prehistoric tsunamis. Section “**Estimation of Tsunami Source**” describes methods of modeling and quantifying tsunami source, and of analyzing tsunami travel times, amplitudes and waveforms, including some historical developments. Section “**Estimation of Earthquake Fault Parameters**” focuses on the estimation of earthquake fault

parameters, including the waveform inversion of tsunami data to estimate heterogeneous fault motion and its application for tsunami warning.

### Tsunami Generation by Earthquakes

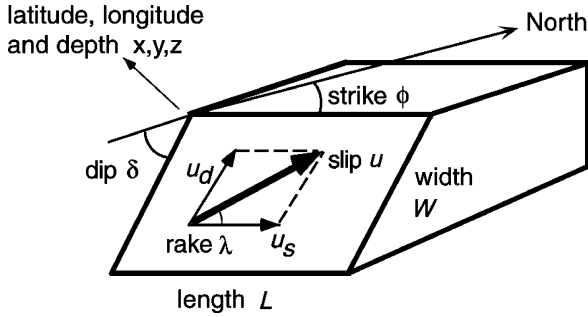
#### Fault Parameters and Seafloor Deformation

The seafloor deformation due to earthquake faulting can be calculated by using the elastic theory of dislocation. The displacement,  $u_k$ , in an infinite homogeneous medium due to dislocation  $\Delta u_i$  across surface  $\Sigma$  is given by the Volterra’s theorem as [1]

$$u_k = \frac{1}{8\pi\mu} \int_{\Sigma} \Delta u_i \{ \lambda \delta_{ij} u_k^{n,n} + \mu (u_k^{i,j} + u_k^{j,i}) \} v_j dS \quad (1)$$

where  $\lambda$  and  $\mu$  are Lamé constants,  $\delta_{ij}$  is Kronecker’s delta,  $v$  is the unit normal to the surface. The expression  $u_i^j$  denotes the  $i$ th component of the displacement due to the  $j$ th component of a point force at the source whose magnitude is  $8\pi\mu$ , and  $u_i^{j,k}$  indicates its spatial derivative with respect to the  $k$ th coordinate. For a half-space with free surface, a mirror image can be used to cancel the stress components on the free surface. The explicit formulas are given by Mansinha and Smyle [2] or Okada [3].

The fault parameters needed to compute surface deformation are summarized in Fig. 2. They are: fault length ( $L$ ), width ( $W$ ), strike ( $\phi$ ), dip ( $\delta$ ), rake ( $\lambda$ ), slip amount ( $u$ ) and location ( $x, y, z$ ). The slip  $u$  can be decomposed into strike-slip component  $u_s$  and dip-slip component  $u_d$ . The



Tsunamis, Inverse Problem of, Figure 2  
**Fault parameters.** Seafloor deformation can be computed from these static parameters

strike  $\phi$  is measured clockwise from North, dip angle  $\delta$  is downward from horizontal, and rake angle  $\lambda$  is a movement of hanging wall measured counter-clockwise from horizontal (see Fig. 2). Therefore, the fault motion is reverse if  $\lambda > 0^\circ$  and normal if  $\lambda < 0^\circ$ . The fault motion has left-lateral component if  $|\lambda| < 90^\circ$  and right-lateral component if  $|\lambda| > 90^\circ$ .

The physical parameter to quantify the fault motion is the seismic moment,  $M_0$ , defined as

$$M_0 = \mu u S = \mu u L W. \quad (2)$$

More conventional parameter of earthquake size is a magnitude scale, which has been determined from amplitudes of seismograms. To relate the seismic moment and magnitude scales, the moment magnitude scale,  $M_w$ , is defined as [4,5]

$$M_w = \frac{2}{3} \log M_0 - 10.7 \quad (3)$$

where  $M_0$  is given in dyne.cm ( $10^{-7}$  Nm).

Most of the above fault parameters can be estimated from seismic wave analysis. The location and depth of fault ( $x, y, z$ ) correspond to hypocenter, which is estimated from arrival times of seismic waves. The fault geometry ( $\phi, \delta, \lambda$ ) is estimated from the polarity distribution of body wave first motions or azimuthal distribution of surface wave amplitudes. The seismic moment is estimated from waveform modeling of seismic waves. The fault size,  $L$  and  $W$ , are more difficult to estimate; they are usually estimated from aftershock distribution or detailed waveform modeling of seismic body waves. The slip amount,  $u$ , is indirectly estimated, from seismic moment  $M_0$  by assuming  $\mu$  and fault size ( $L$  and  $W$ ). All such estimates assume that the faulting is planar and continuous, which most often is a simplification of real, more complex faulting.

The 2004 Sumatra–Andaman earthquake was the largest earthquake since the 1960 Chilean earthquake ( $M_w$  9.5) or 1964 Alaskan earthquake ( $M_w$  9.2). The seismic moment estimates range  $4\text{--}12 \times 10^{22}$  Nm, and the corresponding moment magnitude  $M_w$  ranges 9.0–9.3 from the seismological analyzes [6,7,8]. The aftershock area extended from off Sumatra through the Nicobar to Andaman Islands with the total fault length of 1,200 to 1,300 km [6]. For such a gigantic earthquake, multiple fault planes with different strike and slip amounts are needed to represent the fault motion, as shown later (Sect. “**Estimation of Earthquake Fault Parameters**”).

### Effect of Fault Parameters on Tsunami Generation

Among the above static fault parameters, the slip amount has the largest effect on the vertical seafloor deformation and the tsunami amplitude. The dip angle and fault depth are also important parameters to control tsunami amplitude [9,10]. Dynamic parameters such as rupture velocity are found to be insignificant for tsunami generation. However, for a gigantic earthquake such as the 2004 Sumatra–Andaman earthquake with the source length over 1,000 km, rupture propagation effect on tsunamis is not negligible [11].

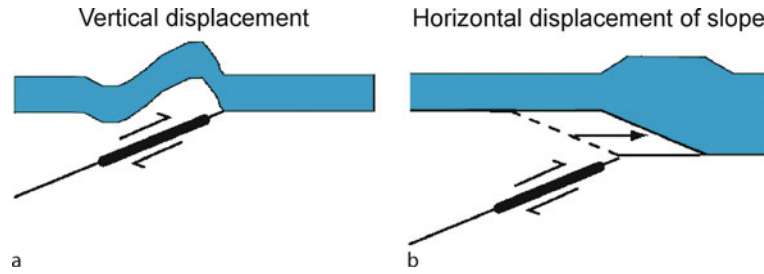
Amplitude of far-field seismic waves, either body waves or surface waves, is controlled by seismic moment, while amplitude of tsunami is controlled by fault slip. Satake and Tanioka [12] found for the 1998 Papua New Guinea tsunami that the far-field tsunami amplitudes are proportional to the volume of displaced water, while the near-field tsunami amplitudes are controlled by the potential energy of the displacement.

Traditionally, only vertical component of seafloor deformation has been considered for tsunami generation (Fig. 3a). If an earthquake occurs on a steep ocean slope such as trench slope, horizontal displacement due to faulting moves the slope and contributes the tsunami generation [13]. The effective vertical movement (positive upward) due to faulting can be written as follows (Fig. 3b),

$$u_z + u_x \frac{\partial H}{\partial x} + u_y \frac{\partial H}{\partial y} \quad (4)$$

where  $u_z$  is vertical component and  $u_x$  and  $u_y$  are horizontal components of seafloor deformation, and  $H$  is water depth (measured positive downward).

Recently, Song et al. [14] proposed that horizontal motion of slope also transfers kinetic energy from seafloor to water. They claimed that the kinetic energy thus transferred was about five times larger than the potential energy



Tsunamis, Inverse Problem of, Figure 3  
**Seafloor deformation and tsunami source. a Vertical seafloor deformation becomes the tsunami source. b When the seafloor is not flat, horizontal displacement of the slope also affects the tsunami generation [13]**

for the 2004 Indian Ocean tsunami, from comparisons of the observed tsunami (sea surface height) data with the computed ones by using an ocean-general-circulation-model.

**Tsunami Propagation**

**Shallow Water Theory**

The equation of motion, or conservation of momentum, for shallow water, or long-wave, theory is given as follows.

$$\frac{\partial \mathbf{V}}{\partial t} + (\mathbf{V} \cdot \nabla) \mathbf{V} = -g \nabla h + C_f \frac{\mathbf{V} |\mathbf{V}|}{d + h} \tag{5}$$

where  $\mathbf{V}$  is the depth-averaged horizontal velocity vector,  $h$  is the water height or tsunami amplitude,  $d$  is the water depth and  $g$  is the gravitational acceleration. On the left-hand side, the first term represents local acceleration and the second term represents nonlinear advection. On the right-hand side, the first term represents pressure gradient, or restoring force due to gravity, and the second term represents nonlinear bottom friction where  $C_f$  is the non-dimensional frictional coefficient.

The equation of continuity, or conservation of mass, can be written as

$$\frac{\partial (d + h)}{\partial t} = -\nabla \cdot \{(d + h) \mathbf{V}\} . \tag{6}$$

These equations can be linearized as

$$\frac{\partial \mathbf{V}}{\partial t} = -g \nabla h \tag{7}$$

$$\frac{\partial h}{\partial t} = -\nabla \cdot (d \mathbf{V}) \tag{8}$$

when the tsunami amplitude  $h$  is small compared to water depth  $d$ , and the bottom friction can be neglected. Such an assumption is valid for deep ocean, or most of the tsunami

propagation path. Near the coasts, nonlinear terms play important roles, hence linearization may not be valid. The major advantage of the linear theory is the superposition principle; the computational results can be easily scaled to estimate with different initial water heights.

From Eqs. (7) and (8), the wave equation with wave velocity (celerity)  $\sqrt{gd}$  can be derived. This indicates that the tsunami speed is controlled by water depth. Once the water depth distribution, or ocean bottom bathymetry, is known, then the tsunami propagation can be computed numerically.

**Numerical Computations**

Equations (5) and (6), or (7) and (8) for linearized case, can be directly solved by numerical methods, once the initial condition is given. The tsunami wave velocity distribution, which is given by the bathymetry, is much better known than the velocity distribution of seismic waves, hence actual values can be used. Finite-difference method with staggered grids is popularly used [15,16], while use of other methods such as finite-element methods have been also proposed [17]. Grid size for finite-difference computations is typically a few km for deep ocean, but grids as fine as several tens to hundreds of meters are used near coasts. The temporal changes in water height at grid points corresponding to the observation points are used as computed tsunami waveforms.

The database of global water depth or bathymetry data such as ETOPO2 (NOAA/NGDC) or GEBCO (British Oceanographic Data Centre) are popularly used. The ETOPO2 database is based on predicted bathymetry from satellite altimetry data [18] with interval of 2 minutes (about 3.5 km), while GEBCO data are digitized from nautical charts with grid interval of 1 minute. Higher resolution bathymetry data near coasts are open to public in some countries such as US (NOAA/NGDC) or Japan (JODC).



## Tsunami Observations

### Instrumental Data

Traditional instrumental data for tsunami observation are tide gauge records. Tide gauges are typically installed on ports or harbors to define datum or to monitor ocean tides. The temporal resolution is usually low with a sampling interval of several minutes or longer. For the tsunami monitoring, higher sampling rate, at least 1 min or shorter interval, is required. While the recorded tsunami waveforms contains coastal effects such as coastal reflections or resonance particularly for the later phase, the initial part of tsunami signals is more dominant by the tsunami source effect hence the source information can be retrieved. Currently, sea level measurement data from many tide gauge stations are transmitted through weather satellite and available in real time. Figure 4 (left) shows some of tide gauge records of the 2004 Indian Ocean tsunami.

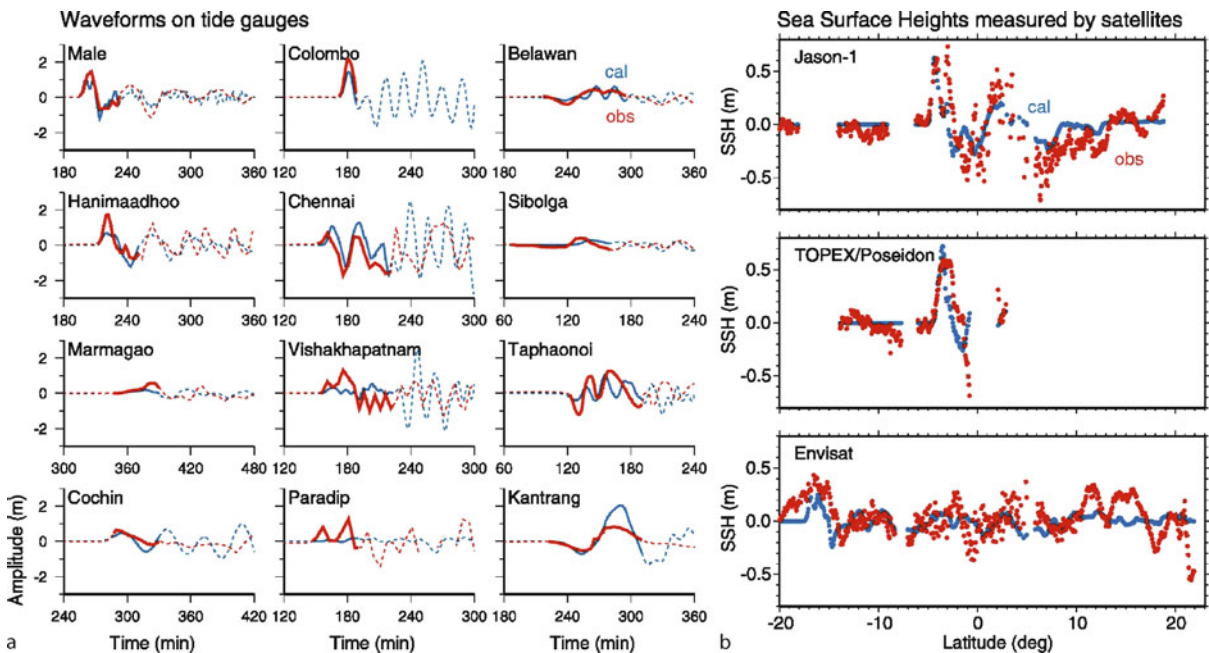
Tsunami waveforms are simpler offshore or in deep ocean, free from nonlinear coastal effects, though the signal is smaller. Offshore and deep ocean tsunami observation facilities have been significantly developed recently. Offshore tsunami gauges such as GPS tsunami gauge have been developed and deployed around Japan [19]. In 2003,

cabled bottom pressure gauges have recorded the tsunami generation process in the source area [20]. The US NOAA deployed dozens of bottom pressure gauges, called Deep-ocean Assessment and Reporting of Tsunamis (DART) or simply Tsunameters [21]. The bottom pressure signals are sent to surface buoys via acoustic telemetry in the ocean, then to land station via satellite. As described later, the DART data are used for real-time data assimilation and tsunami warning purposes [22]. After the 2004 Indian Ocean tsunami, many DART-type bottom pressure gauges have been deployed in the Pacific and Indian Oceans.

Satellite altimeters captured the propagation of the 2004 Indian Ocean tsunami (Fig. 4). Three satellites flew over the Indian Ocean at a few hours after the earthquake and measured the sea surface height (SSH) of about 0.8 m in the middle of Indian Ocean. The tsunami amplitudes in deep ocean are much smaller than the maximum coastal heights of more than 10 m. The SSH data are used to study the tsunami source [11,23].

### Modern, Historical and Prehistoric Tsunami Heights

After damaging tsunamis, tsunami height distribution is often measured by survey teams [24]. Measurements are usually made for flow depth above ground, on the basis of



Tsunamis, Inverse Problem of, Figure 4

The sea level data from the 2004 Indian Ocean tsunami [11]. **a** Tsunami waveforms on tide gauges. *Red curves* indicate observed waveforms and *blue ones* are computed. Data shown in *solid lines* are used for the waveform inversion. **b** Sea surface heights measured by three different satellites (see Fig. 5 for the tracks). *Red* shows the observed data and *blue* is for computed surface heights

various watermarks, then converted to inundation height above sea level [25]. The tsunami inundation heights are usually not constant along a profile from beach, and the height at most inland point is called runup height.

For historical tsunamis, coastal tsunami heights can be estimated from descriptions of tsunami or its damage recorded in historical documents. Such estimates include various assumptions on sea levels and a relationship between tsunami damage and flow depth, but provide important tsunami data for historical tsunamis. For example, date and size of the last gigantic earthquake in the Cascadia subduction zone off North America were estimated as January 26, 1700 and  $M_w \sim 9.0$  from the Japanese tsunami records [26].

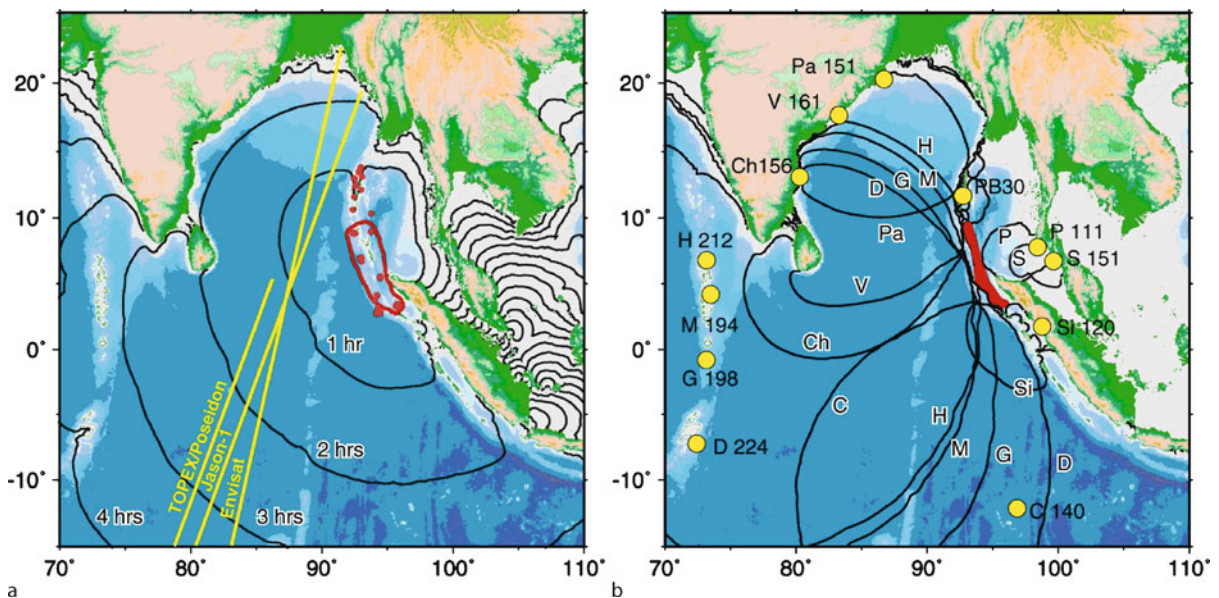
Geological traces such as tsunami deposits can also be used to estimate tsunami heights for prehistoric tsunamis. In the last few decades, many studies of tsunami deposits, combined with numerical computations, have been made to analyze prehistoric tsunamis [27,28]. For example, in Hokkaido, prehistoric tsunami deposits indicate past tsunamis with larger inundation area and longer recurrence interval than those from the recent plate-boundary earthquakes along the southern Kuril trench, which were attributed to multi-segment earthquakes with  $M_w \sim 8.4$  [29].

## Estimation of Tsunami Source

### Refraction Diagram

Tsunami propagation can be computed and described as a refraction diagram. When the tsunami wavelength is smaller than the scale length of velocity heterogeneity, or the water depth variation is smooth, then the geometrical ray theory of optics can be applied. The wavefronts of propagating tsunami can be drawn on the basis of Huygens' principle. Alternatively, propagation of rays, which is orthogonal to wavefronts, can be traced from an assumed source. While refraction diagram do not provide information on water height, the relative amplitudes can be estimated from density of rays [30].

Refraction diagrams can be prepared for major tsunami sources and used for tsunami warning; as soon as the epicenter is known, the tsunami arrival times can be readily calculated. The refraction diagrams are usually drawn from a point source, but it is possible to draw it from an extended source for a great or giant earthquake. Figure 5a shows the refraction diagram from the 2004 Sumatra–Andaman earthquake with wavefronts at each hour. To the east of the assumed source, the tsunami is expected to arrive at the Thai coast in about two hours through Andaman Sea. To the west, through deeper Bay



Tsunamis, Inverse Problem of, Figure 5

**a** Tsunami refraction diagram for the 2004 Sumatra–Andaman earthquake. Red dots indicate aftershocks within 1 day according to USGS. The red curve shows the assumed tsunami source. Tracks of three satellites with altimeters are shown by yellow lines. Black curves indicate tsunami wavefronts at each hour after the earthquake. **b** Tsunami inverse refraction diagram for the same event. Station code and tsunami arrival times (in min) are attached to tide gauge stations (yellow circles) where the tsunami was instrumentally recorded. Black curves are the travel-time arcs computed for each station. Red area indicates inferred tsunami source

of Bengal, the tsunami is expected to arrive at Sri Lanka also in two hours. The predicted tsunami arrival times are similar to the actually observed values [31].

### Inverse Refraction Diagram

Refraction diagram can be drawn backwards from coasts. Such a diagram is called inverse refraction diagram and is used to estimate the tsunami source area. When the tsunami travel time, that is tsunami arrival time minus earthquake origin time, is known, the corresponding wavefront, or travel-time arc, drawn from the tsunami observation point (typically tide gauge stations) would indicate the initial wavefront at the tsunami source. The tsunami inverse refraction diagram was first drawn for the 1933 Sanriku tsunami [32], although the estimated tsunami source was much larger than modern estimates, because both tsunami travel times and the bathymetry were poorly known.

The 2004 Indian Ocean tsunami was observed at many tide gauge stations in the Indian Ocean [33,34]. The tsunami arrival times were read from the tide gauge records and tsunami travel times were calculated from the earthquake origin time. The tsunami propagation was then computed from each tide gauge station, and wavefronts corresponding to the travel time were drawn as travel-time arcs (Fig. 5b). These travel-time arcs surround the tsunami source, and the source area was estimated as about 900 km long [6,35].

### Estimation of Tsunami Source

Tsunami data can be used to study earthquake source processes in a similar way that seismic waves are used. This was first demonstrated for the 1968 Tokachi-oki earthquake ( $M_0 = 2.8 \times 10^{21}$  Nm or  $M_w = 8.3$ ) [36]. The tsunami source area estimated from an inverse refraction diagram agrees well with the aftershock area (Fig. 6a, b). In addition, the initial water surface disturbance was estimated as uplift at the southeastern edge and subsidence at the northwestern edge, from the first motion of recorded tsunami waveforms on tide gauges. This pattern is very similar to the vertical bottom deformation due to the faulting, which was independently estimated from seismological analysis (Fig. 6c).

### Green's Law and Tsunami Heights

The water height in the tsunami source area can be estimated from the observed tsunami heights along the coasts, by using the Green's law. The Green's law is derived from the conservation of potential energy along rays [38],

$$b_0 d_0^{1/2} h_0^2 = b_1 d_1^{1/2} h_1^2 \quad (9)$$

where  $d$  is the water depth,  $b$  is the distance between the neighboring rays,  $h$  is the tsunami amplitude, and the subscripts 0 and 1 indicate two different locations on the same ray. If the tsunami amplitude at location 0 (e. g., on the coast) is known, the tsunami amplitude at location 1 (e. g., at the source) can be estimated as

$$h_1 = \left(\frac{b_0}{b_1}\right)^{1/2} \left(\frac{d_0}{d_1}\right)^{1/4} h_0. \quad (10)$$

The ratio  $b_0/b_1$  represents the spreading of rays, which can be graphically obtained from refraction diagrams. For the Tokachi-oki earthquake, the average tsunami height at the source was estimated as 1.8 m using the Green's law, which is very similar to 1.6 m, the average vertical seafloor displacement computed from the fault model [36].

The Green's law is also used to estimate the shoaling effects. For plane waves approaching the coast, the spreading ratio is unity, hence the amplitude is proportional to a 1/4 power of water depth change. For example, when the water depth becomes a half, the amplitude becomes 1.18 times larger.

### Tsunami Magnitude

Tsunami magnitude scale,  $M_t$ , was introduced to quantify earthquake source that generated tsunamis [39]. The formulas were calibrated with the moment magnitude scale,  $M_w$ , of earthquakes. It is different from other tsunami magnitude or intensity scales that simply quantify the observed tsunamis. The definition of  $M_t$  for a trans-Pacific tsunami is [39]:

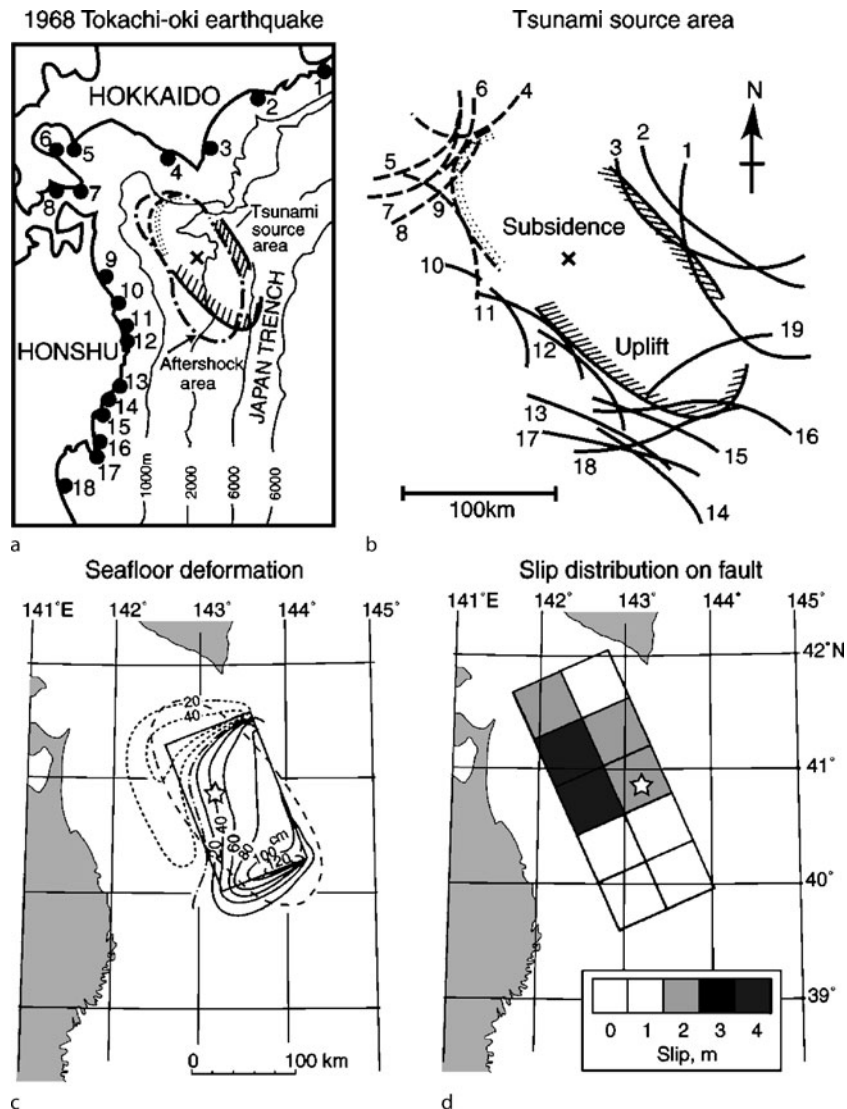
$$M_t = \log H + C + 9.1 \quad (11)$$

and for a regional ( $100 \text{ km} < \Delta < 3500 \text{ km}$ ) tsunami is [40]:

$$M_t = \log H + \log \Delta + 5.8 \quad (12)$$

where  $H$  is a maximum amplitude on tide gauges in meters,  $C$  is a distance factor depending on a combination of the source and the observation points, and  $\Delta$  is the nautical distance in km. The tsunami magnitude  $M_t$  was assigned as  $M_t = 8.2$  for the 1968 Tokachi-oki earthquake and  $M_t = 9.0$  for the 2004 Sumatra-Andaman earthquake.

Because the tsunami magnitude scale  $M_t$  is defined from tsunami amplitudes, it can be used to characterize "tsunami earthquakes" that produce much larger tsunamis than expected from seismic waves (see Polet and



Tsunamis, Inverse Problem of, Figure 6

The tsunami source area and seafloor deformation of the 1968 Tokachi-oki earthquake [36,37]. **a** Estimated tsunami source, aftershock area and distribution of tide gauge stations. **b** Travel-time arcs drawn by inverse refraction diagram. The numbers correspond to tide gauge stations in (a). Solid and dashed curves show uplift and subsidence, respectively. **c** Seafloor deformation computed from a seismological fault model. **d** Slip distribution on fault estimated by an inversion of tsunami waveforms

Kanamori: ► **Tsunami Earthquakes**). Abe [41] defined “tsunami earthquakes” for such events with tsunami magnitude  $M_t$  larger than surface wave magnitude  $M_s$  by more than 0.5. It should not be confused with “tsunamigenic earthquake” which refers to any earthquake that generates tsunami.

### Estimation of Earthquake Fault Parameters

For earthquake tsunamis, the fault parameters can be estimated by inverse modeling of tsunamis. Such attempts

were first made by a trial and error approach. In order to estimate the heterogeneous fault parameters, inversion of tsunami waveforms or runup heights has been introduced.

### Trial and Error Approach

Numerical simulation of tsunami has been carried out for many tsunamigenic earthquakes around Japan [42]. For the 1968 Tokachi-oki earthquake, tsunami waveforms were computed from two models, one based on

seismological analysis (Fig. 6c) and another horizontally shifted by 28 km, and were compared with the observed tsunami waveforms recorded on tide gauges. Comparison of waveforms indicates that the latter model, shifted from that based on seismological analysis, shows better match in terms of tsunami arrival times. The slip amount on the fault was estimated as 4 m.

The best fault models are judged by comparison of the observed and computed tsunami waveforms. A few statistical parameters used to quantify the comparison are geometric mean, logarithmic standard deviation and correlation coefficient. The geometric mean  $K$  of the amplitude ratio  $O_i/C_i$ , where  $O_i$  and  $C_i$  are the observed and computed tsunami amplitudes at station  $i$ , is given as

$$\log K = \frac{1}{n} \sum_i \log \frac{O_i}{C_i} \tag{13}$$

The logarithmic standard deviation  $\kappa$  is defined as

$$\log \kappa = \left[ \frac{1}{n} \sum_i \left( \log \frac{O_i}{C_i} \right)^2 - (\log K)^2 \right]^{1/2} \tag{14}$$

If the logarithmic amplitude ratios obey the normal distribution  $N(\log K, \log \kappa)$ , then parameter  $\kappa$  can be considered as an error factor, because its logarithm shows the standard deviation. The geometric mean  $K$  indicates the relative size of the observed and computed tsunami models. The logarithmic standard deviation  $\kappa$  indicates the goodness of the model; the smaller  $\kappa$  means the better model. The arrival times of observed and computed waveforms are also compared, as indicated in the above example. Another parameter is correlation coefficient of the observed and computed waveforms, which are also used for the comparison of models.

While the above parameters ( $K$  and  $\kappa$ ) were originally defined for maximum amplitudes of waveforms, they are also used for comparison of observed and computed runup heights. For tsunami hazard evaluation of nuclear power plants in Japan, tsunami source models need to satisfy  $0.95 < K < 1.05$  and  $\kappa < 1.45$  for the observed and computed coastal heights [43].

**Heterogeneous Fault Motion**

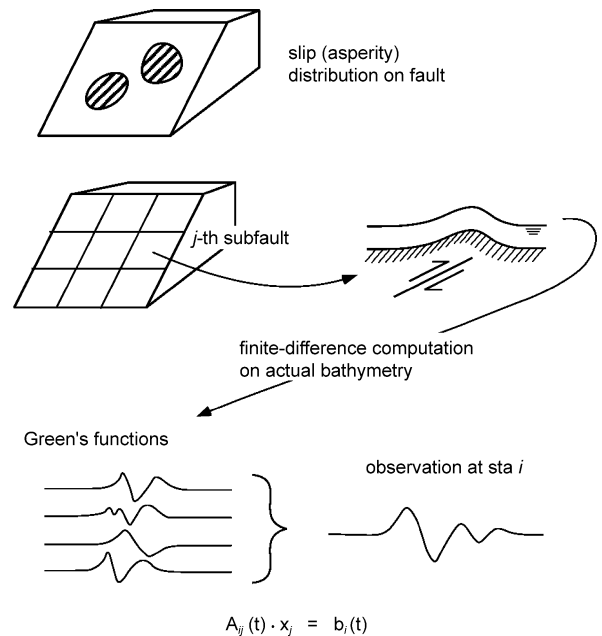
Seismological studies of large earthquakes have indicated that the slip amount is not uniform but heterogeneous on faults. For the 1968 Tokachi-oki earthquake, inversion of far-field body waves or regional Rayleigh waves showed the slip distributions similar to that estimated from tsunami waves [44,45]. The large slip area, called asperity, produces high-frequency seismic waves thus important for

strong-motion prediction for earthquake hazard assessments. The asperity produces large seafloor deformation, hence it is also important for tsunami generation and its hazard estimation. Lay and Kanamori [46] suggested that characteristic size of asperities differs from one subduction zone to another and is controlled by geological setting. Yamanaka and Kikuchi [47], from studies of recurrent earthquakes off northern Honshu, showed that the same asperity ruptures in repeated earthquakes. Their asperity map can be used for earthquake and tsunami hazard assessment.

**Waveform Inversion**

The asperity distribution can be estimated by an inversion of tsunami waveforms. In this method (Fig. 7), the fault plane is first divided into several subfaults, and the seafloor deformation is computed for each subfault with a unit amount of slip. Using these as the initial conditions, tsunami propagation is numerically computed for actual bathymetry and the waveforms at tide gauge stations, called Green’s functions, are computed. Assuming that the tsunami generation and propagation are linear process, the observed tsunami waveforms are expressed as a linear superposition of Green’s functions as follows,

$$A_{ij}(t) \cdot x_j = b_i(t) \tag{15}$$



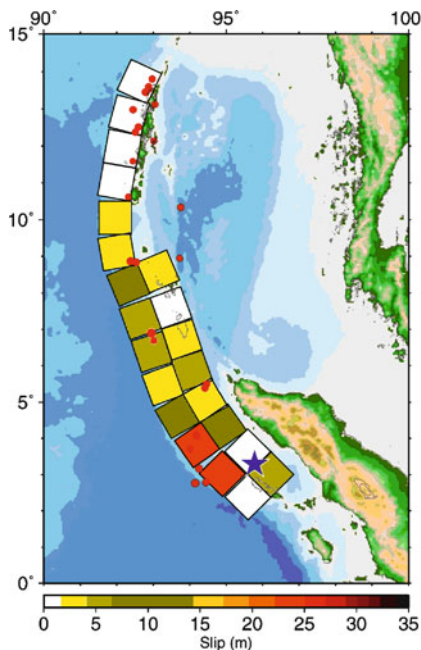
Tsunamis, Inverse Problem of, Figure 7  
Schematic illustration of tsunami waveform inversion method

where  $A_{ij}(t)$  is the computed waveform as a function of time  $t$ , or Green's function, at the  $i$ th station from the  $j$ th subfault;  $x_j$  is the amount of slip on the  $j$ th subfault; and  $b_i(t)$  is the observed tsunami waveform at the  $i$ th station. The slip  $x_j$  on each subfault can be estimated by a least-square inversion of the above set of equations, by minimizing the  $l_2$  norm of the residuals  $J$

$$J = \|A \cdot x - b\| \rightarrow \min \quad (16)$$

where  $A$ ,  $x$ , and  $b$  indicate matrix representations of elements in Eq. (15). Figure 6d shows the slip distribution on the fault for the 1968 Tokachi-oki earthquake. The source fault was divided into 10 subfaults, and slip distribution on the subfaults were estimated. The largest slip, about 3.7 m, was estimated on the subfaults to the west of epicenter, but the average slip is 1.2 m, much smaller than that estimated by the trial and error method (4 m) which compared the maximum tsunami amplitudes [37].

The 2004 Indian Ocean tsunami, caused by the Sumatra–Andaman earthquake, was recorded by satellite altimeters, as well as tide gauges. A joint inversion of tsunami waveforms recorded at 12 tide gauge stations and the sea surface heights measured by three satellites indicates that the tsunami source was about 900 km long [11].



Tsunamis, Inverse Problem of, Figure 8  
Slip distribution on 22 subfaults of the 2004 Sumatra–Andaman earthquake estimated from a joint inversion of tsunami waveforms on tide gauges and sea surface heights measured by satellite altimeters [11]

The estimated slip distribution (Fig. 8) indicates that the largest slip, about 13 to 25 m, was located off Sumatra Island and the second largest slip, up to 7 m, near the Nicobar Islands. Inversion of satellite altimeter data alone supports a longer, about 1,400 km long, tsunami source [23], but such a model produces much larger tsunami waveforms than observed at Indian tide gauge stations. Inversion of tide gauge data alone does not support tsunami source beneath Andaman Islands [48]. The slip distribution estimated by the joint inversion is similar to those estimated from seismological analyses. The fault slip was the largest near off the northern Sumatra, followed by off Nicobar Islands [49]. Fault slip around Andaman Islands was estimated to be small from seismological analysis [50].

### Nonlinear Inversion Methods

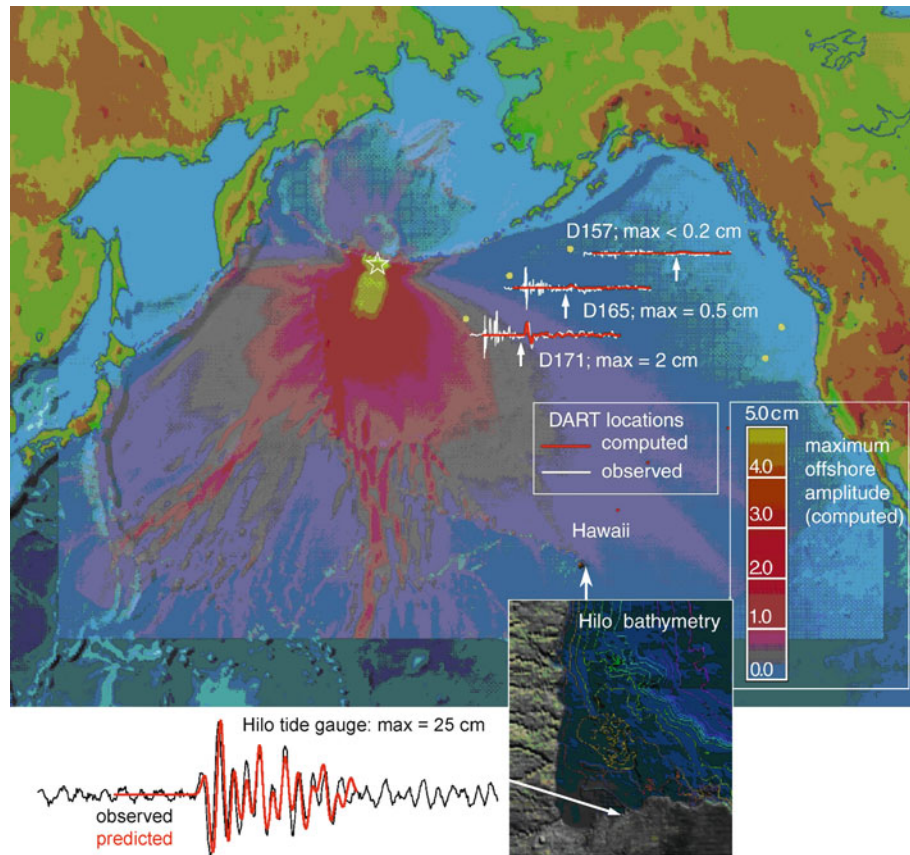
In the above inversion method, both tsunami generation and propagation process are assumed to be linear. Because slip amount  $u$ , among the nine static fault parameters, is linearly related with the seafloor deformation and tsunami amplitudes, and it has the largest effect on tsunamis, other parameters are fixed in the above method. However, the tsunami propagation, particularly near coasts, might be affected by some nonlinear process such as advection or bottom friction.

Nonlinear inversion methods to overcome these limitations have been proposed. Pires and Miranda [51] proposed an adjoint method, which consists of four steps: source area delimitation by backward ray-tracing, the optimization of the initial sea state, nonlinear adjustments of the fault model, and final optimization of fault parameters. The minimum of residual  $J$  is obtained iteratively through gradient descent method using the partial derivative or gradient of  $J$  with respect to parameters to be inverted.

### Inversion of Tsunami Heights

Maximum tsunami heights on coasts, rather than tsunami waveforms, have also been used for tsunami inversion. Distribution of maximum tsunami heights along the coasts is available from field surveys, historical or geological studies, and is valuable to study tsunami source. Pitanesi et al. [52] used coastal tsunami heights of the 1992 Nicaragua “tsunami earthquake” and estimated the slip distribution on the fault, as well as the mean amplification factor of computed coastal heights and measured runup heights.

Annaka et al. [53] proposed a method of joint inversion of tsunami waveforms and runup heights. As a residual to be minimized, they used a weighted sum of difference in waveforms (similar to Eq. 16) and logarithm



Tsunamis, Inverse Problem of, Figure 9

Real-time data assimilation for the November 17, 2003 Rat Island Tsunami [22]. *Star* indicates the epicenter. *Yellow dots* are locations of DART systems. The tsunami waveforms recorded on DART (*white curves*) are compared with computed waveforms (*red*). *Filled colors* on ocean show computed maximum tsunami amplitudes of the updated source. The *bottom plot* shows location of Hilo tide gauge (*right*) and comparison of the predicted tsunami waveforms (*red*) with the recorded tide gauge signal (*black*)

of runup heights. They first tried linear inversion to estimate the initial value, then estimated the perturbation by the nonlinear inversion. Similar nonlinear inversion methods, based on the initial solution estimated by either linear inversion or other data such as geodetic data, have been proposed [54,55].

### Real-Time Data Assimilation

The tsunami waveform inversion can be done in real-time for the purpose of tsunami warning. The real-time data assimilation using the DART records and tsunami forecast have been made by NOAA [22]. They first use seismological information to determine the source location and parameters, then using the database of pre-computed simulation results, invert the DART data to estimate the tsunami source size (slip amounts). The tsunami forecast

is made for farther locations where the tsunami has not arrived. For the 2003 Rat Island (Aleutians) earthquake, they successfully forecasted tsunami waveforms at Hilo, Hawaii, before the tsunami arrivals (Fig. 9).

### Future Directions

Inverse modeling methods of tsunami need to be further developed to better understand the tsunami generation process. Future developments are expected in each field of observation, propagation modeling and application to seismic and nonseismic tsunamis.

The tsunami observation system has been improved recently, particularly after the 2004 Indian Ocean tsunami. Many instrumental data, both coastal and offshore, become available for the studies of tsunami generation process [56]. Maintenance of the systems, particularly for off-

shore systems, is sometimes costly, but essential to record infrequent tsunami. Open and real-time availability of such data is also important for tsunami studies as well as for tsunami warning purposes.

For the past tsunamis, more studies are needed to estimate tsunami heights from historical documents, as well as geological data such as distribution of tsunami deposits. Such historical tsunami database has been developed, e. g., at NOAA/NGDC ([http://www.ngdc.noaa.gov/seg/hazard/tsu\\_db.shtml](http://www.ngdc.noaa.gov/seg/hazard/tsu_db.shtml)).

For modeling tsunamis recorded on coastal tide gauges or runup heights, nonlinear computations with very fine bathymetry data are essential. While computational methods have been developed, fine bathymetry data are not always available. Developments of nonlinear inversion methods are also important.

Finally, inversion of tsunami data can be applied to tsunamis generated from submarine processes other than earthquakes, such as volcanic eruptions or landslides. For such nonseismic tsunamis, parametrization is essential to quantify the geological process and to solve inverse problems.

## Bibliography

### Primary Literature

- Steketee JA (1958) On Volterra's dislocations in a semi-infinite elastic medium. *Can J Phys* 36:192–205
- Mansinha L, Smylie DE (1971) The displacement fields of inclined faults. *Bull Seism Soc Am* 61:1433–1440
- Okada Y (1985) Surface deformation due to shear and tensile faults in a half-space. *Bull Seism Soc Am* 75:1135–1154
- Hanks T, Kanamori H (1979) A moment magnitude scale. *J Geophys Res* 84:2348–2350
- Kanamori H (1977) The energy release in great earthquakes. *J Geophys Res* 82:2981–2987
- Lay T, Kanamori H, Ammon CJ, Nettles M, Ward SN, Aster RC, Beck SL, Bilek SL, Brudzinski MR, Butler R, DeShon HR, Ekstrom G, Satake K, Sipkin S (2005) The great Sumatra–Andaman earthquake of 26 December 2004. *Science* 308:1127–1133
- Stein S, Okal EA (2005) Speed and size of the Sumatra earthquake. *Nature* 434:581–582
- Tsai VC, Nettles M, Ekstrom G, Dziewonski AM (2005) Multiple CMT source analysis of the 2004 Sumatra earthquake. *Geophys Res Lett* 32. doi:10.1029/2005GL023813
- Geist E (1998) Local tsunamis and earthquake source parameters. *Adv Geophys* 39:117–209
- Yamashita T, Sato R (1974) Generation of tsunami by a fault model. *J Phys Earth* 22:415–440
- Fujii Y, Satake K (2007) Tsunami source model of the 2004 Sumatra–Andaman earthquake inferred from tide gauge and satellite data. *Bull Seism Soc Am* 97:S192–S207
- Satake K, Tanioka Y (2003) The July 1998 Papua New Guinea earthquake: Mechanism and quantification of unusual tsunami generation. *Pure Appl Geophys* 160:2087–2118
- Tanioka Y, Satake K (1996) Tsunami generation by horizontal displacement of ocean bottom. *Geophys Res Lett* 23:861–864
- Song YT, Fu L-L, Zlotnicki V, Ji C, Hjorleifsdottir V, Shum CK, Yi Y (2008) The role of horizontal impulses of the faulting continental slope in generating the 26 December 2004 tsunami. *Ocean Modell* 20:362–379
- Intergovernmental Oceanographic Commission (1997) IUGG/IOC TIME Project Numerical Method of Tsunami Simulation with the Leap-frog Scheme. UNESCO, Paris
- Mader CL (1988) Numerical modeling of water waves. University of California Press, Berkeley
- Yeh H, Liu P, Synolakis C (1996) Long-wave runup models. World Scientific, Singapore
- Smith WHF, Scharroo R, Titov VV, Arcas D, Arbic BK (2005) Satellite altimeters measure tsunami, early model estimates confirmed. *Oceanography* 18:10–12
- Kato T, Terada Y, Kinoshita M, Kakimoto H, Isshiki H, Matsuishi M, Yokoyama A, Tanno T (2000) Real-time observation of tsunami by RTK-GPS. *Earth Planet Space* 52:841–845
- Mikada H, Mitsuzawa K, Matsumoto H, Watanabe T, Morita S, Otsuka R, Sugioka H, Baba T, Araki E, Suyehiro K (2006) New discoveries in dynamics of an M8 earthquake-phenomena and their implications from the 2003 Tokachi-oki earthquake using a long term monitoring cabled observatory. *Tectonophysics* 426:95–105
- Gonzalez FI, Bernard EN, Meinig C, Eble MC, Mofjeld HO, Stalin S (2005) The NTHMP tsunameter network. *Nat Hazard* 35: 25–39
- Titov VV, Gonzalez FI, Bernard EN, Eble MC, Mofjeld HO, Newman JC, Venturato AJ (2005) Real-time tsunami forecasting: Challenges and solutions. *Nat Hazard* 35:41–58
- Hirata K, Satake K, Tanioka Y, Kuragano T, Hasegawa Y, Hayashi Y, Hamada N (2006) The 2004 Indian Ocean tsunami: Tsunami source model from satellite altimetry. *Earth Planet Space* 58:195–201
- Synolakis CE, Okal EA (2005) 1992–2002: Perspective on a decade of post-tsunami surveys. In: Satake K (ed) *Tsunamis: Case studies and recent developments*. Springer, Dordrecht, pp 1–29
- Intergovernmental Oceanographic Commission (1998) Post-tsunami survey field guide. UNESCO, Paris
- Satake K, Wang KL, Atwater BF (2003) Fault slip and seismic moment of the 1700 Cascadia earthquake inferred from Japanese tsunami descriptions. *J Geophys Res* 108. doi:10.1029/2003JB002521
- Atwater BF, Musumi-Rokkaku S, Satake K, Tsuji Y, Ueda K, Yamaguchi DK (2005) The orphan tsunami of 1700. *USGS Prof Paper* 1707:133
- Dawson AG, Shi SZ (2000) Tsunami deposits. *Pure Appl Geophys* 157:875–897
- Nanayama F, Satake K, Furukawa R, Shimokawa K, Atwater BF, Shigeno K, Yamaki S (2003) Unusually large earthquakes inferred from tsunami deposits along the Kuril trench. *Nature* 424:660–663
- Satake K (1988) Effects of bathymetry on tsunami propagation – Application of ray tracing to tsunamis. *Pure Appl Geophys* 126:27–36
- Rabinovich AB, Thomson RE (2007) The 26 December 2004 Sumatra tsunami: Analysis of tide gauge data from the world ocean Part 1, Indian Ocean and South Africa. *Pure Appl Geophys* 164:261–308



32. Miyabe N (1934) An investigation of the Sanriku tsunami based on mareogram data. *Bull Earthq Res Inst Univ Tokyo Suppl* 1:112–126
33. Merrifield MA, Firing YL, Aarup T, Agricole W, Brundrit G, Chang-Seng D, Farre R, Kilonsky B, Knight W, Kong L, Magori C, Manurung P, McCreery C, Mitchell W, Pillay S, Schindele F, Shillington F, Testut L, Wijeratne EMS, Caldwell P, Jardin J, Nakahara S, Porter FY, Turetsky N (2005) Tide gauge observations of the Indian Ocean tsunami, December 26, 2004. *Geophys Res Lett* 32. doi:10.1029/2005GL022610
34. Nagarajan B, Suresh I, Sundar D, Sharma R, Lal AK, Neetu S, Shenoi SSC, Shetye SR, Shankar D (2006) The great tsunami of 26 December 2004: A description based on tide-gauge data from the Indian subcontinent and surrounding areas. *Earth Planet Space* 58:211–215
35. Neetu S, Suresh I, Shankar R, Shankar D, Shenoi SSC, Shetye SR, Sundar D, Nagarajan B (2005) Comment on “The great Sumatra–Andaman earthquake of 26 December 2004”. *Science* 310:1431a
36. Abe K (1973) Tsunami and mechanism of great earthquakes. *Phys Earth Planet Inter* 7:143–153
37. Satake K (1989) Inversion of tsunami waveforms for the estimation of heterogeneous fault motion of large submarine earthquakes – The 1968 Tokachi-Oki and 1983 Japan Sea earthquakes. *J Geophys Res* 94:5627–5636
38. Mei CC (1989) *The applied dynamics of ocean surface waves*. World Scientific, Singapore
39. Abe K (1979) Size of great earthquakes of 1873–1974 inferred from tsunami data. *J Geophys Res* 84:1561–1568
40. Abe K (1981) Physical size of tsunamigenic earthquakes of the northwestern Pacific. *Phys Earth Planet Inter* 27:194–205
41. Abe K (1989) Quantification of tsunamigenic earthquakes by the Mt scale. *Tectonophysics* 166:27–34
42. Aida I (1978) Reliability of a tsunami source model derived from fault parameters. *J Phys Earth* 26:57–73
43. Yanagisawa K, Imamura F, Sakakiyama T, Annaka T, Takeda T, Shuto N (2007) Tsunami assessment for risk management at nuclear power facilities in Japan. *Pure Appl Geophys* 164: 565–576
44. Kikuchi M, Fukao Y (1985) Iterative deconvolution of complex body waves from great earthquakes – The Tokachi-oki earthquake of 1968. *Phys Earth Planet Inter* 37:235–248
45. Mori J, Shimazaki K (1985) Inversion of intermediate-period Rayleigh waves for source characteristics of the 1968 Tokachi-oki earthquake. *J Geophys Res* 90:11374–11382
46. Lay T, Kanamori H (1981) An asperity model of large earthquake sequences. In: Simpson DW, Richards PG (eds) *Earthquake prediction – An international review*. American Geophysical Union, Washington DC, pp 579–592
47. Yamanaka Y, Kikuchi M (2004) Asperity map along the subduction zone in northeastern Japan inferred from regional seismic data. *J Geophys Res* 109:B07307, doi:10.1029/2003JB002683
48. Tanioka Y, Yudhicara, Kusunose T, Kathioli S, Nishimura Y, Iwasaki S-I, Satake K (2006) Rupture process of the 2004 great Sumatra–Andaman earthquake estimated from tsunami waveforms. *Earth Planet Space* 58:203–209
49. Ammon CJ, Ji C, Thio HK, Robinson D, Ni SD, Hjorleifsdottir V, Kanamori H, Lay T, Das S, Helmberger D, Ichinose G, Polet J, Wald D (2005) Rupture process of the 2004 Sumatra–Andaman earthquake. *Science* 308:1133–1139
50. Velasco AA, Ammon CJ, Lay T (2006) Search for seismic radiation from late slip for the December 26, 2004 Sumatra–Andaman (Mw=9.15) earthquake. *Geophys Res Lett* 33:L18305, doi:10.1029/2006GL027286
51. Pires C, Miranda PMA (2001) Tsunami waveform inversion by adjoint methods. *J Geophys Res* 106:19773–19796
52. Piatanesi A, Tinti S, Gavagni I (1996) The slip distribution of the 1992 Nicaragua earthquake from tsunami run-up data. *Geophys Res Lett* 23:37–40
53. Annaka T, Ohta K, Motegi H, Yoshida I, Takao M, Soraoka H (1999) A study on the tsunami inversion method based on shallow water theory. *Proc Coastal Engin JSCE* 46:341–345
54. Yokota T, Nemoro M, Masuda T (2004) Estimate of slip distribution by tsunami height data inversion. *Abst Jpn Earth Planet Sci Joint Meeting* S043-P0005
55. Namegaya Y, Tsuji Y (2007) Distribution of asperities of the 1854 Ansei Nankai earthquake. *Abst Jpn Earth Planet Sci Joint Meeting* S142-009
56. Satake K, Baba T, Hirata K, Iwasaki S, Kato T, Koshimura S, Takenaka J, Terada Y (2005) Tsunami source of the 2004 off the Kii peninsula earthquakes inferred from offshore tsunami and coastal tide gauges. *Earth Planet Space* 57:173–178

## Books and Reviews

- Lawson CL, Hanson RJ (1974) *Solving least squares problems*. Prentice-Hall, Englewood Cliffs. (Republished by Society for Industrial and Applied Mathematics, 1995)
- Lay T, Wallace TC (1995) *Modern global seismology*. Academic Press, San Diego
- Menke W (1989) *Geophysical data analysis: Discrete inverse theory* (revised edition). Academic Press, San Diego
- Satake K (2007) *Tsunamis*. In: Kanamori H (ed) *Treatise on Geophysics*, vol 4. Elsevier, Amsterdam

## Volcanic Eruptions, Explosive: Experimental Insights

STEPHEN J. LANE, MICHAEL R. JAMES  
Lancaster Environment Centre, Lancaster University,  
Lancaster, UK

### Article Outline

Glossary  
Definition of the Subject  
Introduction  
Volcanic Materials  
Volcanic Processes  
Analogue Approach  
Future Directions  
Acknowledgments  
Bibliography

### Glossary

**Conduit** Subterranean pathways to the surface created and sustained by the flow of volcanic materials are known as volcanic conduits. In the laboratory, tubes are used as scale equivalents. The two terminologies ('conduits' and 'tubes') should be kept distinct to avoid inadvertent acceptance that a laboratory experiment is a direct replica of a volcanic process.

**Magma** When rock melts underground it becomes magma. When magma flows on the surface it is known as lava and when it is projected into the atmosphere it forms pyroclasts. Most magma is silicate-based and comprises a three-phase mixture of solid (crystals, called microlites when small), liquid (melt) and gas (bubbles or vesicles).

**Rheology** When a force is applied to matter it deforms. The rheology of a material describes the relationship between applied force and the material response. Volcanic materials have a wide range of rheological behaviors dependent on the proportions of gas, liquid and solid, temperature, chemical composition, deformation history and rate of deformation.

**Scaling** Explosive volcanic eruptions involve temperatures of order 1000°C, pressures of many MPa and physical sizes from  $\mu\text{m}$  to km. Laboratory study of these processes generally involves observation and measurement of processes operating at lower temperature or pressure, or at smaller physical size. Scaling is the means of making laboratory-scale analogues representative of volcano-scale natural processes.

**Volatiles** Explosive volcanic eruptions are driven by the exsolution of a dissolved gas phase as magma pressure falls. These gaseous species, dominated by water vapor on Earth, are known as volatiles.

### Definition of the Subject

The magnitude of volcanic eruptions may be classified using the Volcanic Explosivity Index (VEI) [95,99]. The lowest VEI index of zero covers non-explosive activity (such as the production of lava flows) and is generally associated with low magma effusion rates or low viscosity magmas, for example, carbonate-based carbonatites [127]. For silicate-based magmas, which are generally more viscous than carbonatites, the boundary between effusive and explosive volcanism is commonly marked by Hawaiian firefountaining activity [96]. Magmas with a high silica content, or a lower temperature (and thus higher viscosity), can erupt either explosively or non-explosively as lava flows or dome-building eruptions [128]. Some eruptions switch between effusive and explosive behavior dependent on prevailing physicochemical conditions (e.g., [54,62]), with magmas tending to become more explosive at higher effusion rates. High VEI eruptions are, therefore, associated with high flow rates, large erupted volumes and substantial eruption plumes.

Such explosive activity comprises a wide range of phenomena, which result from complex non-linear interactions and feedback mechanisms. Processes that initiate in the subsurface plumbing or conduit system determine the nature of following events and control eruption styles. On emerging from the vent, volcanic material enters the atmosphere and the ensuing interactions are key in determining the consequent transport of volcanic debris that defines the impact on the environment and on human lives and infrastructure. On Earth, our capability to mitigate such volcanic hazard relies in large part on forecasting explosive events, a process which requires a high degree of understanding about the physicochemical factors operating during explosive volcanism.

In common with many science disciplines, the approaches taken to gain an understanding of explosive volcanism have relied on a combination of field observations, theoretical models and laboratory models of materials and mechanisms. Field observations form the 'ground truth' for numerical and experimental modeling (and provide vital historical data), but are often very difficult to unambiguously interpret due to their scarcity in both space and time. Numerical modeling complements field observations by providing much greater information density and, hence, the potential to examine processes in great de-

tail. However, the accuracy of the results depends on correctly including all relevant materials properties and system processes in the model and the results can be difficult to verify against often-unobservable field phenomena. Laboratory investigation of the properties of volcanic materials, their modes of interaction and the behaviors of analogous systems provides the third complementary approach; this is the area we focus on here.

## Introduction

In the context of explosive volcanic activity, laboratory experiments cover a wide range of research, the most fundamental of which is to gain an understanding of the properties of the volcanic material involved. Without this to underpin further work, neither accurate computer nor laboratory simulations could be carried out. The materials properties that are relevant to explosive volcanism include the solubility in silicate melts of volatiles (most commonly water and carbon dioxide), melt viscosity as a function of chemical composition, magma rheology as a function of the size, shape and proportion of crystals and bubbles, gas permeability of magma, magma deformation mode (e. g. brittle or ductile) as a function of strain-rate, and volatile diffusivity. Experimental determination of these parameters and their interdependencies results in empirically formulated relationships that can be used in numerical models and compared with the properties of analogue fluids. We first review aspects of the volcanic materials literature in Sect. “[Volcanic Materials](#)”, with the aim of illustrating the nature of molten rock, the complexity of which underpins most explosive volcanic processes.

Experimental modeling of these processes can then build on the materials understanding. Such experiments involve investigation of the behavior of natural volcanic products at laboratory time and length scales, including the response of magma samples to rapid changes in pressure and temperature, the fall behavior of silicate particles in the atmosphere, and the generation and separation of electrostatic charge during explosive eruptions. Here, we encounter the concept of scaling [[11,20](#)], which links the applicability of experimental results using small samples of natural materials to natural scales that are often several orders of magnitude larger. For experiments that replace natural volcanic products with analogue materials (for example, in order to relax the experimental pressure and temperature requirements), scaling arguments are crucial. A scaling approach permits adjustment of experimental durations and sizes, whilst maintaining similitude with much larger and longer events. Consequently, analogue exper-

imentation has allowed investigation of explosive phenomena in the broader context of fluid-dynamic behavior, and is a powerful means of giving ‘insight’ into volcanic processes.

Burgisser et al. [[20](#)] review and discuss scaling criteria for experiments and natural phenomena involving dilute multiphase mixtures in magmatic systems, and conclude “... that, despite numerous experimental studies on processes relevant to magmatic systems, some and perhaps most geologically important parameter ranges still need to be addressed at the laboratory scale”. This emphasizes that small-scale experiments aiming to mimic larger-scale natural processes need to be considered in two contexts. Firstly, laboratory experiments yield valid results in their own right and provide a means of validating numerical models. Secondly, laboratory experiments are not facsimiles of the natural process, but can give perception into how aspects of natural systems behave provided those aspects are scaled; a process normally considered at design stage. Consequently, comparing experimental results with the natural process must always be undertaken with due caution. The analogue experiment literature is reviewed in Sects. “[Volcanic Processes](#)” and “[Analogue Approach](#)”.

In Sect. “[Definition of the Subject](#)”, the combined tactics of field observation, numerical model and experimental model was discussed. In practice, complete overlap of these techniques can be difficult to achieve because the parameters measurable in the field are often incompatible with experimentally accessible variables, which in turn may not coincide with the output from numerical models. Explosive volcanism is also a difficult phenomenon to observe in action, and there are three reasons for this. Firstly, explosions do not happen very often, and the larger the explosion the more infrequent it is, with, for example, repeat timescales of order  $10^3$  years for VEI 6 events like the Krakatau, 1883 eruption. This makes detailed syn-event measurements of large explosive episodes infrequent on the timescale of contemporary volcanology. Secondly, high VEI events tend to destroy near-field observation equipment, making measurements very difficult to obtain even when an event does occur. Thirdly, processes that comprise an explosive event can be difficult to access directly; perhaps a prime example of this is how to obtain data about the nature of flow in the volcanic conduit system during explosive activity. Direct measurements are not possible and our main source of information is the deformation of the conduit wall created by unsteadiness in the flow, which in turn is calculated from the interpretation of ground motion signals detected by seismometers (e. g., [[31](#)]).

These considerations mean that there are many field observations, numerical and experimental models that await comparison with their companion approaches, however, the detailed study of volcanoes with low VEIs is one area where combined tactics are possible. As an example, in Sect. “**Analogue Approach**” we review the contribution that experimental modeling has made to the understanding of fluid flow in the conduit feeding VEI 1–2 explosions at Stromboli Volcano, Italy. Here it has been possible to measure signals from many thousand events, over the period of a few years (e. g., [7]), with a network of seismometers and other instruments, as well as detailed sampling of eruption products (e. g., [72]). Such field data are key to testing both numerical and experimental models of explosive volcanic activity wherever it occurs.

## Volcanic Materials

### Magma Rheology

Rheology describes the way in which materials deform and flow under the influence of an applied stress. Explosive eruptions only occur because considerable stress is applied to magma. The rheology of natural silicate materials was being measured before World War II (e. g., [14]), with empirical models (derived from experimental data) becoming established in the 1960s and 1970s (e. g., [18,92,107] and references therein).

When stress is applied to a solid it deforms or strains elastically, the strain being the macroscopic result of atoms in the solid moving closer together or further apart. These two parameters are related by a constant called a modulus, for example, in solids under tension  $\sigma = E\gamma$ , where  $\sigma$  is tensile stress,  $\gamma$  is tensile strain and  $E$  is Young’s modulus. Similar relationships exist for deformation under shear or by pressure change. These moduli are likely to be a constant for a range of values of applied stress, but could become dependent on strain, or indeed strain rate.

When stress ( $\sigma$ ) is applied to a fluid (the collective term for gas and liquid), viscosity ( $\eta$ ) is the parameter that relates the rate of flow (strain rate,  $d\gamma/dt$ ) to the stress that is being applied to it, giving the relationship  $\sigma = \eta d\gamma/dt$ . The flow of a fluid is the macroscopic result of fluid molecules moving past each other; therefore, changes that influence molecular motion also affect viscosity. The viscosity can be independent of the rate of flow, a case known as Newtonian behavior, and is best exhibited by gases and liquids of low molecular weight. Fluids whose viscosity is a function of strain-rate are described as non-Newtonian, and many foodstuffs, for instance, exhibit this behavior (e. g. tomato ketchup). Viscosity is also a function of temperature, generally decreasing as temperature

increases and molecules move further apart (thermal expansion). An everyday example would be Golden or Maple syrup that becomes easier to pour on heating. The concept of viscosity fails when molecular interactions are rare, for example in gases at low pressure.

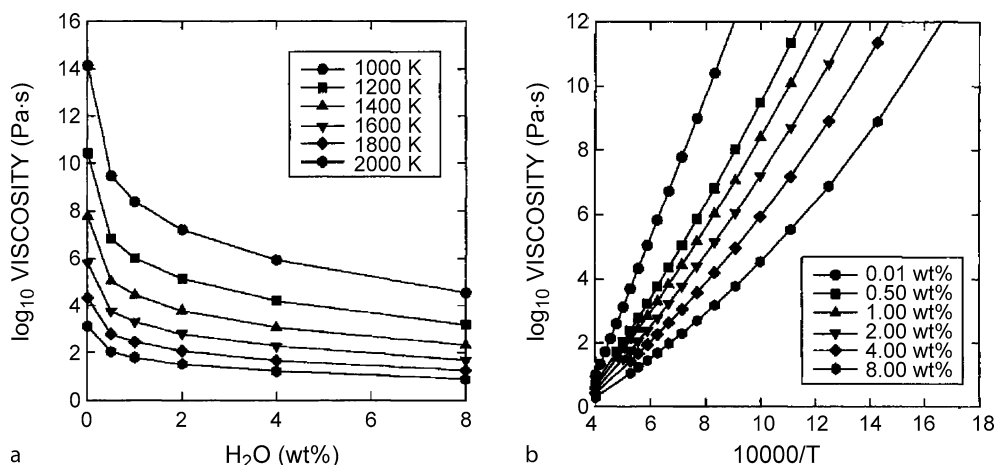
**Newtonian Viscosity** Natural silicates in the liquid state have been assumed to be Newtonian in behavior and demonstrate a strong viscosity dependence on temperature. The relationship was initially represented by an Arrhenius-type equation, such as  $\eta = A \exp[E/RT]$ , where the constants  $A$  and  $E$  are dependent on the chemical composition of the silicate melt (e. g., Murase and McBirney [92]),  $R$  is the gas constant and  $T$  the absolute temperature. With considerable further laboratory effort in measuring the viscosities of a wide compositional range of natural silicate liquids, it became clear that viscosity was not strictly an Arrhenian function of temperature. Hess and Dingwell [49] developed an empirical model to describe laboratory values of viscosity for leucogranitic melts as a function of temperature and water content. The compilation of data used by Hess and Dingwell [49] is shown in Fig. 1, and the empirical equation generated to fit the data is:

$$\log \eta = \frac{[-3.545 + 0.833\ln(w)] + [9601 - 2368\ln(w)]}{\{T - [195.7 + 32.25\ln(w)]\}}$$

where  $w$  is the water % w/w. The constants in this equation will be different for silicate melts of different composition; see Zhang et al. [133] for an updated and modified version of this model. Water is important here because it is a ubiquitous component of explosive volcanism on Earth. The concentration of water has a large and non-linear impact on viscosity (Fig. 1a) because it acts as a network modifier by breaking Si-Si bonds in a silicate melt, thus decreasing the average molecular weight.

The use of laboratory measurement to provide quantitative understanding of the dependence of viscosity on temperature, water content and other parameters has provided a key component in the numerical modeling of the explosive flow behavior of natural silicates, as well as pointing the way to identifying fluids for analogue experiments.

**Non-Newtonian Viscosity** There is laboratory evidence that the viscosity of liquid magmas is not a constant as once assumed, but may be subtly non-Newtonian [110]. The contents of a volcanic conduit are rarely 100% liquid, with crystals present in many silicate melts below 1250°C, considerably higher than most eruption temperatures. The



Volcanic Eruptions, Explosive: Experimental Insights, Figure 1

Model curves, based on experimental measurement, of the dependence of viscosity of metaluminous leucogranitic melts on dissolved water content (a) and temperature (b) from a compilation of data (Hess and Dingwell [49]). Note the logarithmic viscosity scale and non-linear nature of the relationships. Explosive volcanic activity often results in water loss from the melt, and cooling results from expansion and interaction with the atmosphere; melt viscosity increases result. The role this plays on the dynamics of expanding vesiculated flows may be studied by analogue experimental means. Figure reproduced with kind permission of the Mineralogical Society of America

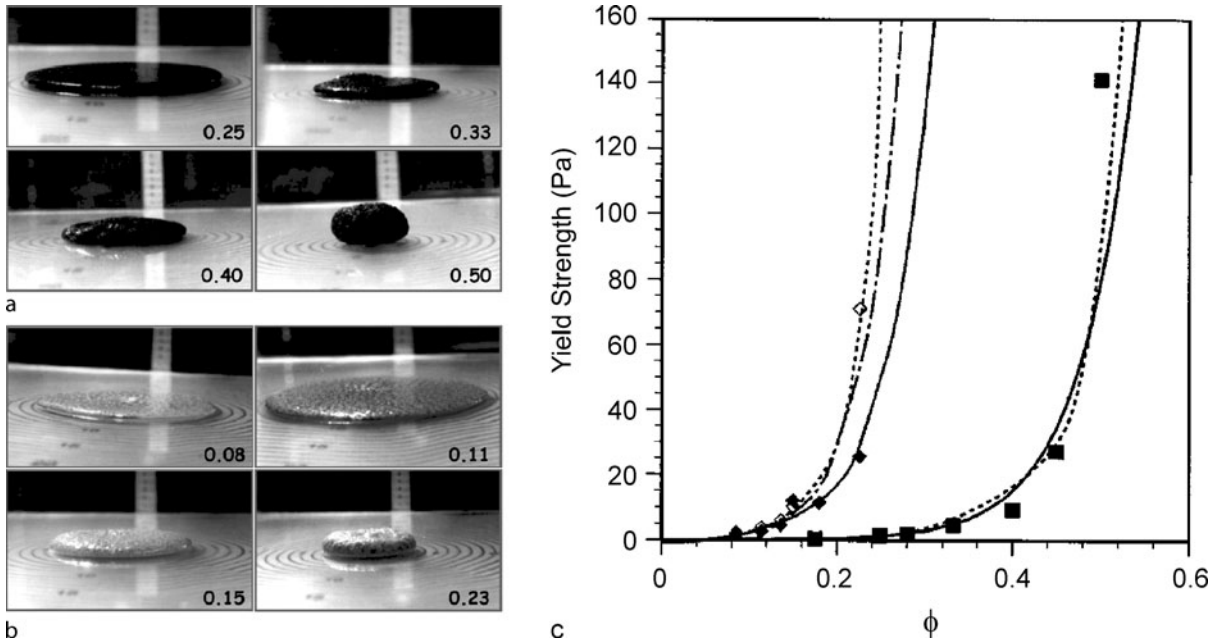
effect of crystallization is to significantly change the relationship between viscosity and temperature [91], especially as crystal content is both temperature and time dependent. The presence of bubbles is also ubiquitous in explosive volcanic materials, and these too alter magma behavior in response to an applied stress. The presence of both yields a rheologically complex 3-phase material.

Combined laboratory and field-measurement of the viscosity of lavas from Mount Etna [98] showed that above 1120°C Newtonian behavior was exhibited. Eruption temperatures ranged from 1084–1125°C; therefore, for much of this range the magma behaved in a non-Newtonian fashion as a consequence of the interaction between crystals suspended within the liquid silicate melt. The Etnean basalt was found to be thixotropic, meaning that the viscosity decreased the longer it was deformed, perhaps indicating that crystals were becoming aligned and could consequently move past each other more readily. The basalt also showed pseudoplastic or shear-thinning behavior, meaning that viscosity decreased as the rate of flow increased, perhaps again, because of the dynamics of crystal alignment. Finally, the basalt showed a small but measurable yield stress of < 78 Pa, below which it behaved more like a solid material. The yield stress is likely to be a consequence of crystals weakly binding together into a mesh-like structure. The small nature of the yield stress means that the viscosity of these magmas could be described by a power law equation (i. e., the yield stress was ignored)

where  $\sigma = \eta (d\gamma/dt)^n$ . Above 1120°C,  $n = 1$  and Newtonian behavior was observed, but as temperature falls to 1084°C,  $n$  decreases to 0.46, meaning that viscosity decreases as strain rate increases. Pinkerton and Norton [98] also noticed that physical motion of the silicate melt could induce crystallization as seen by Kouchi et al. [68]. Similar behavior has been found more recently for laboratory measurements of basalts from Japan [55].

Hoover et al. [51] showed that crystals induce yield strength in magmas at volume fractions as low as 0.2, although this is strongly dependent on crystal shape. These experiments were carried out on natural basaltic samples, as well as using an analogous system of neutrally buoyant particles in corn syrup (Fig. 2). Natural samples do, by definition, provide data directly applicable to the natural system, but analogue experiments are complementary because particles have controlled size, shape and concentration. These experiments provide a good example of the kind of complex interplay of time-dependent processes that can occur in magmatic systems that tend only to be quantified by detailed experimentation.

At eruption temperatures, the non-Newtonian behavior of magmas is due in large part to the presence of crystal networks; however, during flow, bubbles must also move past each other and contribute to fluid rheology. Explosive eruptions only occur because bubbles form and expand; therefore, understanding bubble behavior is central to understanding explosive processes. When both bubble size



Volcanic Eruptions, Explosive: Experimental Insights, Figure 2

Magma has complex rheology partly because of its crystal content. Hoover et al. [51] found that basalt with crystals showed solid-like behavior at low stress. Analogue experiments using near-spherical poppy seeds (a) and highly irregular shavings of kaolin/wax (b) in corn syrup demonstrate the role of particle concentration and shape in giving yield strength to solid-liquid mixtures, the number in each panel being the volume fraction of particles. Yield strength increased rapidly (c) for volume fractions of round poppy seed (filled squares) above about 0.4, allowing extruded material to retain considerable vertical topography (a). The same effect occurs at volume fractions around 0.2 for irregular particles (b) because they interact more strongly with each other than spherical particles. Crystals in magma not only act as nucleation sites for bubbles, but also have a very strong rheological influence in their own right. Reprinted with kind permission from Elsevier

and the fluid flow rate are small, then the surface tension between the silicate liquid and the water vapor in the bubble resists viscous forces and is able to minimize the bubble surface area maintaining spherical bubbles; the bubbles act as if they are solid spheres suspended in the silicate melt. However, if bubbles and flow rate are larger, then surface tension is no longer able to maintain a spherical shape against the viscous drag, and bubbles elongate in the flow field. Under conditions of steady flow, when bubble shape is stable with time, these two behaviors may be distinguished by the capillary number,  $Ca = \lambda \gamma'$  where  $\lambda = a\eta/\sigma$ ,  $a$  being bubble diameter,  $\eta$  viscosity,  $\gamma'$  strain rate and  $\sigma$  surface tension (e.g., Llewellyn and Manga [74]). When  $Ca$  is significantly less than unity, surface tension forces dominate and bubbles remain near spherical. However, when  $Ca$  is significantly greater than unity surface tension is overcome by the combination of liquid shear and viscosity, forcing bubbles to elongate.  $Ca$  provides an example of a parameter that may be used to scale between volcanic behavior and a laboratory exper-

iment. If bubbles preserved in pumice from an eruption were highly elongated, then any experiment simulating this process would need to also produce highly elongated bubbles and not spherical bubbles. The shape-change undergone by bubbles provides an added complexity over the effects of crystals, whose shapes are generally constant.

Bagdassarov and Dingwell [8,9] experimentally measured the rheological behavior of vesicular rhyolite, using oscillatory rheometry, with a liquid viscosity between  $10^9$  and  $10^{12}$  Pa s and gas bubble volume fractions in the region between zero and 0.3. Under these conditions, where  $Ca \gg 1$ , it was shown that vesicular rhyolite decreased in viscosity with increasing gas volume fraction. Such rheological experiments on hot, high-viscosity natural silicate melts are difficult, and the scaling of the small-strain results to larger scales is not straightforward. Using an oscillating viscometer, Llewellyn et al. [75] measured the viscosity of bubbly Golden Syrup, a fluid analogous to molten silicate with water vapor bubbles, but with liquid viscosity between  $10^1$  and  $10^2$  Pa s. The analogue fluid has the ad-

vantage of being liquid at room temperature, thereby placing much less stringent demands on any measurement apparatus. Llewellyn et al. [75] showed the effect of increasing gas volume fraction from zero to 0.5 with small deformations ( $Ca \ll 1$ ) caused viscosity to change by over an order of magnitude. They also discriminated between steady flow (bubbles not changing shape) and unsteady flow (bubbles in the process of changing shape) by proposing the dynamic capillary number  $Cd = \lambda\gamma''/\gamma'$ , where  $\lambda$  represents the timescale over which a deformed bubble can relax. Under steady flow conditions bubbles tended to increase viscosity, but under conditions of rapidly changing strain rate increasing gas volume fraction reduced viscosity; a result consistent with the oscillatory rheometry results of Bagdassarov and Dingwell [9]. This experimental result largely unified the seemingly contradictory literature that preceded it, and the consequences of these two regimes of behavior are discussed lucidly by Llewellyn and Manga [74], where the influence of bubbles is shown to change some model outcomes by factors of two or more. This is an excellent example of the role that analogue experimental measurement can play in helping to understand complex and inaccessible natural phenomena.

**Viscosity or Modulus?** The concept of a relaxation time was encountered above in relation to how long a deformed (non-spherical) bubble takes to approach the relaxed spherical shape. This same approach can also be applied at the molecular level as liquids flow. During viscous flow, molecules move past each other, and this requires them to be able to reorient or deform. If the timescale of molecular deformation is much shorter than the timescale required by the strain rate during flow, then liquid or viscous behavior ensues. However, if the strain rate is such that molecules cannot reorient or deform, then the material behaves, at least in part, like an elastic solid, with strain being accommodated by bond stretching, or fracture, if yield stress is exceeded.

Materials that demonstrate aspects of liquid and solid behavior are known as viscoelastic. Viscoelasticity was first investigated in the nineteenth century using materials like rubber and glass. Viscoelastic behavior in volcanic materials begins with the onset of non-Newtonian liquid behavior at deformation timescales about 100 times longer than the relaxation time of silicate molecules in the melt [39]. The onset of non-Newtonian behavior suggests that there is a range of relaxation timescales for different molecules, and this may be a consequence of silicate melts having a range of molecular weights rather than a single value.

The *timescale* of molecular relaxation is given by the Maxwell relation, *viscosity/modulus*, i. e., the relative

weighting of liquid to solid response when stress is applied. In volcanic materials the modulus is effectively constant as temperature and composition change, but viscosity changes by a factor of about  $10^9$  or more. At low magma viscosities, the relaxation timescale is short and viscous liquid behavior occurs at volcanically plausible strain rates. However, magmas of high viscosity have long relaxation timescales. Here, the onset of non-Newtonian behavior, and ultimately the glass transition to brittle solid behavior, occurs at much lower strain rates. Dingwell [38] discusses the relevance of this to explosive volcanic eruptions, and proposes the onset of brittle failure in high-viscosity magma under significant strain rates as a criterion for magma fragmentation. Zhang [130] showed that magma fragmentation also depends on the number density and size of bubbles within the melt.

### Volatiles Solubility

Volatiles are defined as those compounds that are dissolved within the matrix of a silicate melt at high pressures, but become supersaturated and, consequently, then exsolve in the supercritical or gaseous state as pressure drops towards atmospheric. On Earth,  $H_2O$  is the most common magmatic volatile, with  $CO_2$ ,  $SO_2$ , HCl, and the halogens also present. For water, exsolution occurs into the supercritical state, or gas phase at depths less than about 1 km; therefore, understanding the physicochemical behavior of these compounds at volcanic temperatures and pressures is a key component in the modeling of explosive volcanic eruptions. Water itself has been the subject of extensive laboratory and numerical experiments (with Kalinichev [65] providing a review) because of its anthropogenic usefulness as well as geological importance. The densities of volatile species are less than that of the silicate melt, for instance at the depths of exsolution the density of water is about  $200\text{--}400\text{ kg m}^{-3}$ , a factor of  $\sim 10$  less than magma. The bulk modulus of water (order  $10^7$  Pa) is much less than that of silicate melt (about  $10^{10}$  Pa, Alidibirov et al. [5]). Consequently, as pressure falls, any exsolved volatile phase expands by about three orders of magnitude more than that of the surrounding silicate melt; this is the fundamental driving mechanism of explosive volcanism.

The solubility of volatiles in silicate melts has, and continues to be, the subject of extensive experimentation. Carroll and Holloway [28] review volatiles in magmas, and Liu et al. [73] provide a well-referenced introduction to water and carbon dioxide in rhyolite, and give new experimental data and empirical relationships so important to numerical models of explosive volcanic activity. Empirical  $H_2O$  and  $CO_2$  solubility in rhyolitic melt ( $700\text{--}1200^\circ\text{C}$  and

0–500 MPa) are expressed by Liu et al. [73] as

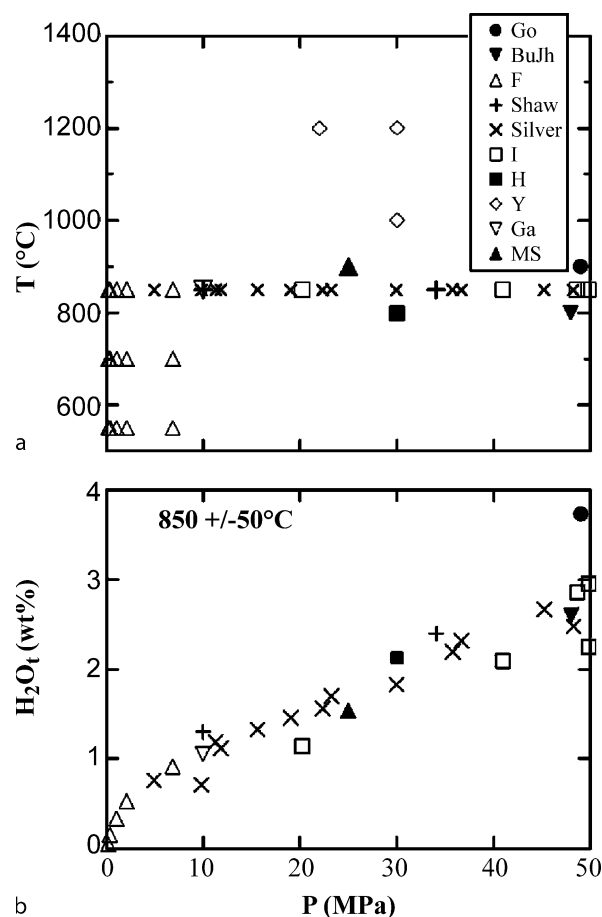
$$\begin{aligned}
 \text{H}_2\text{O}_t &= \frac{(354.94P_w^{0.5} + 9.623P_w - 1.5223P_w^{1.5})}{T} \\
 &+ 0.0012439P_w^{1.5} \\
 &+ P_{\text{CO}_2} (-1.084 \times 10^{-4}P_w^{0.5} - 1.362 \times 10^{-5}P_w), \\
 \text{CO}_2 &= \frac{P_{\text{CO}_2}(5668 - 55.99P_w)}{T} \\
 &+ P_{\text{CO}_2} (0.4133P_w^{0.5} + 2.041 \times 10^{-3}P_w^{1.5}),
 \end{aligned}$$

where  $\text{H}_2\text{O}_t$  is the total dissolved  $\text{H}_2\text{O}$  % w/w,  $\text{CO}_2$  content is in mass-ppm,  $T$  is absolute temperature and  $P$  the partial pressure of water or carbon dioxide.

Figure 3 shows how the solubility of  $\text{H}_2\text{O}$  in rhyolite varies with pressure from a range of experimental work. Experiments focused on explosive volcanism have concentrated on the solubility of water at a temperature of  $850^\circ\text{C}$ , because this is a common eruption temperature for rhyolitic melts. Although water solubility increases as pressure increases, this relationship is not linear, especially at lower pressures, and the nature of the non-linearity is dependent on magma composition (see [86] for review). The root of the complexity of water solubility is that water chemically reacts with the silica glass matrix as a network modifier. This reduction in the average molecular weight is responsible for decreasing the viscosity (Fig. 1), as well as having a range of complex geochemical effects (see [12] for popular review).

### Nucleation and Diffusion

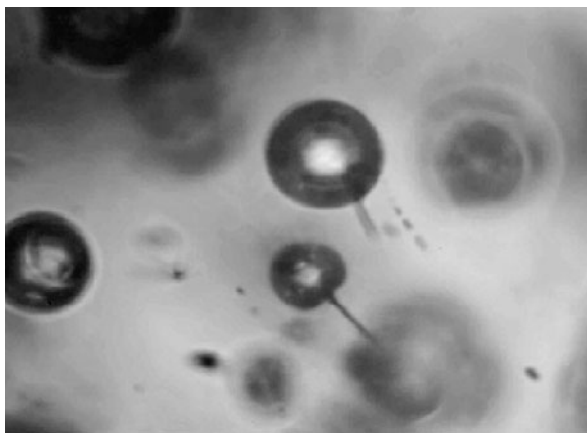
Decreasing the pressure applied to a volatile-saturated magma causes supersaturation of the volatile species. For explosive volcanism to occur, the volatile must make the transition from the dissolved state to a lower-density (and therefore higher volume) exsolved state, i. e., the process of bubble nucleation. Determining the degree of volatile supersaturation required for nucleation to take place has been the subject of significant laboratory effort. Magmas are complex chemical compounds that are rarely free of interfaces. Hurwitz and Navon [52] found experimentally that heterogeneous bubble nucleation dominated, with only small supersaturations of water vapor ( $< 1$  MPa, or  $\sim 50$  m head of magma) being required to nucleate bubbles at the surfaces of Fe-Ti oxide microlites. Other crystal-melt interfaces required supersaturations  $> 10$  MPa for nucleation of water bubbles. Even at supersaturations of  $> 130$  MPa, nucleation was heterogeneous even in apparently crystal-free melts.



Volcanic Eruptions, Explosive: Experimental Insights, Figure 3 **a** shows pressures and temperatures of experiments carried out on rhyolite melts to investigate water solubility as a function of pressure, see Liu et al. [73] for sources. The bulk of the experiments are carried out at  $850^\circ\text{C}$  because this is a common temperature for rhyolitic eruptions. **b** The solubility of water decreases with pressure, becoming increasingly non-linear below 10 MPa (shallower than about 50 m depth for lithostatic pressure). The transition, and consequent expansion, of water going from the dissolved to the vapor (or supercritical) states as pressure decreases is the driving mechanism of explosive volcanic eruptions. Reprinted with kind permission from Elsevier

More recently, Gardner [43] found experimentally that certain crystal interfaces or features acted as nucleation sites, in particular blocky shaped magnetite or the ends of needle-shaped hematite (Fig. 4). Gardner's [43] study, and others cited therein, indicate that bubble nucleation is a complex function of the composition and temperature of the melt, the composition and morphology of crystals, and the degree of supersaturation prevailing. Accurate modeling of bubble nucleation requires detailed knowledge of the melt as it becomes supersaturated. Nucleation could





Volcanic Eruptions, Explosive: Experimental Insights, Figure 4 As pressure decreases, the melt becomes supersaturated in water. During the decompression of small samples of hydrated rhyolite, bubbles of water vapor nucleated then grew at the ends of needle-shaped hematite crystals, but not along their lengths (scale bar = 20  $\mu\text{m}$ , Gardner [43]). Such ‘imperfections’ allow heterogeneous bubble nucleation and prevent large supersaturations developing. Reprinted with kind permission from Elsevier

occur at low supersaturations, with conditions never far from equilibrium. Conversely, in the absence of suitable nucleation sites, volatile supersaturations of the order of 100 MPa could develop as pressure in the melt decreases, and when nucleation does occur, rapid bubble growth can be expected.

Once bubbles of volatile have nucleated, growth is promoted by diffusion of volatiles from the supersaturated melt, and by expansion during pressure decrease. Any decrease in pressure also acts to increase volatile supersaturation (Fig. 3), encouraging either new bubble nucleation or, if bubbles are nearby, the diffusion of volatile into existing bubbles [17]. The diffusivity of volatiles is, therefore, a key parameter in determining the nature of the degassing process and has been reviewed by Watson [124]. Volatile diffusion into a bubble from the local melt surrounding it implies that the bubble will be surrounded by a shell of melt with relatively low volatile content and, thus, with elevated viscosity (Fig. 1a). This viscous melt shell may play an important role in the dynamics of bubble expansion during eruption, further adding to the importance of understanding diffusivity. Zhang and Behrens [131] experimentally investigated the diffusivity of water in rhyolitic melts, which, in combination with previous studies, covers the parameter range 400–1200°C, 0.1–810 MPa, and 0.1–7.7% w/w total  $\text{H}_2\text{O}$  content. The resulting empirical relationship between diffusivity of molecular water ( $D$ ) and pressure ( $P$ ), temperature ( $T$ ) and water concentration ( $X$ )

is given by

$$D_{\text{H}_2\text{O}_m} = \exp \left[ \left( 14.08 - \frac{13128}{T} - \frac{2.796P}{T} \right) + \left( -27.21 + \frac{36892}{T} + \frac{57.23P}{T} \right) X \right].$$

Such experimental determination of water diffusivity then forms key aspects of models of the behavior of silicate melts that undergo decompression, for instance, in how rapidly bubble pressure recovers after a perturbation [32], or in the nature of the resulting explosive activity [84].

### Permeability

Erupting magma usually includes gas bubbles, which can approach and interact with each other due to differences in their rise velocities within the melt and due to expansion. If the residence timescale of magma in the plumbing system is long compared to the time required for bubbles to move through it, then the bubbles can be considered to be separating from the melt. This is often the situation in low VEI eruption styles such as Strombolian activity. Here, gas bubbles can ascend through the low viscosity magma relatively rapidly and are able to collide and concentrate, possibly forming bubble rafts or foam layers. Bubble coalescence may then create fewer, but larger bubbles with significantly increased ascent velocity that, on reaching the surface, result in Strombolian activity.

Conversely, if magma residence timescale is short in comparison to bubble rise velocity, then bubble interactions cannot take place so readily. Under these conditions, the rapid depressurization and expansion of the trapped gas bubbles often results in high VEI events, but not universally. The ability of some such magmas to erupt relatively passively indicates that, as they ascend, they are able to lose gas by some other mechanism. One way for this to happen is for the vesiculated magma to become increasingly able to accommodate gas flow and separation as the volume fraction of bubbles (porosity) increases, allowing gas to move from bubble to bubble to the surface or into the surrounding rock. Such gas escape can be quantified using permeability, which is proportional to the effective pore diameter. Establishing how bubbly magmas become permeable, and the nature of that permeability is, therefore, of key importance to understanding eruption mechanisms.

The permeability of eruption products has only been measured experimentally in recent decades, and by few workers (e. g., [42,67,87,104]). Eichelberger et al. [42] determined the room-temperature gas permeability of rhyolite with between 37% v/v and 74% v/v porosity. Over this

range, permeability increased by over nearly four orders of magnitude from  $10^{-16}$  to  $10^{-12}$  m<sup>2</sup>, precipitating the hypothesis that high permeability can suppress explosive behavior in silicic volcanism, provided gas can escape. Klug and Cashman [67] made laboratory measurements on 73 pumice samples with porosities ranging from 30% v/v to 92% v/v. Permeability ranged from  $10^{-14}$  to  $10^{-12}$  m<sup>2</sup>, with the relationship that  $k \propto \phi^{3.5}$ , where  $k$  is permeability and  $\phi$  is porosity. Permeabilities at high porosity values agree between these two investigations, but disagreement at low porosities suggests greater complexity. Saar and Manga [104] demonstrated that there is not a straightforward relationship between permeability and the volume fraction of bubbles. Samples of basaltic scoria, which expand isotropically then cool rapidly, and therefore contain near-spherical bubbles, are well behaved in their permeability-porosity relationship and may be modeled by established percolation theory. However, many multiphase volcanic products do not have spherical bubbles because anisotropic flow causes high values of the capillary number. The elongate nature of the resulting bubbles causes high permeabilities at low porosities [104]. However, the permeability, like the porosity, is no longer isotropic, and gas escapes in directions determined by the local direction of flow as well as the prevailing pressure gradient. Laboratory measurements of volcanic products have been used to scale and validate theoretical models of magma permeability as a function of porosity. These models have, amongst others, been based on analogy with networks of resistors in electric circuits [16], and combination of classical approaches with fractal pore-space geometry [35]. Such models can then be incorporated into numerical simulations of flow in volcanic conduits that test the importance of permeability on explosive-effusive transitions.

### Consequences

The consequences of the materials properties of magmas responsible for explosive volcanic flows might now be explored. At pressures higher than the total volatile vapor-pressure, magma remains bubble-free. When pressure is reduced sufficiently below that of the volatile vapor pressure, then bubbles of volatile nucleate in the magma. These bubbles grow, and as bubble density is less than that of the magma, the bulk magma density decreases. This may act to increase magma ascent rate, creating a mechanism of positive feedback. Simultaneously, the reducing volatile content of the magma rapidly increases the liquid viscosity, reducing the strain rate required for viscoelastic and brittle behavior. This process occurs at the length-scale of diffusion into bubbles as well as within the liquid as a whole

and, depending on the timescales, crystal growth may also be changing, or indeed driving, the rheological evolution of the system.

These non-linear interactions underpin the complexity of understanding how magma flows, and illustrate the subsurface control on the nature of eruptions. In order to investigate how the properties of volcanic materials interact during the eruption process, experiments of a different nature are required.

### Volcanic Processes

Experiments investigating materials properties are mainly designed to measure parameters under static or steady conditions. However, explosive eruptions are not static and only pseudo-steady at best; therefore, different experiments are needed to explore bulk flow behaviors. Experiments investigating the explosive response of volcanic materials to possible eruption triggers are best conducted using natural silicates that have been well characterized by previous study. These experiments can be designed to minimize the inherent complexity of the overall process by concentrating on specific aspects. Models verified by experiments then contribute to larger models of the whole process.

### Explosive Processes

Magma is a hydrated silicate liquid (normally also containing some solid crystals) that, during the explosive eruption process, is converted into silicate fragments ranging in size from  $10^1$  to  $10^{-6}$  m. The mechanism responsible is fragmentation and understanding fragmentation processes is a major goal of physical volcanology.

Explosive volcanic activity predominantly requires the formation and growth of gas bubbles within magma. Exsolution processes are initiated at pressures generally  $> 10$  MPa, and may extend to pressures of several hundred MPa for water. Such pressures result from overburdens of several kilometers and flow experiments carried out at volcanic temperature and pressure (1100–1500K, up to at least 100 MPa) place severe constraints on containment vessels, measurement transducers and the volume of experimental samples. Another explosive mechanism occurs when volatile species mingle with hot magma at lower pressures resulting in hydrovolcanic activity. This is driven by expansion as liquid turns to gas on heating.

Direct observation of explosive processes is unlikely, but experiments mimicking volcanic conditions give valuable insight. Experimental exploration of explosive processes using natural samples requires that appropriate pro-

portions of gas, liquid and solid are present together with suitable pressures and temperatures.

**Hydrovolcanism** Experiments investigating the response of hot natural silicate melts to explosive ‘molten fuel-coolant interactions’ (MFCIs) with liquid water have been carried out by Wohletz [126] and Zimanowski et al. [134]. MFCIs are not confined to geological processes and are a dangerous industrial hazard, for example, on 4 November 1975, at the Queen Victoria Blast Furnace, Appleby–Frodingham Steelworks, Scunthorpe, UK, up to 90 tons of molten metal were thrown over a wide area after an MFCI explosion involving about 2 tons of water and 180 tons of molten metal. Berthoud [13] reviews MFCIs from the perspective of the nuclear power and other industries, but the physics described also applies to hydrovolcanism.

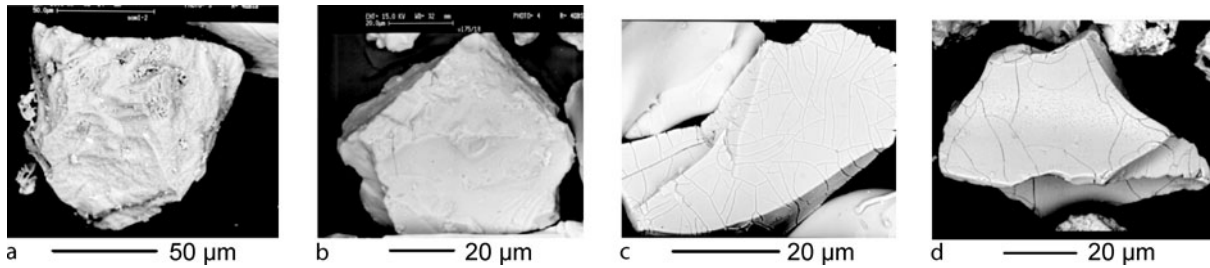
Volcano-specific research into MFCIs has studied the size and morphology of silicate particles from volcanic eruptions and compared them to those generated experimentally by MFCI processes (Wohletz [126]). The ratio of water to magma was found to be a key parameter in determining the nature of the particles resulting from the explosive process, with particle size reducing as explosivity increases. Such experiments provide insight into hydrovolcanic eruptions where magma may interact explosively with surface water (Surtseyan eruptions) or ground water (phreatomagmatic eruptions). Two processes dominantly produce the explosivity: (1) the rapid ‘flashing’ of liquid water to vapor; and (2) rapid cooling of the silicate melt causing thermal shock. Zimanowski et al. [135] studied how explosive fragmentation of a remelted volcanic rock could be interpreted from the size and shape of expelled silicate fragments. Two methods of fragmentation were used: (1) injection of air at 5 or 10 MPa into the base of the crucible containing the melt (at 1653 K); and (2) initiation of a MFCI using liquid water and a triggering perturbation. With an air injection at 5 MPa pressure, fine particles were uncommon and the particles generally rounded; however, the particle size distribution suggested that two unspecified fragmentation processes had operated. Three unspecified fragmentation modes were implied for the 10 MPa air injections, with finer, but still rounded, particles produced. Interaction with water (MFCI) also had three fragmentation mechanisms, but produced the finest particles, many of which were blocky. The MFCI experiments resulted in a much more complex fragmentation process than the air injection experiments (Zimanowski et al. [135]), with a key difference being the rapid cooling of the silicate melt by interaction with water. The combination of rapid cooling and high strain rates generated by

large pressures is considered to cause brittle failure of the melt and generate the fine, blocky fragments during the MFCI experiment.

Although not hydrovolcanism, the air-injection method of Zimanowski et al. [135] has been used by Taddeucci et al. [117] to experimentally investigate the role of crystal volume fraction during the explosive fragmentation of magma. This is of interest because ascending magmas are thought to have both axial and radial heterogeneity in their crystal contents, both before and during explosive activity. Preliminary results showed a strong correlation between the fragment morphology and the crystal content in both experimental and natural samples, with increasing crystallinity promoting brittle fracture. This suggests a number of interpretations: (1) a higher viscosity liquid phase naturally exists in more crystalline samples; and/or (2) strain-rates in the smaller liquid volume of high crystallinity samples are higher, i.e., there is less liquid to accommodate a similar motion; and/or (3) the macroscopic rheological behavior of the two-phase mixture promotes brittle processes [51].

Comparisons between experimental and natural particles generated by MFCI experiments [22] show that quite subtle surface features can be used to distinguish between events involving either partial or complete water vaporization (Fig. 5). Where vaporization is complete, both experimental and natural particles are blocky and equant in shape, and a brittle fragmentation process is implied from observations of stepped surfaces on the particles, created by mechanical etching. Particles resulting from incomplete vaporization of water show similar features, except that their surfaces host a network of cracks that are considered to be the result of rapid surface cooling (thermal shock), created by interaction with excess liquid water. Büttner et al. [22] comment on the experimental relevance of their research; “In general, the artificial pyroclasts produced in the experiments cover all grain sizes and shape properties of the natural ones, thus illustrating the quality of the experimental scaling in terms of geometry and energy release.” Such phenomenological similarities are a powerful indicator that experimental processes are mimicking the intended natural processes, and this is additional evidence of similarity. However, the experiments did not mimic all the post-explosion fragmentation processes that lead to the final natural product, demonstrating that care must be taken in comparing the outcomes of controlled experiments with the final product of the volcanic process.

Grunewald et al. [48] introduced a chemical aspect to MFCI experiments by using near-saturated NaCl solution as well as pure water, with the intention that the addi-



Volcanic Eruptions, Explosive: Experimental Insights, Figure 5

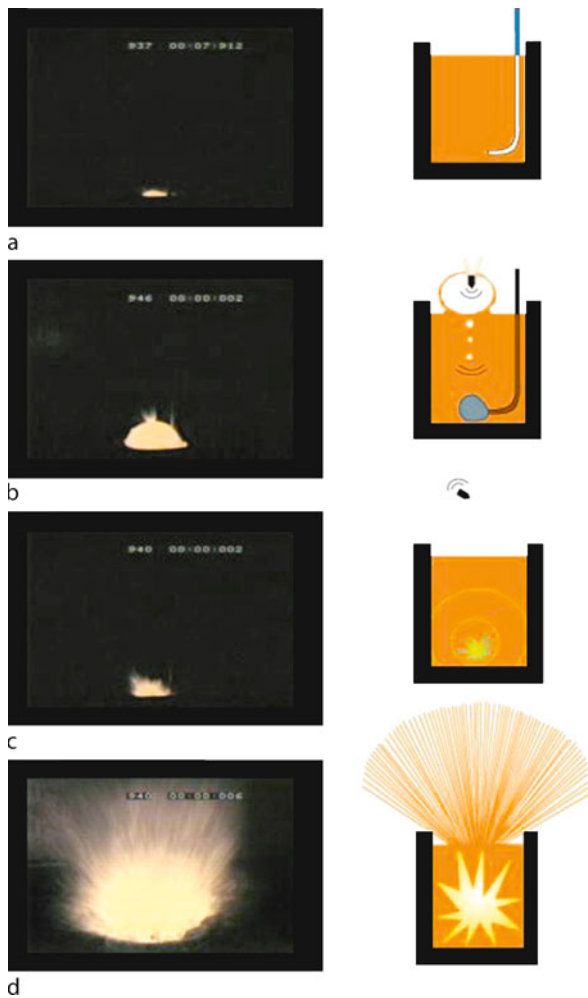
Magma and water may interact explosively during phreatomagmatic eruptions. Molten fuel-coolant interactions (MFCI) or magma-water interactions (MWI) can be carried out experimentally (Fig. 6). **a** shows a fragment generated by experimental MWI (Büttner et al. [22]) where all water was vaporized. **b** shows a fragment from a natural phreatomagmatic eruption that has not interacted with excess water. **c** shows an experimental fragment that 'erupted' through excess water, and **d** the product of phreatomagmatic activity thought to have interacted with excess water. The phenomenological similarities between **a** and **b**, and **c** and **d** provide powerful evidence that experiments are mimicking the processes that occurred during natural phreatomagmatic eruptions. Reprinted by kind permission from Macmillan Publishers Ltd: Nature (Büttner et al. [22]), copyright (1999)

tion of other species would add complexity. Experimental methods were similar to that of Zimanowski et al. [135] (Fig. 6), with the triggering of a rapid thermal interaction by the injection of hot molten silicate with cool liquid brine. It was found that NaOH and HCl were formed during the heating of the NaCl solution, and that explosion energies were lower than with pure water. The lower explosion energies resulted in more rounded and larger silicate fragments as found by Zimanowski et al. [135]. The reason for the reduced physical energy of fragmentation was attributed to energy coupling into the endothermic chemical process that produced NaOH and HCl, but the heating of brine may well also have introduced a liquid NaCl phase (1074 K–1738 K) that altered the physical MFCI process. These experiments raise the interesting prospect of being able to use the chemistry of volcanic gases and particle surfaces to make inferences about physical processes occurring during hydrovolcanism.

A different experimental approach to studying magma-water interactions (MWI) was used by Trigila et al. [121] in which experiments could be carried out at pressures up to 200 MPa, and temperatures up to 1473 K. In contrast to the method of Zimanowski et al. [135], where water is injected into hot (1650 K), liquid-dominated silicate melt at 0.1 MPa, Trigila et al. [121] 'infused' water into permeable sintered samples of solid-dominated basalt at lower temperature (1100 K), but high pressure (8 MPa). The samples were manufactured by grinding basalt to grain sizes of < 10 µm (powder) or about 0.7 mm (granular). These were heated to experimental temperatures and allowed to sinter for 30 minutes to form a permeable cylinder (about 2 cm in diameter and 6 cm in length) into which water was introduced. Only the more perme-

able granulated samples underwent spontaneous explosive behavior on mixing with water, generating particles similar in size to the original grains. These exploratory experiments indicate the importance of permeability to MFCI processes where groundwater could interact with hot, porous country rock as well as magma. We have already explored the importance of permeability in allowing exsolving volatiles to escape and reduce explosivity; under different circumstances we now we see that permeability can also increase explosivity.

**Exsolution** The nature of geological materials at pressures and temperatures existing within the crust and mantle has been the subject of considerable research by petrologists. Experimental petrology generally studies samples of material in small (millimeter-scale) capsules that are sealed, then pressurized up to tens of GPa and heated up to 2000 K; easily covering volcanically relevant conditions. Adaptation of this experimental technology to volcanic processes by Hurwitz and Navon [52] was used to study the response of hydrated rhyolite to decompressions up to 135 MPa (equivalent to about 5 km silicate overburden) at eruption temperatures (780°C to 850°C). Extended flow processes cannot be simulated with such small samples at these experimentally challenging conditions, but the pseudo-static response of hydrated silicate melt to pressure decreases of varying rate and magnitude provides important dynamic information about the processes of volatile exsolution (bubble nucleation and growth) and diffusion, as well as crystal growth. Hurwitz and Navon [52] revealed that nucleation of water bubbles (as a supercritical fluid or gas) in natural magmas is likely to be dominated by various heterogeneous processes operating over a range of supersaturations.



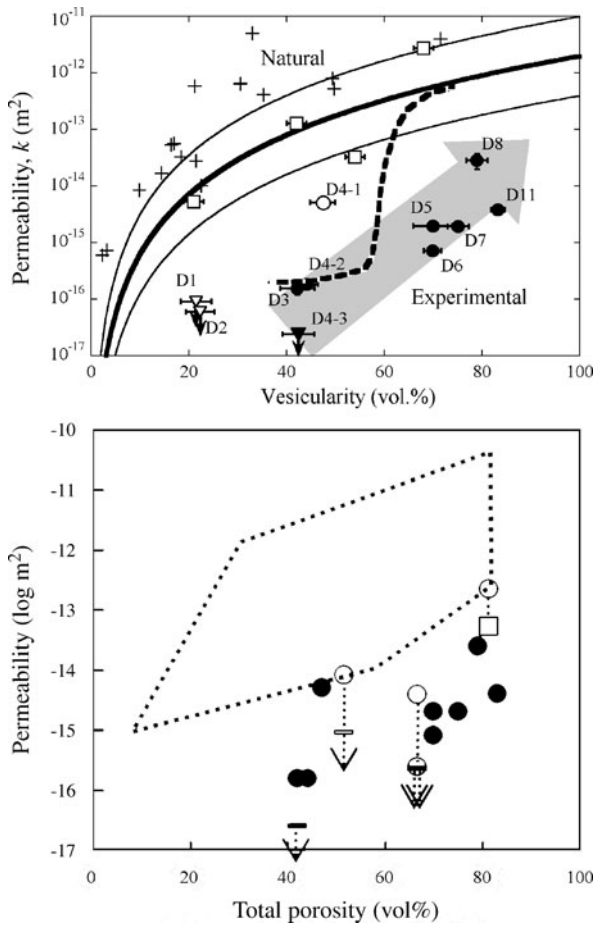
Volcanic Eruptions, Explosive: Experimental Insights, Figure 6 Schematic diagrams and frames from high-speed video illustrate one experimental method of simulating explosive phreatomagmatic eruptions (Grunewald et al. [48]). **a** Pre-experiment conditions with water injection tube in crucible of hot (1300°C) basalt melt. **b** 15 ml of water is injected, with the air expelled from the tube forming a bubble. The bubble was impacted by an air gun pellet to trigger **(c)** the explosive mixing of water and melt. **d** rapid transition of water from liquid to vapor results in pressure rise. The melt is fragmented by the subsequent rapid expansion. Reprinted with kind permission from Elsevier

Lyakhovsky et al. [76] carried out further analysis of the experimental results of Hurwitz and Navon [52] to reveal the controls on bubble growth following nucleation. Numerical modeling of the experimental data showed that diffusion of water from the melt into the bubbles controlled bubble growth at modest supersaturations in hydrated melts (3–6% w/w water, viscosity  $10^4$ – $10^6$  Pa s). The viscosity of the melt played a controlling role only

when bubble surface area was small in the initial stages ( $< 1$  s) of growth. At high supersaturations, the rapidly increasing area of bubble wall physically moves larger volumes of melt closer to the bubble and volatile advection is a significant contributor to bubble growth under certain conditions. As bubbles grow, the water content of the melt declines and the melt viscosity rapidly increases (Fig. 1a). Bubble growth then again becomes influenced by viscosity. These experimental results indicated that at water concentrations  $> 3\%$  w/w in rhyolite, equilibrium degassing was a good approximation. However, as water content declines then equilibrium is not maintained and melt some distance away from a bubble wall, and therefore out of range of significant diffusion, becomes supersaturated in water. Lyakhovsky et al. [76] postulated that these supersaturated melt pockets then encourage additional episodes of bubble nucleation; a process supported by Blower et al. [17] using analogue experimentation and computer modeling. Constrained rapid-decompression experiments provided key insights into the role of continuous nucleation in near-equilibrium degassing of high-viscosity, low-diffusivity melts, with supersaturation limited to that required by heterogeneous nucleation sites.

The role of high viscosities ( $10^7$ – $10^9$  Pa s) on bubble nucleation was studied by Gardner [43], using methods similar to those of Hurwitz and Navon [52]. Heterogeneous and continuous nucleation of water bubbles in a rhyolite with less than 1% v/v microlites of Fe–Ti oxides and plagioclase were again exhibited (Fig. 4). Increasing viscosity and decreasing supersaturation were found to increase bubble nucleation times from a few seconds to greater than 100 s. The controlling parameter for nucleation rate was found not to be temperature, diffusivity or viscosity, but the surface tensions of interfaces between melt, crystal and bubble.

All these experiments were fully sealed, with no opportunity for the exolving volatile phase to escape. This condition was relaxed by Burgisser and Gardner [19] who decompressed 3 mm diameter, 7 mm long cylinders of rhyolite hydrated within gold capsules at 150 MPa and 825°C, with some capsules modified to incorporate a sink for escaping water (pseudo-open conditions). Three bubble growth regimes were identified that depended on balances between diffusivity and viscosity in relation to the rate of sample decompression. The open system experiments showed that, given sufficient time (about 180 s here) water bubbles begin to coalesce once bubble volume fraction exceeds about 43%; note that the permeability of natural samples tends to decline rapidly below about 40% porosity. This is important because bubble coalescence promotes



an increase in permeability as preferential degassing pathways form in the melt. The limited volume of the sample, and its confining capsule, prevents the establishment of the range of flow behaviors that exist in volcanic conduits during explosive eruptions. However, the experiments of Burgisser and Gardner [19] provide useful insight into the role of flow on bubble coalescence for this very reason; features seen in the experimental samples are likely to have been difficult to distinguish in any experiments aimed at flow behavior. This demonstrates the usefulness of experiments for investigating spatially and temporally small (but tractable) aspects of a complex bigger picture. The results of such experiments can be used to verify models of the partial-process, with the facility to eventually combine these into a description of the full process.

The link between porosity and permeability was further explored by Takeuchi et al. [119] who isothermally decompressed 5 mm diameter, 15 mm long samples of Usu dacite in gold tubes from 150 MPa at 900°C, to pressures ranging from 50–10 MPa. Samples were then measured for

◀ Volcanic Eruptions, Explosive: Experimental Insights, Figure 7 As bubble volume fraction (vesicularity) increases, more bubbles coalesce to allow easy gas exchange between them. The result of this is increasing sample permeability with increasing vesicularity. Permeability allows gas to escape from the magma and reduces explosive potential, making permeability an important parameter in understanding the transition from effusive to explosive volcanism. Experiments carried out by Takeuchi et al. [119] (circles and squares with error bars, the solid and open symbols indicate isotropic and deformed vesicular textures of the products, respectively) are compared with measurement of natural samples in the top panel (dashed curve indicates the relationship of Eichelberger et al. [42], the bold solid curve with two fine satellite lines indicates the relationship of Klug and Cashman [67] with the upper and lower limit of the scattered data, the crosses and squares represent permeabilities of Melnik and Sparks [87], and Takeuchi et al. [119] respectively). Permeabilities have been corrected at a later date because of measurement artefact (bottom panel, Takeuchi et al. [118]). The trend of increasing permeability with vesicularity is demonstrated, but experiments generally show lower permeability than natural samples. One explanation for this could be that most of the natural samples have been subject to significant flow, whereas the experimental samples were generally not (open symbols show some evidence of flow). The role that flow plays in the development of permeability remains unclear. Reprinted with kind permission of AGU

permeability along the tube axis and, with their bubble fraction ranging from 22 to 83% v/v, covered the porosity range over which permeability has been measured in natural samples. However, the experimental samples showed significantly lower permeabilities than natural samples, except for the low porosity data of Eichelberger et al. [42], and for experiments where bubbles were elongated along the permeability measurement direction (Fig. 7). This suggests that flow processes act to increase permeability by 1–2 orders of magnitude, and that most of the natural samples measured have undergone significant flow; or as Takeuchi et al. [119] state “The quenched experimental products can be considered as snapshots in vesiculating magma during decompression without deformation of vesicular structure in the late stages of eruption”. Describing experiments as snapshots of a larger, more-complex process is an apt analogy. More recently, Takeuchi et al. [118] have revisited their experimental samples and, after removing the gold tubes, found that the sample permeabilities were further reduced (Fig. 7), especially at low vesicularity. The original measurements had included some gas leakage between the silicate sample and the gold tube and, when operating at the limits of permeability detection, the leakage added a significant systematic error to the measurements. The new results further emphasize the importance of the flow history of vesicular materials in the development of permeability, and illustrate the complex-

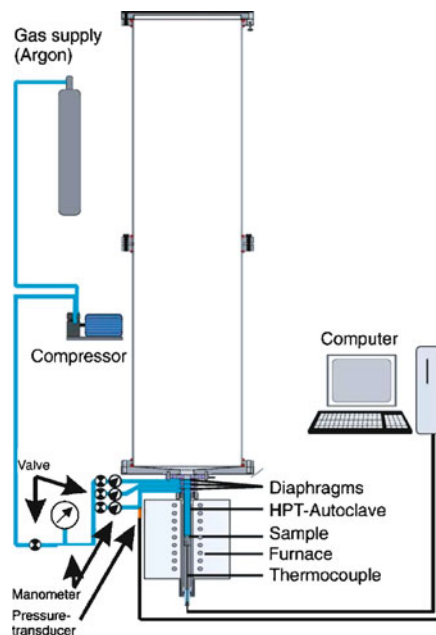
ity of interactions in the magmatic degassing process that determine the eruptive behavior.

**Decompression-Driven Flows** At Mt. St. Helens in 1980, a landslide acted to reduce the pressure applied to a body of vesiculated magma. Because the vesicles contained gas, the reduced pressure drove extensive magma expansion. This created the explosive lateral blast that preceded sustained explosive activity on May 18th 1980. Similar behavior may also be exhibited when portions of lava domes collapse. The response of vesiculated magma samples to relatively rapid decompression ( $10^{-3}$ – $10^1$  GPa  $s^{-1}$ ) (Alidibirov and Dingwell [3]) forms a major area of experimental volcanology.

Laboratory apparatus known as a ‘fragmentation bomb’ (Fig. 8) was conceived by Alidibirov and Dingwell [1,2] and has been developed by others since. This shock-tube apparatus rapidly decompresses experimental samples at 850°C from up to 35 MPa to 0.1 MPa (Spieler et al. [112]). The samples, cylinders of approximate dimensions of 2 cm diameter by 20 cm length, are natural volcanic rocks heated to the required temperature in an atmosphere of Ar within a pressure vessel. On reaching the temperature of interest, the sample is further pressurized

with Argon until a system of calibrated rupture discs fail. On failure, the sample rapidly decompresses ( $1$ – $100$  GPa  $s^{-1}$ ) and any explosive products freely expand into the collecting tank for further analysis. The use of Ar for these experiments removes any additional complexities that may arise from the nucleation, diffusion and growth of water bubbles allowing the detailed study of the mechanical response of vesiculated silicates to a rapid pressure drop.

Alidibirov and Dingwell [3] provide a review of the findings of experiments to that date. Samples of Mt. St. Helens Dacite from the May 18th 1980 eruption were decompressed by up to 18.5 MPa at temperatures up to 950°C. Observation of the resulting fragments showed angular material characteristic of brittle failure of the samples. Calculations of the strain rates imposed on the samples indicated that, even at the highest temperatures used, the dacite would behave as a brittle solid. Alidibirov and Dingwell [3] analyze three mechanisms that could explain explosive fragmentation of the sample. The dacite samples had interconnected pores and were, therefore, permeable. Rapid removal of the Ar gas above the sample creates a pressure gradient within the sample. If permeability is very high then the Ar gas can readily escape and pressure gradients will be relatively low. At the other ex-

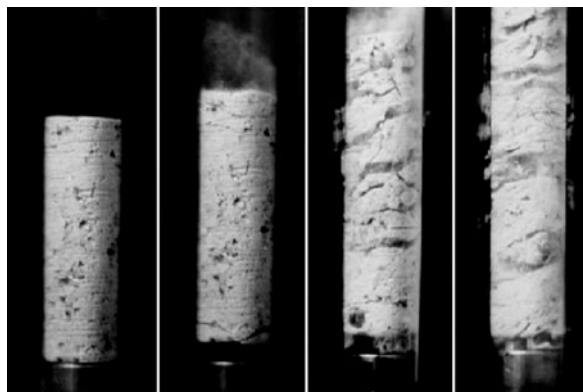


Volcanic Eruptions, Explosive: Experimental Insights, Figure 8

Schematic diagram and photograph of ‘fragmentation bomb’ shock-tube apparatus (Kueppers et al. [69]). Natural samples are pressurized up to 50 MPa with gas and heated up to 850°C in the autoclave. Samples can be used at room temperature with a transparent experimental section for high-speed imaging (Spieler et al. [111] (Fig. 9); Taddeucci et al. [116] (Fig. 22)). The diaphragms are ruptured to decompress the sample to atmosphere (0.1 MPa) and the resulting flow is collected in the large tank. Pressures, images and fragment morphology then provide information on flow dynamics. Reprinted with kind permission from Elsevier

treme, if bubbles are not interconnected then pressure gradients within the sample will be maximized (note, the experimental samples must have some permeability in order to pressurize them with Ar, but this does not need to be the case volcanically). In reality, the samples have a degree of permeability that allows pressure gradients to develop within the dacite. The pressure gradient was found to be at its highest at the fragmentation front, where the yield strength of the sample was exceeded. The fragmentation front was found to intrude the sample at tens of  $\text{m s}^{-1}$ , much slower than the sound speed and known as the fragmentation wave. In this instance, fragmentation can be described as a pressure-gradient driven change in flow regime from permeable filtration flow where only the gas moves, to two-phase flow where gas and solid are both in motion.

Martel et al. [82] synthesised 2 cm long by 2 cm diameter samples of hydrated haplogranitic glass with water contents ranging from 1.4 to 5.7% w/w. Water bubbles were nucleated and grown within the samples under a range of temperatures and pressures within the fragmentation bomb apparatus (Fig. 8) on timescales of about 1 hour, giving vesicularities ranging from zero to 0.91 v/v. Samples that developed large bubbles and high vesicularities underwent significant expansion and tended to develop elongated, or tube, bubbles during this process. Subsequent rapid depressurizations of between 5 and 18 MPa resulted in explosive fragmentation of the pre-foamed samples purely from expansion of the gas phase within the bubbles; timescales were too short for any significant further diffusion and exsolution of water from the glass. The resulting fragments were collected and analyzed for size, shape and vesicularity. The primary fragmentation process was found to be a sequential brittle spalling of the upper sample surface exposed to the decompression and was imaged by Spieler et al. [111] using Santorini pumice at room temperature in a transparent pressure vessel (reproduced here in Fig. 9). Low vesicularities and small decompressions were found to yield larger fragments than high vesicularities and large decompressions, a result consistent with the amount of energy released during the decompression process. The fragment size distributions may be dominated by the primary fragmentation mechanism (surface spalling), but impacts between fragments, and with the apparatus wall, may increase the proportion of the smallest fragments at the expense of the largest, especially with large decompressions (giving high fragment velocities). Interestingly, the presence of tube bubbles acted to decrease fragment size, demonstrating again that flow history is important in determining the products of explosive events.



Volcanic Eruptions, Explosive: Experimental Insights, Figure 9 The rapid decompression of a sample of Santorini pumice, at room temperature, by at least 9 MPa (3 times higher than at 850°C) in the ‘fragmentation bomb’ (Fig. 8) results in spalling fragmentation (Spieler et al. [111]). The *first frame* shows the 50-mm length by 20-mm diameter sample just prior to decompression. 0.25 ms after diaphragm rupture (*second frame*) the ejection of fine dust from the sample surface indicates that pressure is falling in the *upper reaches* of the sample and gas is starting to escape from the pores. The combined elastic response of the sample and apparatus has detached the sample from the holder, fracturing the sample base. 0.5 ms after decompression (*third frame*), spalling fragmentation propagates through the sample perpendicular to the axis of decompression. Fragmentation results from the build up of a pressure gradient within gas in the upper sample sufficient to exceed the tensile strength of the bubbly solid. After 0.75 ms (*fourth frame*) primary fragmentation is complete and fragments are ejected into the large tank (Fig. 8). The timescale of this transient flow is of order 1 ms, but such visualizations give great insight into possible volcanological fragmentation mechanisms. Reproduced with kind permission from Springer Science and Business Media

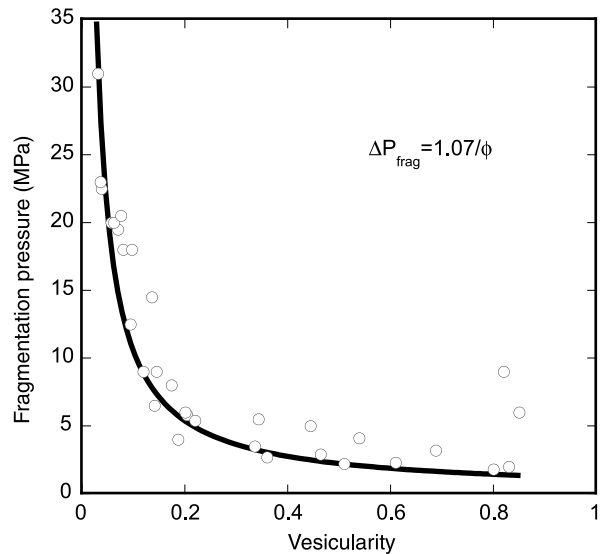
In a refinement of these experiments, Martel et al. [83] investigated the effect of adding relatively inert alumina particles to the hydrated haplogranitic samples as an analogue to crystals in a silicate melt. Solid particles influenced the final particle size distribution because they tended to remain intact during the experimental fragmentation process. An increasing proportion of alumina particles acted to increase the fragment size for a given decompression, suggesting that energy release was reducing. This would be consistent with the reduced proportion of gas phase that results from an increasing proportion of solids. This effect may also account for the observed increase in the decompression needed to fragment samples as the proportion of crystals increased, with the limit that a zero-vesicularity sample (100% crystals) would not fragment under decompression. Interestingly, tube bubbles were not formed in these experiments due to enhanced heterogeneous bubble nucleation acting to reduce bubble diameter.



Spieler et al. [112] carried out a systematic investigation of the rapid pressure drop required to cause the brittle spalling fragmentation of natural samples with a wide range of chemistry, porosity, permeability and crystallinity. The major control on fragmentation threshold pressure ( $\Delta P_{\text{frag}}$ ) was found to be the sample porosity ( $\phi$ ). Figure 10 shows the relationship to be approximated by  $\Delta P_{\text{frag}} = 1/\phi$ , with pressure given in MPa. Such a relationship makes qualitative sense because, as vesicularity tends to zero,  $\Delta P_{\text{frag}}$  will tend to a large value because of little or no gas phase to drive the fragmentation. At large vesicularities, there will be little melt to resist the expansion of the volatile and fragmentation will take place with small decompressions. Figure 10 suggests that the rapid removal of order 100 m of overburden from a vesiculated magma with  $\phi > 0.2$  will result in explosive fragmentation caused by pressure-retention in the pores. Such an event could happen by slope failure, for example, at Mt. St. Helens in 1980, or by a dome collapse. At lower vesicularities, the rapid removal of an overburden thickness of order 1000 m is required to initiate explosive fragmentation. However, although smaller overburden removals may not result in immediate spalling fragmentation, slower diffusively driven bubble growth will act to expand the magma and could thus trigger more sustained explosive activity.

Another mechanism of increasing pore pressure in magma is by crystallization of the silicate melt (e.g., Sparks [108]). This occurs because water is not incorporated into the structure of growing crystals and concentrates in the declining proportion of melt. Taddeucci et al. [115] used the ‘fragmentation bomb’ to ascertain the fragmentation threshold of crystal-rich magma from the 2001 basalt erupted at Mt. Etna, as a function of porosity. This was combined with models of pressures generated by crystal growth, and calculations of conduit pressures required to eject magma blocks observed during the explosions. This approach supported an explosion model sourced in the heterogeneous fragmentation of crystallizing magma plugs. The particularly interesting aspect of this model is the introduction of porosity heterogeneity within the magma. Figure 10 demonstrates that as porosity increases, the pressure change required for fragmentation decreases. Therefore, the competence of a body of magma will depend on its weakest link, namely any regions of high porosity. Once these begin to fragment, then the less vesicular regions will be removed as large blocks.

Although volcanically accurate materials were used in fragmentation bomb experiments, the effects of reductions in length and timescale by three orders of magnitude remain unknown. Nevertheless, these important experiments give specific insight into key mechanisms that



Volcanic Eruptions, Explosive: Experimental Insights, Figure 10 Samples of natural volcanic material with a range of vesicularities were heated to 850°C and decompressed in the ‘fragmentation bomb’ (Fig. 8) by Spieler et al. [112]. The pressure drop required to cause explosive brittle fragmentation was inversely related to the sample vesicularity, with the constant of proportionality (here, about 1.07 MPa) being the tensile strength of the solid phase. Intriguingly, this suggests that heterogeneity in vesicularity will result in selective fragmentation of more vesicular regions of magma, particularly at vesicularities below 0.3. Data replotted from Spieler et al. [112]

could operate during the rapid decompression (or internal pressurization) of a vesiculated body of magma, and general insight into the behavior of volcanic materials undergoing explosive eruption.

### Explosion Products

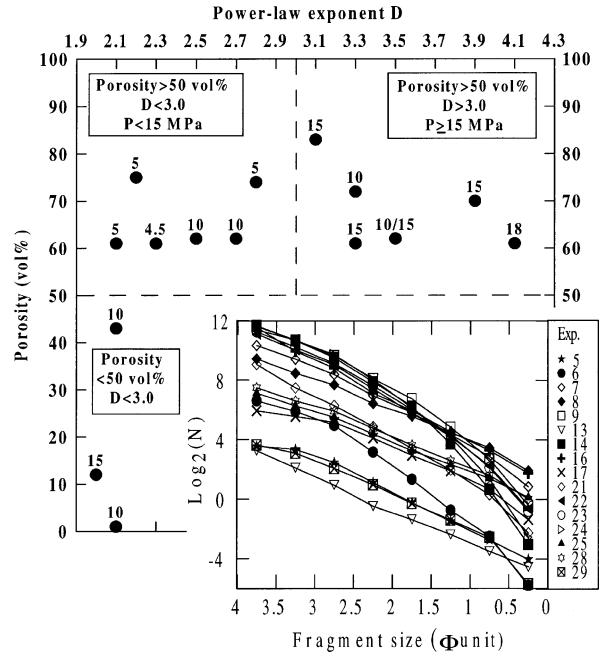
Explosive expansion of magma, driven by the low bulk modulus of gas-phase volatile species, generates a range of consequent phenomena. The primary magma fragmentation (one mechanism is illustrated in Fig. 6, and another in Fig. 9) involves a large increase in surface area within the volcanic system. The mechanism of this increase is the breaking of molecular bonds.

On exiting the volcanic vent, hot silicate fragments and volatile gas (mainly water vapor on Earth) interact with any atmosphere and cool. In the absence of a significant atmosphere, for example on Io, the volcanic ejecta expands, then the silicate fragments follow ballistic trajectories once they have decoupled from any volcanic gases and cool by radiation. Interaction between silicate fragments is

now unlikely and their trajectories end when they impact Io's surface; some may escape Io altogether. On Earth, the presence of an appreciable atmosphere results in different behavior. Explosion products emerge from the vent and interact with the atmosphere. If the silicate fragments are large and the explosive power low then fragment trajectories tend towards ballistic control. Conversely, if eruption power is high and fragments small, then atmospheric interaction may dominate. The mixing of air with hot eruption material and its consequent expansion may result in a thermally buoyant plume that ascends many kilometers into the atmosphere and may circumnavigate the globe by long-range atmospheric transport. If insufficient atmospheric air is entrained in order to develop and maintain buoyancy then some ejecta may fall back to ground to form a pyroclastic flow. On a smaller scale, the presence of electrostatic fields and liquid phases encourages sub-millimeter silicate fragments to aggregate. Many experimental studies have been carried out on the processes that follow explosive volcanic events and we review a selection here.

**Fragment Size Distributions** The fragment size distribution resulting from explosive volcanic eruption has been found to follow a power law distribution,  $N = \lambda r^{-D}$ , where  $N$  is the number of fragments greater or equal in size to  $r$ , and  $\lambda$  is a constant of proportionality. For a single or primary fragmentation event the exponent ( $D$ ) is predicted to be less than 3.0. Experiments carried out by Kaminski and Jaupart [66] impacted a sample of Minoan pumice with a solid piston to give an exponent of 2.7. James et al. [57] impacted two pumices together and found an exponent of 3.5. Pyroclastic flow or fall deposits from explosive volcanic eruptions exhibit an exponent range between 2.9 and 3.9.

Martel et al. [82] rapidly decompressed hot vesiculated rhyolite in a 'fragmentation bomb'. Decompressions of less than 15 MPa produced exponents of less than 3.0 for the resulting fragment size distribution, regardless of initial sample vesicularity. However, decompressions of > 15 MPa, for samples of vesicularity > 0.5, resulted in exponents > 3.0 and ranging up to 4.1 (Fig. 11). These results suggest that secondary fragmentation mechanisms, such as those induced by fragment-wall and fragment-fragment collision, could be occurring within the apparatus after the primary spalling-fragmentation event (Fig. 9). Alternatively, as high exponents were observed at high vesicularities, the highly heterogeneous nature of a foamed silicate could invalidate the assumptions made in suggesting 3.0 as the exponent boundary for primary fragmentation, at least for fragment size ranges smaller than the order of bubble size. Interpretation of fragment size distribu-



Volcanic Eruptions, Explosive: Experimental Insights, Figure 11 Hot samples of bubbly rhyolite glass with a range of vesicularities were decompressed through various pressures in the 'fragmentation bomb' (Fig. 8), and the fragment size distribution (FSD) measured (Martel et al. [82]). The FSD was compared with a power-law distribution given by  $N = \lambda r^{-D}$ , where  $N$  is the number of fragments with radii larger than  $r$ . The power-law exponent,  $D$ , was found to increase with the magnitude of decompression (numbers next to data points in MPa), as well as with higher vesicularity. Compare this to an exponent of 3.5 for colliding pumices in Fig. 12. Theoretically,  $D > 3$  suggests more than one fragmentation process, but these are complex, heterogeneous, multiphase materials and established fragmentation theory based on dense homogeneous materials may not give the full picture. Reprinted with kind permission from Elsevier

tions between experimental and natural systems in terms of fragmentation processes is further complicated by difficulties in obtaining complete natural samples from explosive events. For example, winnowing in the atmosphere separates fragments as a function of fall velocity, which is a function of size and shape; therefore, the products of any individual fragmentation process are subsequently separated.

**Fall Velocity of Silicate Fragments** Understanding the behavior of silicate fragments in the atmosphere requires knowledge of the fall velocities of real volcanic pyroclasts, which rarely conform to an idealized spherical shape. Walker et al. [123] dropped samples of pumice from heights up to 30 m and found that terminal fall velocities were best approximated by the fall equation for cylindri-

cal bodies. Wilson and Huang [125] carried out extensive experiments on silicate crystals and glass fragments in the size range 30 to 500  $\mu\text{m}$ . A fragment shape factor ( $F = (b + c)/2a$ ) was defined in terms of the longest ( $a$ ), intermediate ( $b$ ) and shortest ( $c$ ) principal axes of the fragment, allowing definition of the empirical formula  $Cd = (24/\text{Re})F^{-0.828} + 2(1.07 - F)^{0.5}$ , where the fragment drag coefficient is given by  $Cd$ , as a function of Reynolds number,  $\text{Re}$ , and the shape factor. Suzuki [114] revisited the experimental data and proposed a modified equation,  $Cd = (24/\text{Re})F^{-0.32} + 2(1.07 - F)^{0.5}$ . This empirical, experimentally informed approach allowed calculation of the fall velocity of a fragment in the atmosphere and provided input into numerical models of the dispersal of products from explosive volcanism. Further information can be found in [101,109,120].

**Fragment Electrification** James et al. [56] investigated possible mechanisms by which silicate fragments could become electrically charged during explosive activity. Two natural pumices were collided together resulting in the generation of particles generally smaller than 70  $\mu\text{m}$  in diameter as the silicate foam fragmented in brittle fashion. Fracto-emission generated ions and charged silicate fragments. Ions were found to have one net charge and silicate fragments the opposite net charge. Experiments carried out at atmospheric pressure produced charge densities similar to those found on ash particles resulting from explosive terrestrial volcanism, whilst those carried out at 0.1 Pa sustained at least an order of magnitude more charge. The absolute amount of charge was dependent on the impact energy, (a result comparable to [23] findings of increasing electric field with greater fragmentation energy), with the net charge representing a slight imbalance between positively and negatively charged fragments. The mechanism of producing charged silicate fragments from pumice-pumice collision adequately accounts for the fragment charges and electric fields measured in proximity to explosive volcanic eruptions. This suggests that pumice-pumice collision, i. e., secondary fragmentation, could be a major source mechanism for sub-100  $\mu\text{m}$  silicate fragments.

Experimental hydrovolcanic explosions were found to produce an electrically charged fragment and ion cloud [24]. Interestingly, fragmentation using high-pressure Argon gas produced electrical effects of smaller magnitude than magma-water interaction, even though expansion rates of the ejecta from the experimental crucible were similar. This suggests that the cooling effect of water does produce much more brittle failure resulting in enhanced fracto-emission [37,40], consistent with the production of

smaller blocky fragments from MFCI experiments [135]. It is also possible that polar water molecules dissociated into reactive ions during the MFCI process, a mechanism not so accessible to Argon atoms. Büttner et al. [24] found striking similarities between experimental electric fields and those generated during explosive events at Stromboli volcano.

**Silicate Fragments and Their Aggregation** Accretionary lapilli are widespread in the deposits from explosive volcanic eruptions, and anomalous variation in the deposit thickness and grain size distribution as a function of distance from the explosive vent is commonly observed. Both of these processes are explainable if silicate fragments smaller than about 100  $\mu\text{m}$  clump together, or aggregate, to form a larger entity which has different aerodynamic properties than its constituent particles. Sparks et al. (see Chapter 16 in [109]) provides a review of particle aggregation in volcanic plumes prior to 1997.

Accretionary lapilli are obvious in the geologic record as millimeter to centimeter sized spheroidal aggregates (density range 1200 to 1600  $\text{kg m}^{-3}$ ) of silicate fragments and secondary minerals such as gypsum, and they formed the early focus of study into aggregation in volcanic plumes. Two complementary experimental studies provided insight into the formation mechanisms of accretionary lapilli. Gilbert and Lane [45] focused on formation in atmospheric plumes. A wind tunnel was used to demonstrate experimentally that particles of volcanic ash collide and adhere to the surfaces of objects falling through the atmosphere that are covered with a thin layer of liquid. The liquid layer was found to act as a control on the sizes of silicate fragments that adhere, with high sticking efficiencies at 10  $\mu\text{m}$  diameter, reducing by two orders of magnitude for fragments 100  $\mu\text{m}$  in diameter. A key experimental finding was the importance of hygroscopic chemical species in maintaining thin liquid films. Experimentally this was achieved using NaCl; volcanically, sulphuric acid is the most likely hygroscopic compound that could maintain liquid films at low relative humidity and down to temperatures of  $-70^\circ\text{C}$ . These experiments strongly suggest that accretionary lapilli are diagnostic of a three-phase environment within the volcanic plume in which they formed, supporting earlier hypotheses (e. g., [88]).

Schumacher and Schmincke [105] focused on accretionary lapilli formation in pyroclastic flow processes. Experiments were carried out providing insight into interpreting the structure of accretionary lapilli in terms of their formation mechanisms. Accretionary lapilli were synthesized by a method similar to the industrial production of fertilizer pellets, with mixtures of volcanic ash and

variable proportions of water being rotated in a steel pan. The resulting spheroidal agglomerates varied with the proportion of water present, formation being optimal at between 15 and 25% w/w water, and grain size sorting was found to occur at low water mass fractions. The mechanism of formation, namely capillary binding, is identical to that of Gilbert and Lane [45], but under conditions of much higher fragment number-concentration.

The occurrence of secondary thickness maxima and bimodality of grain size distribution, exemplified by Carey and Sigurdsson [27], implies the mechanism of silicate fragments being transported within aggregates. These aggregates could be accretionary lapilli, but these are often absent from the deposits suggesting that the aggregates are fragile and lose their identities once incorporated into a volcanic deposit. Fragile aggregates have been observed falling from volcanic plumes and one explanation for their formation is the presence of electrostatic charge. During fragmentation, silicate particles become electrostatically charged (see Sect. “Fragment Electrification”), and whilst in the atmosphere these charges have no path to escape to earth unless electric fields exceed the level required for electrostatic discharge in air. This leads to long-lived attractive electrostatic forces between silicate fragments [81], and consequently the formation of aggregates. Once on the ground an electrically conductive path to earth can be established, especially in the presence of moisture, and the aggregates disintegrate.

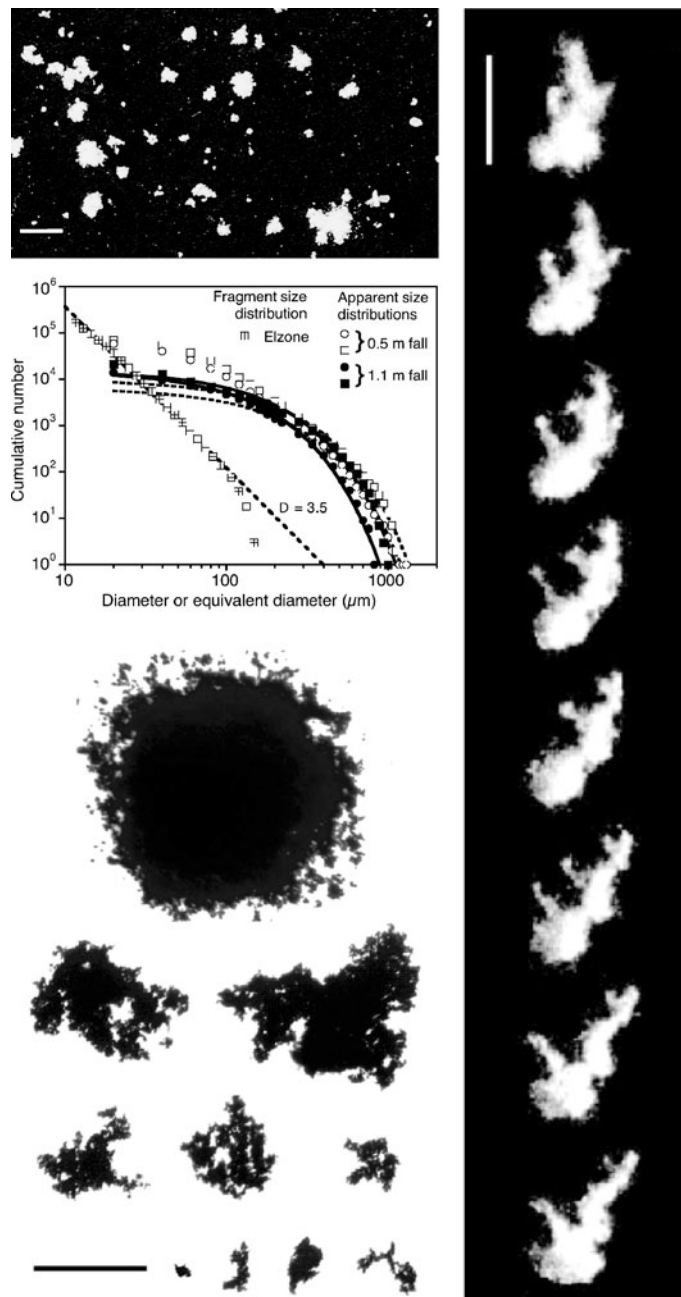
Schumacher and Schmincke [105] carried out experiments using natural volcanic ash and a stream of ions from an electrostatic paint gun. Silicate fragments smaller than about 180  $\mu\text{m}$  diameter were found to rapidly aggregate into loose fragile clusters demonstrating the viability of the electrostatic aggregation process for fine volcanic particles.

James et al. [57] studied the generation of aggregates resulting from the interaction of ions and charged silicate fragments using the fragment generation and charging mechanism of James et al. [56] with pumice from the May 1980, Mount St. Helens fall deposit. Silicate fragments were allowed to fall and interact over a distance of 0.5 to 1.1 m onto a variety of analysis platforms. Fallout was collected and analyzed for fragment size (Fig. 12) using an agitated dispersion of the particles in a conducting electrolyte to maintain disaggregation. The particle size distribution showed power law behavior with exponent 3.5 (Fig. 12), somewhat above the theoretical maximum of 3.0 for a primary fragmentation event. These experiments indicated that the aggregation process happens shortly after the fragmentation event has generated charged silicate fragments and ions (Fig. 12). The aggregation process significantly changes the size distribution

relevant to aerodynamic behavior (Fig. 12) from the single-fragment form  $N = \lambda r^{-3.5}$  to aggregate form  $N = 15\,089 \exp[-0.011d']$ , where  $d'$  is the aggregate diameter. Experimental measurement of aggregate fall velocities enabled estimation of aggregate densities between 80 and 200  $\text{kg m}^{-3}$ , at least an order of magnitude less than both the fragments that comprise aggregates and accretionary lapilli. The main consequence of electrostatic aggregation for sub-100  $\mu\text{m}$  silicate fragments is to reduce atmospheric residence time because aggregation acts to increase fall velocity above the single fragment value. The size distribution of individual fragments within an aggregate was experimentally examined by James et al. [58], who found empirically that for aggregate diameters less than 140  $\mu\text{m}$ ,  $n = 0.000272d_a^3[\exp(-0.22d_p) + 0.008 \exp y(-0.083d_p)]$ , where  $n$  is the number of fragments larger than diameter  $d_p$  within an aggregate of diameter  $d_a$ . As electrostatic aggregates exceed about 140  $\mu\text{m}$  in diameter the size distribution of their constituent fragments was found to change. This was attributed to the aggregate growth mechanism becoming dominated by aggregate-aggregate interaction rather than aggregate-fragment interaction (Fig. 12), and stabilization of the size distribution of material being incorporated into the aggregate. Electrostatic aggregates can be represented as spheres of density about 200  $\text{kg m}^{-3}$  and an experimentally determined empirical relationship between drag coefficient ( $Cd$ ) and Reynolds number ( $Re$ ) for aggregate diameter range 50 to 500  $\mu\text{m}$  is  $Cd_{\text{agg}} = 23Re^{-0.637}$  (James et al. [58]). Such experimental data can then be used within numerical models of the transport of sub-100  $\mu\text{m}$  volcanic ejecta in the atmosphere, although at the time of writing this has yet to be undertaken.

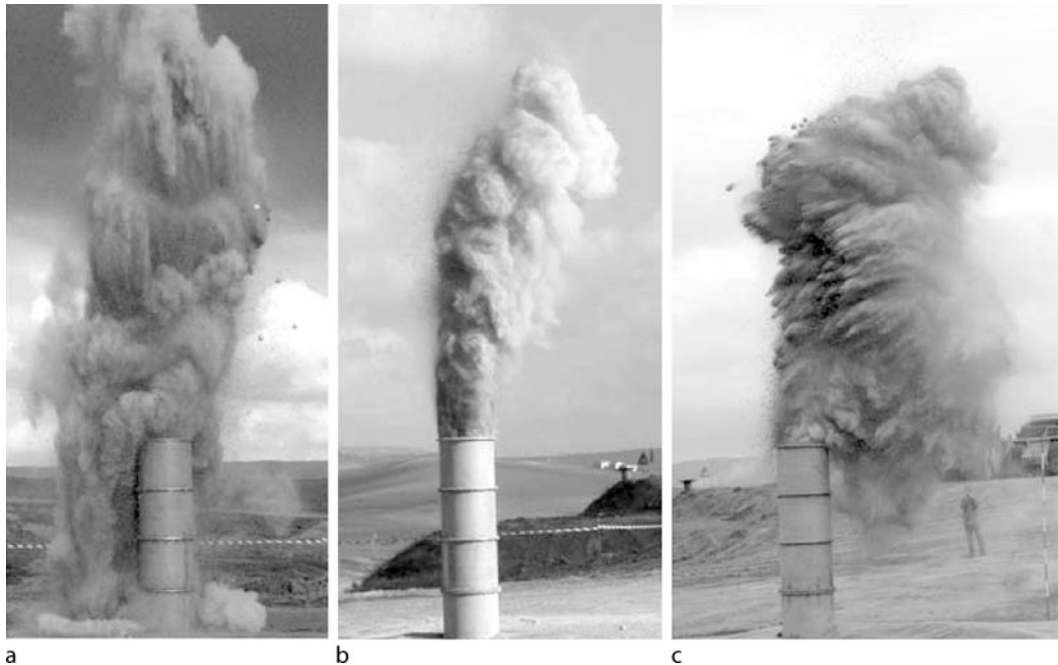
**Pyroclastic Flows** Pyroclastic flows comprise a hot mixture of silicate particles and gas formed by the collapse of negatively buoyant regions of a volcanic plume or avalanching of material from a growing lava dome. By nature, pyroclastic flows are destructive to infrastructure and almost invariably fatal to plants and animals. Insight into the behavior of pyroclastic flows may be gained by studying their deposits. However, it is difficult to definitively identify fluid dynamic processes operating during a pyroclastic flow from the complex structures within deposits that are revealed once motion has ceased. Here, experiments can help in isolating flow processes and their subsequent effect on deposits. However, scaling between experiment and natural process requires care to ensure behavioral similarity as particles of a range of sizes interact with the gas phase to different degrees.

Dellino et al. [36] designed experiments to investigate the mechanics of pyroclastic flows on a scale com-



Volcanic Eruptions, Explosive: Experimental Insights, Figure 12

Fragmentation during explosive eruptions produces electrically charged fragments and ions by fracto-emission processes (James et al. [57,58]). Electrostatic forces are significant in comparison to gravity for fragments less than about 100  $\mu\text{m}$  in size, resulting in fragment aggregation. Aggregates have different aerodynamic properties than their constituent fragments, thus altering the fallout behavior of sub-100  $\mu\text{m}$  volcanic ash and consequently its atmospheric transport. The top image shows experimental aggregates collected on a plate (scale 1 mm), with the graph showing the size distribution of the aggregates (*curved lines*), as well as that of the constituent fragments (*straight line*). The *bottom panel* (scale 0.5 mm) shows that experimental aggregates of different sizes have different morphologies, and the *right-hand panel* (scale 0.5 mm) shows strobe images of a falling and rotating experimental aggregate. Reprinted with kind permission of AGU



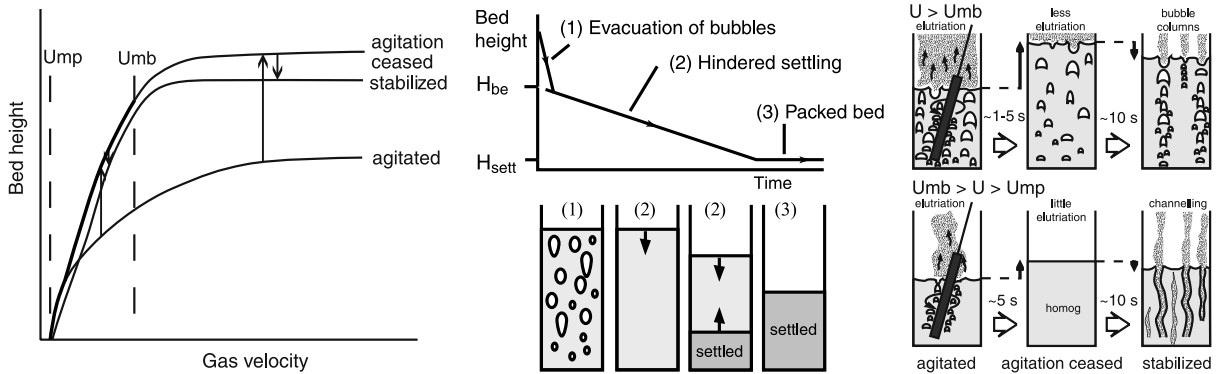
Volcanic Eruptions, Explosive: Experimental Insights, Figure 13

Ejection of natural pyroclastic material by a burst of gas can be used to study both volcanic plumes and pyroclastic flows (Dellino et al. [36]), this also looks enormous fun! a shows an experiment with low specific mechanical energy (SME) generating a collapsing column and (pyro)clastic flow. c shows a high SME experiment generates a plume that, whilst not buoyant, disperses on the prevailing wind, and b shows the transitional case. Such large-scale investigation helps to overcome some of the scaling difficulties imposed by smaller-scale laboratory experiments. Reprinted with kind permission of AGU

parable to the natural process. In order to approach this, the physical scale of the experiments was such that they required an outdoor laboratory. The experimental material was sourced from a natural pyroclastic flow deposit in order to mimic the interaction between fragments and gas as closely as possible. Varying masses of this material were packed into an experimental tube 60 cm in diameter and 2.2 m high. 14 liters of gas at pressures over  $10^7$  Pa was rapidly injected into the base of the tube to eject the experimental material, with experiments triggered and logged under computer control. The experiments were carried out at ambient temperature and were, therefore, not scaled for any thermal effects; this has the benefit of reducing both equipment and process complexity. Ejected material could be sampled for analysis and compared to natural pyroclastic deposits. Preliminary experiments showed two main types of behavior as the specific mechanical energy (SME, given by  $\text{gas pressure} \times \text{gas volume} \div \text{mass of pyroclastic material}$ ) of the system was changed (Fig. 13). At high SME values a dilute particle plume develops with the particles settling out individually. At low SME values the ejected material forms a collapsing column that generates a particle flow similar to a py-

roclastic flow. Intermediate SME values give transitional behavior with aspects of both end-members. These experimental observations have considerable phenomenological similarity to natural explosive eruptions where particle-dilute regions ascend in the atmosphere and more particle-concentrated regions fall under gravity to give pyroclastic flows. Such similarity gives added confidence that mechanisms operating in the natural process also operate in the simulation.

Experiments producing a collapsing-column deposited material at the point of impact of the column with the ground. The resulting deposit was found to be structureless, just as many natural proximal pyroclastic deposits, because grain-sorting processes did not operate in this concentrated particle flow. Other similarities included the formation of a concentrated undercurrent, bed load at the base of the turbulent flow, and continuous atmospheric suspension of small silicate particles; this is very interesting in the absence of thermal processes. These preliminary experiments demonstrate great potential for linking processes in scaled experimental flows with the resulting experimental deposits and, therefore, providing a powerful interpretation tool for natural pyroclastic flows.



Volcanic Eruptions, Explosive: Experimental Insights, Figure 14

The inflation and deflation of a bed of natural pyroclastic material using hot gas shows a range of flow structures (Druitt et al. [41]). During inflation, the bed can be homogeneous or develop gas bubbles or channels, and fine particles can be removed by elutriation. When fluidizing gas velocity exceeds  $U_{mp}$  (left panel), the bed inflates and particles start to segregate on the basis of size and density. Bubbling occurs when gas velocity exceeds  $U_{mb}$  (upper right panel), and the bed expands only weakly with increasing gas velocity (left panel). Agitation of the flow acts to reduce the gas volume fraction. Once the gas supply is closed off then bubble collapse results in rapid deflation ((1) in middle panel). An extended period of hindered settling then generates a layer of settled material at the base of the bed ((2–3) in middle panel). These experiments were used to investigate the gas retention properties of pyroclastic materials as a function of fragment size and temperature. Gas retention was favored by finer fragment sizes, with high temperature having a secondary retention effect. Hot, fine-grained pyroclastic flows are, therefore, predicted to flow further. Reproduced with kind permission from Springer Science and Business Media

Experiments carried out at smaller scale, but which included heating of the volcanic ash [41], have also investigated gas retention in pyroclastic flows. A bed of pyroclastic flow material with a wide fragment size range was fluidized by the drag force of rising hot gas. Both the bed expansion and collapse (degassing) after the gas supply was cut were studied. During expansion, smaller fragment size, lower density and higher temperature all promoted more uniform and easier expansion of the fluidizing bed. Figure 14 illustrates the nature of flow patterns that develop. Collapse experiments were imaged using X-rays and carried out between  $20^{\circ}\text{C}$  and  $550^{\circ}\text{C}$  giving temperatures comparable to those in natural flows. The generic collapse process is illustrated in Fig. 14. Experiments and numerical models identified that an aerated bed loses its gas by permeable flow on a diffusive timescale. A fluidized and uniformly expanded bed was controlled by timescales of hindered settling and diffusive degassing of the resulting sediment layer. The relative magnitudes of these timescales depended on the thickness of the degassing bed, with settling dominating in thin beds and diffusion in thick ones. These experiments indicate that hot pyroclastic flows with small fragment sizes retain gas longer than cool, large fragment flows, with temperature having a secondary effect. The gas retention translates into the distance a flow can travel and these experimental results are consistent with observations of natural pyroclastic flows.

## Analogue Approach

Experiments with volcanic materials can suffer constraints placed by the high temperatures and pressures at which the processes of interest occur. If similar processes can be studied under more amenable environments, then the practical limitations are relaxed and, generally, fluid volumes can be increased and a greater range of transducers are available for measurement. This can be achieved by substituting natural fluids with analogues that can be studied at near room temperatures and pressures. However, one difficulty with this approach is in ensuring the relevance of analogue results to volcanic systems and, perversely, natural volcanic materials are now often characterized more fully than many of the analogue fluids used.

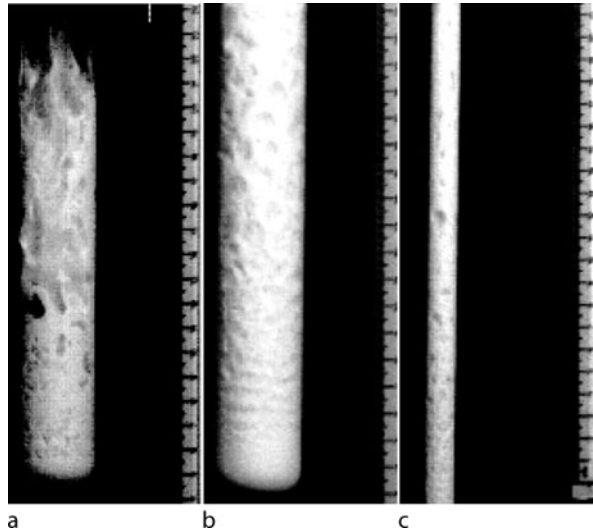
## Explosive Processes

Direct observation of the motion of magma within volcanic conduits would reveal the how and why of volcanic explosive activity. However, direct observation is considered near-impossible and indirect observation open to interpretation. Experiments with natural materials provide constraints on physico-chemical processes within conduits, but extending the experimental parameter space by using analogue materials provides greater insight and enables testing of numerical models.

**High VEI Events** Volcanic eruptions of high VEI involve the transition of a low-velocity, high-density, high-viscosity bubbly magma, for which the liquid phase is considered continuous, to a high-velocity, low-density, low-viscosity ‘dusty gas’ where the gas phase is considered continuous. This fragmentation process represents the point at which the liquid and gas phases start to separate. Studies designed to investigate this process in general, but with volcanic application in mind, form a major part of experimental research using the analogue approach.

As pressure declines, the exsolution and expansion of volatile (specifically water) from magma provides the driving mechanism for most explosive events. In order to study similar behavior in analogue systems the shock tube has been almost universally adopted, as it has for the investigation of fragmentation in natural materials (e. g., [1,2]). Such experiments became of great importance around WW2 when understanding of shock waves and flame fronts from explosions, and supersonic flight, became of interest to the fluid dynamic community; volcanological application of shock tubes emerged from this background. Mader [77] provides a review of early shock tube experimentation relating to flow within volcanic conduits, we briefly summarize these here, but focus on more recent work.

**Post-Fragmentation Flows** The decompression of beds of small, incompressible particles with compressible, gas-filled pore space can give insight into the development of two-phase flows within and above the conduit subsequent to primary fragmentation [6]. The resulting rapidly expanding, but transient flows show development of significant heterogeneity, which declines as expansion reduces particle concentration. The development of heterogeneity within expanding particle-laden gas suggests mechanisms for the formation of pyroclastic flows from high-density sections of eruption columns, whilst lower-density sections become buoyant and form an eruption plume. Using smaller particles, Cagnoli et al. [25] carried out similar experiments to explore the dynamics of short-lived Vulcanian explosions. In similarity (we think, but see Fig. 9) with post-fragmentation volcanic flows, the experimental tube was much larger than the particle size. As in Anilkumar et al. [6], these transient flows demonstrated heterogeneous distribution of particles as the mixture expanded and the gas separated from the particles (Fig. 15). These features include sub-horizontal gas-rich regions that invite comparison with those found by Spieler et al. [111] in decompressed pumice samples; the implication being that the primary fragmentation process may be incidental to the nature of the evolving flow. Heterogeneities also com-



Volcanic Eruptions, Explosive: Experimental Insights, Figure 15 Rapid 20 to 90 kPa decompression of beds of sub-100  $\mu\text{m}$  glass beads generates a particle flow as the gas phase expands in response to the pressure drop (Cagnoli et al. [25]). a shows the decompression of 100 g of beads (with average diameter of 38  $\mu\text{m}$  through 51 kPa) after 28 ms. Note the wispy flow front, indicating escape of gas, and the development of large gas bubbles within the flow (scale shows 1 cm and 0.5 cm gradations). b (150 g of 95  $\mu\text{m}$  beads, 69 kPa decompression, after 23 ms) shows the development of sub-horizontal particle-poor regions, giving some phenomenological similarity to spalling fragmentation (Fig. 9). c shows flow in a narrower tube (100 g of 38  $\mu\text{m}$  beads, 68 kPa decompression, after 20 ms) where large particle-poor regions develop at the flow margin. Reprinted with kind permission from Elsevier

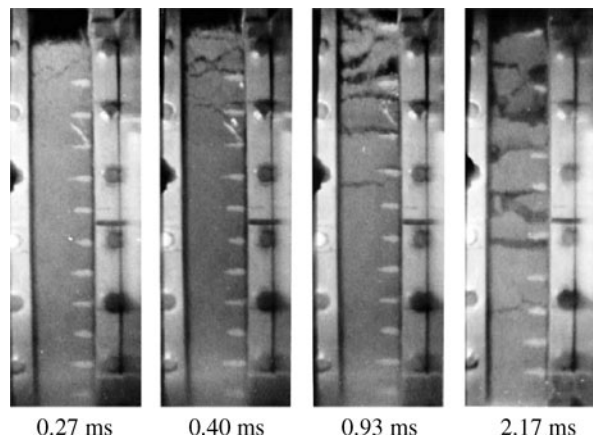
prise gas-rich bubbles, sometimes approaching tube diameter in size, and more elongated towards the flow front. Streamers of particles also emerge from the flow front indicating inhomogeneity in the distribution of gas escaping the particle flow (Fig. 15). As the flow ages and velocities increase, turbulence acts to smear out heterogeneities. The source of the heterogeneities in such transient flows is unclear. They may represent the expansion of gas rich regions in the original packed bed, suggesting that the transient flow reflects heterogeneity in the source material. Variability in permeability, and the dynamics of initial particle acceleration may also act to amplify heterogeneity as the flow expands. These experiments demonstrate the complex nature of even ‘simple’ two-phase flows that are undergoing rapid changes in variables such as pressure and velocity. The volcanic implication is that flow within a volcanic conduit can be highly heterogeneous and that modeling such flows as having smoothly changing density, as a function of time or space, will be far from representing the real



richness of the flow physics. Such modeling may give accurate prediction of time-averaged parameters, but will not reflect variability. However, it could be the variability that ultimately results in hazards like pyroclastic flows.

Chojnicki et al. [29] demonstrated that models of explosive volcanic flows based on pseudogas approximations underestimated the initial shock wave strength and velocity. Pseudogas models, and others based on steady-state experiments, then overestimated the subsequent particle bed expansion rate when applied to laboratory scale experiments. Such discrepancies suggest that one or both of the approaches, numerical and experimental, requires modification to reconcile the differences in order to understand the natural phenomena. In this case, Chojnicki et al. [29] identify processes not accounted for in the numerical approach, but which became apparent from experimental observation. Existing theory was then empirically modified to fit the unsteady experimental flow. Volcanically, these experiments suggest that the pressures present in conduits prior to Vulcanian explosive events were underestimated by a factor of approximately five when calculated using established pseudogas and inviscid shock theories. The pressures calculated from equations based on volcanically relevant experiments are consistent with the overpressures associated with Vulcanian explosions, as well as the rupture strength of volcanic rock. This demonstrates the importance of experimental design to progressing understanding of complex volcanic phenomena, and using the results of these experiments to test, modify and develop the numerical approach.

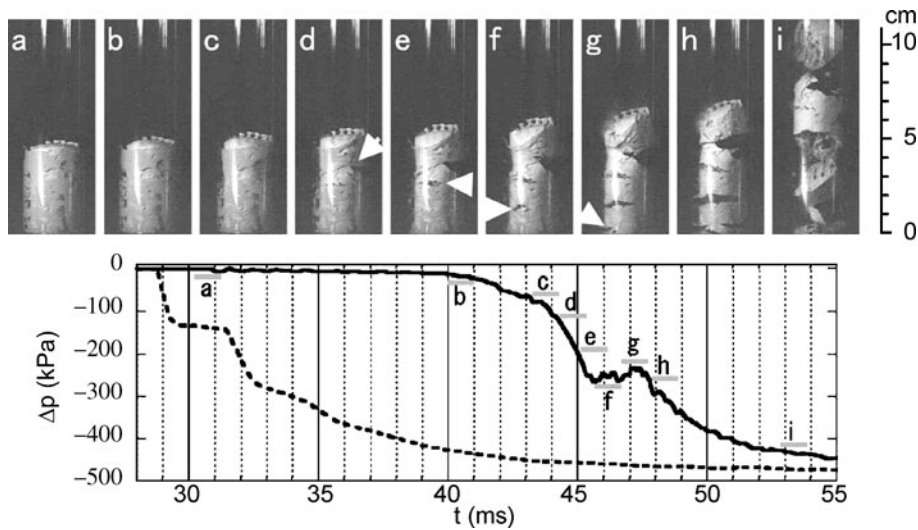
**Decompression-Driven Flows** The rapid decompression of natural vesicular solids causes fragmentation if the pressure drop is large enough for pressure gradients within the sample to exceed the tensile strength (Fig. 9). Such transient experiments have also been carried out on analogue solids in order to further characterize the behavior of this system of flows. Decompression of synthetic organic resin, with about 90% porosity and significant permeability (Alidibirov and Panov [4]), resulted in fragmentation phenomenologically similar to that of natural samples (Figs. 9, 16) undergoing brittle failure. A fragmentation front or wave was observed to propagate through the sample creating spalling-type fragmentation. Fragment size was observed to decrease, and exit velocity increase as the magnitude of decompression increases from 0.1 to 0.8 MPa. The propagation velocity of the observed fragmentation wave was less than sound speed in either the solid resin or the gas in the pores. Combining this observation with the dependence of fragment size on pressure drop suggests that fragmentation occurs as



Volcanic Eruptions, Explosive: Experimental Insights, Figure 16 An 0.8-MPa rapid decompression of vesiculated solid foam (Plastiprin) shows very similar spalling fragmentation behavior to natural vesiculated magma (compare with Fig. 9). This demonstrates the usefulness of analogue experimentation in identifying the universality (or not) of flow behaviors (Alidibirov and Panov [4]). Reproduced with kind permission from Springer Science and Business Media

gas escapes from the decompressed foam surface, creating a pressure gradient within the foam. When this pressure gradient is sufficiently steep, the foam fails revealing the next gas-escape surface. The expanding gas then accelerates the foam fragments. This implies that fragment size will also depend on sample permeability; an impermeable sample fragmenting on a scale comparable with bubble size and at small decompressions, whilst a permeable sample requires large decompressions and fragments on a scale much larger than bubble size. Very permeable materials will not fragment because the gas can escape without applying sufficient tensile force to the foam. Variability in permeability and porosity may well be reflected in fragment size for any decompression event.

Ichihara et al. [53] depressurized a vesiculated silicone compound and studied the expansion of vesicular materials across the viscoelastic transition by varying the decompression rate (Fig. 17). Fragmentation occurred at decompression rates  $> 2.7 \text{ MPa s}^{-1}$ , with only sample expansion occurring at  $< 2.7 \text{ MPa s}^{-1}$ . At high decompression rates, fragmentation occurred before expansion of the fragments. These experiments showed that both the fragmentation threshold and the nature of the fragmentation process depend not on the magnitude of decompression (above a minimum), but on the rate of pressure change. Furthermore, the rate of decompression needed to exceed the critical value for more than about 0.1 s [53], indicating that a pressure change approaching 300 kPa represents



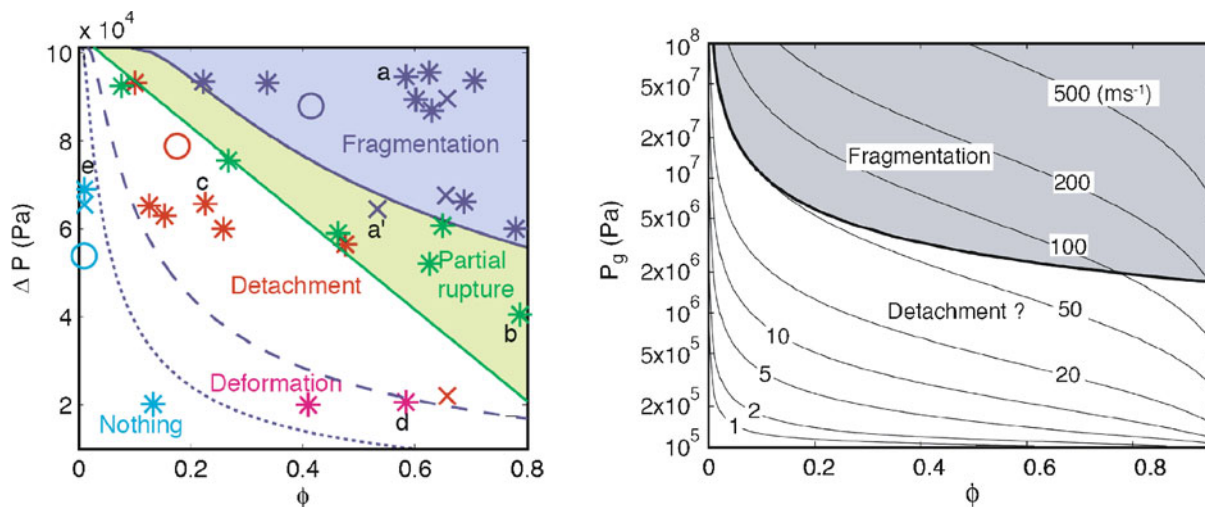
Volcanic Eruptions, Explosive: Experimental Insights, Figure 17

Rapid ( $50 \text{ MPa s}^{-1}$ ) 0.5-MPa decompression of a vesiculated dilatant (*shear-thickening*) silicone compound that acts as a brittle solid at high strain rates ( $> 3\text{s}^{-1}$ ), but a viscous liquid at low strain rates. Pressure above the sample is given by the *dashed line*, whilst that below by the *solid line*. The surface of the sample is decompressing in (a), but pressure beneath the sample remains steady until (b), decreasing more rapidly from (c). Brittle failure of the sample occurs over a  $< 10 \text{ ms}$  time span (d–g) in a spalling-fragmentation manner (Ichihara et al. [53]), about 40 ms after initial decompression. Fractures appear to initiate at the tube wall and then open to expand the flow (h, i). There is little or no expansion of the vesiculated sample or fragments. Reprinted with kind permission of AGU

the minimum pressure drop for fragmentation and also relates to the tensile strength of the experimental material. This fragmentation timescale was similar to the structural relaxation time for the glass transition (about 0.3 s), suggesting that crossing the relaxation timescale results in brittle failure of the sample. Coincidentally, the timescale for viscous control of bubble growth in the experimental material is also in the region of 0.1 s, suggesting that bubble expansion may determine the fragmentation timescale. This coincidence occurs because the rigidity of the silicone fluid is within an order of magnitude of the experimental pressure; in magmas the difference is in the region of three orders of magnitude. The nature of the fracture surfaces (Fig. 17, compare with Figs. 9 and 16), and the observation of minimal expansion before fragmentation suggests, however, that fragmentation occurred because instantaneous strain rate exceeded that required for brittle failure.

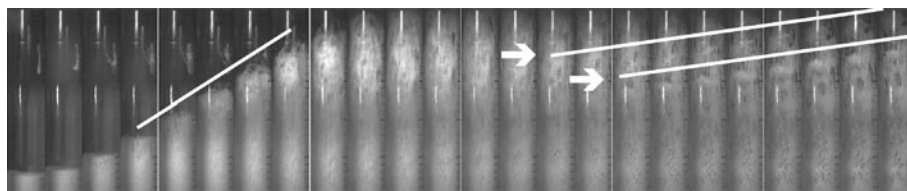
Using viscoelastic and shear-thinning solutions of 0.1%, 0.3% and 0.5% w/w polysaccharide (xanthan gum) in water, transient flow experiments exploring a wider parameter range, including variable porosity, were undertaken by Namiki and Manga [93]. Air bubbles were introduced using a hand-mixer, giving gas volume fractions up to 0.79. Samples were rapidly decompressed from atmospheric pressure to pressures ranging from 0.8 to 0.06

of an atmosphere; pressures much lower than this would encounter the saturated vapor pressure of water, causing the flows to be influenced by the production of water vapor. The shock tube was 0.05 m diameter by 0.25 m length, with fluid added to a depth of about 0.05 m. A range of flow patterns was observed (Fig. 18) as system parameters were varied. These include: fragmentation, partial rupture, detachment, deformation, and nothing; only ‘fragmentation’ results in ejection of fluid from the shock tube. Large decompression of highly vesicular samples leads to fragmenting flows (Fig. 18), which initially expand rapidly, then fragment layer by layer (Fig. 19) showing some similarity to spalling fragmentation (Figs. 9, 16, 17). In ‘partial rupture’, large-scale net-like structures develop that allow gas to escape from the fluid, but the liquid phase does not fragment *en masse*. At lower flow energies, the liquid phase detaches and contracts from the tube wall to yield an inverse annular, or ‘detached’ flow. Transient experiments allow the measurement of flow-front position, thus enabling estimation of strain rates within the flow. In these experiments the flow expansion was found to be independent of viscosity, which ranged from about 30 Pa s to 0.01 Pa s and was strain-rate dependent. For all experiments, it was found that shear rates were sufficient for elastic behavior, consistent with the independence from viscosity.



Volcanic Eruptions, Explosive: Experimental Insights, Figure 18

Aqueous xanthan gum solutions are shear thinning and develop brittle solid behavior at strain rates above about 0.1 to 1  $\text{s}^{-1}$ , depending on concentration. The *left panel* shows that rapid decompressions ( $< 0.1$  MPa) of pre-vesiculated xanthan gum solution results in a range of flow patterns (Namiki and Manga [93]). Combining data established from the decompression of natural samples with the analogue xanthan gum experiments allows the calculated fragmentation field for explosive volcanic eruptions (*right panel*) in terms of gas pressure and vesicularity. It was not possible to define the parameter space for other experimental flow patterns applied to magmatic systems. Reprinted with kind permission from Elsevier



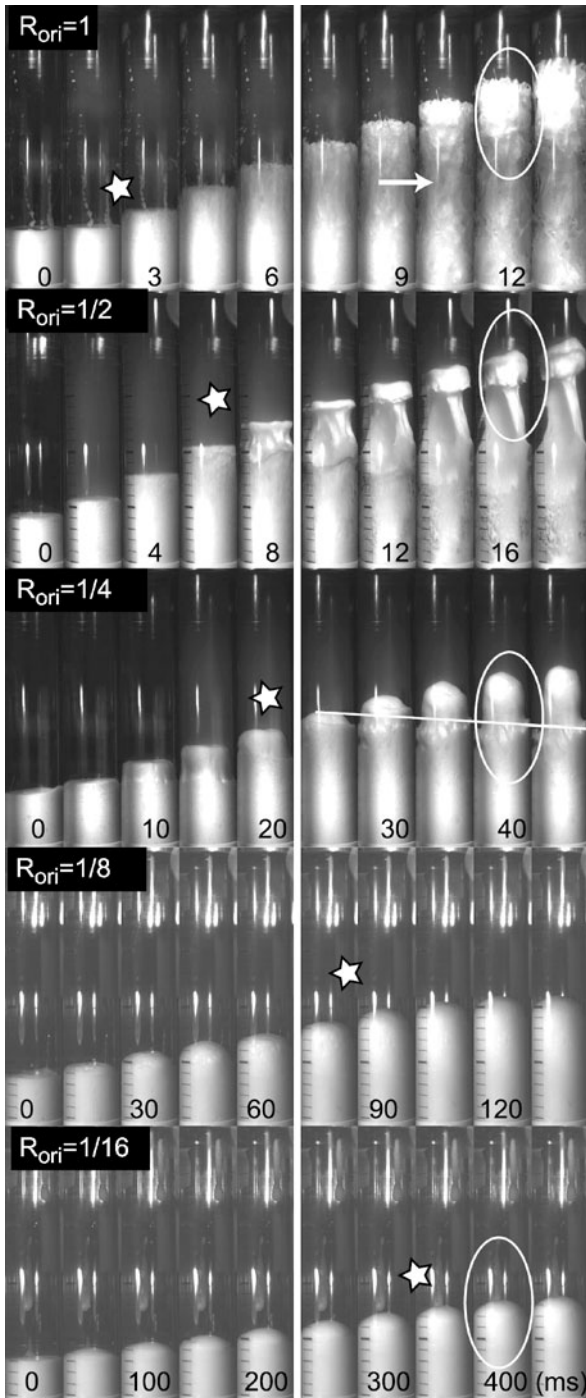
Volcanic Eruptions, Explosive: Experimental Insights, Figure 19

Fragmentation of vesiculated xanthan gum solutions occurs if decompression and vesicularity are high enough (Fig. 18). The *first frame* (zero time) shows the sample of vesiculated solution. On decompression, bubbles expand to give an approximately constant-velocity flow front (Namiki and Manga [93]), shown by the *line in frames 4–8* (3 ms between frames). During expansion, bubbles coalesce through film breakage to increase significantly in size, especially in the *upper parts* of the flow. The foam then fragments layer by layer, as shown by the *two lines* following liquid-poor regions. Most of the xanthan gum is ejected from the shock tube during the experiment. Reprinted with kind permission from Elsevier

The changes in flow pattern are attributable to the behavior of the bubble walls. Detachment of the fluid from the tube walls can be explained by the Poisson's ratio of the axially expanding foam, and represents the development of radial heterogeneity within an axial flow. However, if bubble walls rupture, elastic energy is released resulting in radial expansion and, thus, suppression of detachment. This suggests that fragmentation cannot take place from the detached flow pattern. Failure of bubble walls also increases permeability and allows gas to escape. Failure of the thin film, or plateau, between two bubbles, but not of the thicker plateau borders between three or more bubbles, is considered to result in the partial-rupture flow pattern.

The plateau borders form the net-like structure. Wholesale failure of plateau and plateau borders results in fragmentation. Application of experimental processes to volcanic scales (Fig. 18) by utilizing results from studies of transient flows of natural materials, gives indication of the decompressions needed to fragment vesicular magma for a range of vesicularities, but defining other flow patterns at volcanic scale requires further research.

Using a similar technique to Ichihara et al. [53], Namiki and Manga [94] investigated the role of transient decompression rate on aqueous xanthan gum solution. The variables here also included porosity (0.19–0.82 v/v) and absolute decompression (0.41–0.94 atmo-



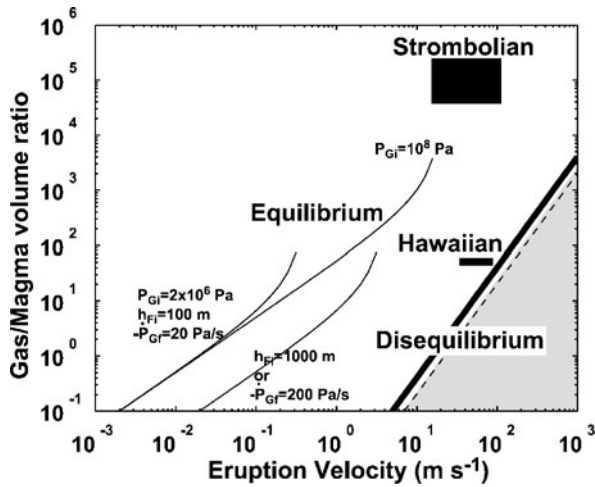
◀ Volcanic Eruptions, Explosive: Experimental Insights, Figure 20

Images of a 60-kPa decompression of xanthan gum solution of initial vesicularity about 0.6 (Namiki and Manga [94]). In order to study the effect of decompression rate, the top of the shock tube has been choked with an orifice whose radius is a fraction of the tube diameter ( $R_{ori}$ ). Experiments with rapid decompressions have short durations (time in ms on individual frames), and slow decompressions significantly longer durations, with the star indicating the time decompression ended. For  $R_{ori} = 1$ , the flow develops a rough surface suggesting rupturing of bubble walls (ellipse), and the arrow indicates the onset of fragmentation. When  $R_{ori} = 1/2$ , the flow top starts to separate (ellipse), but surfaces are smooth indicating that bubble walls are staying intact. Fragmentation does not occur when  $R_{ori} = 1/4$ , but the flow detaches from the tube wall as it expands. Smaller orifices result in flow expansion that retains wall contact. Different values of vesicularity and absolute decompression yield different patterns of behavior. Reprinted with kind permission of AGU

during high decompression rates, indicating that bubble walls are rupturing; conversely, lower rates of decompression leave flows with smooth upper surfaces. These experiments suggest that decompression rate is an important parameter in determining eruption behavior, in particular the transition between effusive and explosive eruption in low-viscosity vesiculated magmas subjected to decompression. Explosive response (fragmentation) to decompression results from disequilibrium expansion, where pressure in the bubbles is significantly higher than in the surrounding fluid and bubble expansion is limited by the enthalpy change resulting from the decompression. Bubbles cannot physically expand fast enough to maintain the same pressure as the surrounding fluid. Effusive response results from bubble pressure being close to fluid pressure.

Models of equilibrium and disequilibrium expansion were tested against experimental results. At high decompressions, equilibrium expansion would be faster than disequilibrium expansion; experiments follow the slower disequilibrium expansion model. At lower decompressions the disequilibrium model predicts more rapid expansion than the equilibrium model; experiments again follow the slower expansion, this time in equilibrium. Namiki and Manga [94] gave the critical decompression rate for invisible explosive behavior as  $-P'_{Ot} > (2\gamma/\rho_L\phi_i P_{Gi}(1-\phi_i)(\gamma-1))^{0.5}(P_{Ot}^2/h_{Fi})$ , where  $P_{Ot}$  is pressure outside bubble during decompression,  $\gamma$  is isentropic exponent,  $\rho_L$  is liquid density,  $\phi_i$  is initial gas volume fraction,  $P_{Gi}$  is internal bubble pressure and  $h_{Fi}$  the initial height of the bubbly fluid column. This experimentally based threshold decompression rate is consistent with the estimated decompression rate for the explosive/effusive transition in natural basaltic eruptions (Fig. 21), with Hawaiian eruptions sug-

sphere). High decompression rates result in a variety of flow patterns, including fragmentation, depending on porosity and absolute decompression. At low decompression rates only sample deformation or expansion takes place (Fig. 20). Rough upper surfaces develop on flows



Volcanic Eruptions, Explosive: Experimental Insights, Figure 21 Application of experiments from Fig. 20 to the volcanic case give an estimate of the magma velocity required to cause sufficient decompression rate to initiate explosive eruption (*gray region*) as a function of magma vesicularity (Namiki and Manga [94]), where  $P_{Gi}$  is initial bubble gas pressure,  $h_{Fi}$  initial height of bubble column and  $-P'_{Gf}$  the bubble decompression rate. Eruptions that result from the rapid expansion of bubbly magma are likely to plot on the thick line separating equilibrium and disequilibrium fields. Hawaiian eruptions approach the transition and, on this basis, result from the rapid expansion of vesiculated basalt magma. Strombolian eruptions, however, are indicated to result from another mechanism. Reprinted with kind permission of AGU

gested as mainly driven by the sufficiently rapid decompression of bubbly magma.

**Degassing-Driven Flows** The decompression and expansion of materials with pre-existing vesicularity provides valuable insight into volcanic behavior. However, magmas also degas on decompression, potentially adding to the explosivity of eruptions. In contrast to experiments with natural materials, where high temperatures and pressures make spatially and temporally extensive flows difficult to achieve, analogue experiments offer the possibility of studying the evolution of flow processes from single phase liquid to multi-phase fragmented flow using high-speed imaging and other data collection. Decompression of an unvesiculated hydrated magma results in, amongst other effects, the nucleation and growth of bubbles of gas phase or supercritical water, complex changes in the rheology of the liquid phase, increasing permeability and decreasing diffusivity. Incorporating all these interdependent variables into an experimental system adds complexity and makes understanding the first order effects more difficult, even if scaling between experiment and volcano

is possible. Designing experiments to investigate the dominant physics using a canonical approach reduces complexity and eases scaling requirements. The first order physical process occurring during these flows is the phase change of one system component from liquid (or more strictly, dissolved) to gas, with the associated reduction in bulk modulus and response to pressure change. Therefore, the aim of analogue experiments here is primarily to study the effect of degassing on flow.

The decompression of a liquid below its saturated vapor pressure results in boiling of that liquid as it becomes superheated. Experiments carried out by Hill and Sturtevant [50] used a shock tube to rapidly decompress a refrigerant. The refrigerant was Newtonian in rheology and of low viscosity, which, whilst dissimilar to the physical properties of magma, reduces the experiments to the first order process. Boiling produced a flow of vapor with entrained liquid drops. The superheating of the experimental liquid can be likened to the supersaturation of water in magma, and the liquid boiling to the exsolution of water from magma. The emerging gas phase expands, entrains liquid droplets and rapidly accelerates the flow. These experiments showed the development of a sharply defined plane at which explosive boiling occurred.

The addition of inert particles to a volatile liquid [113] reduces the proportion of volatile phase and maintains particles in the post-fragmentation flow. Decompression, combined with the imposition of a steep pressure gradient within a centrifuge, results in a well-defined fragmentation front when decompressions are large. However, the mass of inert material acts to reduce post-fragmentation acceleration thus maintaining pressure and suppressing heterogeneous bubble growth below the fragmentation front. Small decompressions result in bubble nucleation on the inert particles throughout the mixture, illustrating that system behavior depends on the rates at which processes respond to imposed changes. Although these experiments probably represent end-member behavior, they provided a tantalizing glimpse of possible explosive fragmentation mechanisms that may operate in volcanic conduits given suitable conditions.

Volcanically, only a few weight percent of the magma degasses to drive the flow, whilst boiling experiments represent an extreme expression of degassing flows where all the liquid can change phase to gas. Experiments that retain low liquid viscosity and Newtonian rheology, but degas only a fraction of their mass, extend our knowledge of degassing flows and are likely to behave in a different, possibly more volcanic fashion. Mader et al. [78,79] rapidly generated bubbles of  $\text{CO}_2$  in water, and throughout the fluid rather than just at the top interface. In volcanic systems,

if bubbles nucleate and grow on timescales over which they can rise and escape then significant fluid expansion does not occur. However, if the  $\text{CO}_2$  can be experimentally generated with sufficient supersaturation to drive bubble growth on timescales much shorter than those needed for bubble rise, then the two-phase mixture rapidly expands and fragments in a ductile fashion, producing liquid and foam drops suspended in gas, as the gas volume fraction increases. Two methods were used to achieve  $\text{CO}_2$  supersaturation up to 455 times ambient. Water was saturated in  $\text{CO}_2$  and decompressed in a shock tube developing high supersaturations. An alternative approach used rapid chemical reaction between  $\text{K}_2\text{CO}_3$  and  $\text{HCl}$  to generate a supersaturated  $\text{CO}_2$  solution. An intriguing positive feedback mechanism was observed whereby agitation of the liquid by the growing gas bubbles enhanced the release of gas thus increasing the violence of the flow. It was found that considerable expansion and acceleration took place before the major change in flow pattern induced by fragmentation. This contrasts with brittle systems that tend to fragment then expand [53]. Accelerations approached  $180\text{ g}$  and were proportional to volatile supersaturation; velocities peaked at  $15\text{ m s}^{-1}$ , and strain rates were in the region of  $30\text{ s}^{-1}$ . These are considered to be in the same range to those experienced volcanically and indicate a degree of similarity between laboratory and natural systems, with strain rates suggesting that high-viscosity magmas will be fragmenting in brittle fashion.

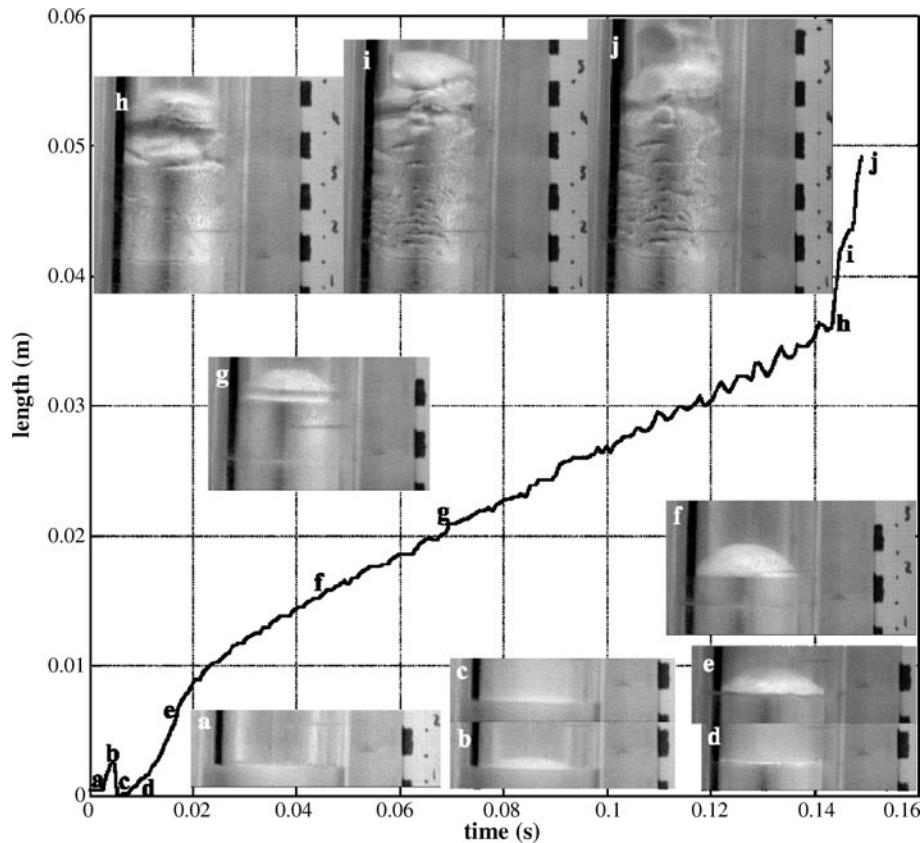
Injection apparatus in the chemical mixing experiments introduces discontinuities in the tube geometry and this appears to play a role in the behavior of the flow. Mader et al. [80] carried out relatively sustained explosive experiments, lasting about  $1.5\text{ s}$ , where a considerable volume of carbonated water was decompressed. The vessel used was a round-bottomed flask. The discharge rate of water was found to fluctuate because of flow instabilities at the base of the flask neck, probably accompanied by pressure fluctuations. The flask neck was also found to pin the point of fragmentation as the flow evolved, with similar results being found by Zhang [129]. These experiments demonstrated the importance of sustained flows in allowing quasi-steady behavior to emerge, as well as demonstrating that conduit geometry influences flow patterns.

Although water- $\text{CO}_2$  experiments illuminate the behavior of explosively degassing systems, water is many orders of magnitude less viscous than magma. Zhang et al. [132] investigated the role of viscosity by adding polymers to the water, increasing liquid viscosity by between 1 and 3 orders of magnitude. Degassing styles ranged between passive degassing with little or no expan-

sion for low supersaturation decompressions, through to rapidly expanding and fragmenting flows when pressure reduces by a factor of 50 or more on decompression. Bubbles were found to appear uniformly throughout the liquid on decompression, and were considered to be the result of heterogeneous bubble nucleation on minute particles in the liquid. This differs from boiling liquid flows and is more consistent with heterogeneous bubble nucleation on crystals in magmas. Initial acceleration of the liquid was found to be constant if ambient pressure was constant, with bubbles growing as  $time^{2/3}$ . Viscosity was found to be less important than the dissolved gas content in determining explosive capability, indicating that the low solubility of  $\text{CO}_2$  in magmas made explosive eruptions driven solely by  $\text{CO}_2$  unlikely. Experimental bubble growth was found to match well with models formulated for bubbles in magma, adding to confidence that the experimental system was, at least in part, similar to the natural phenomena.

Taddeucci et al. [116] increased complexity by carrying out degassing experiments with a viscoelastic analogue material, i. e., one that can change from viscous liquid to brittle solid behavior as strain rate increases. The silicate analogue was polydimethyl siloxane, with various additives, sold commercially as ‘Silly Putty’. Silly Putty is a dilatant material with yield strength, and is viscoelastic [26]. The volatile analogue was the inert gas Argon, which had minimal effect on the rheology of the Silly Putty. Pressures up to  $10\text{ MPa}$  were also found to have minor effect on rheology. The apparatus used was similar to the ‘Fragmentation Bomb’ (Fig. 8) used for decompression of non-degassing samples; however, because the Silly Putty experiments were carried out at room temperature, a visibly transparent shock tube could be used to facilitate high-speed imaging. Samples were loaded into the base of the shock tube, then exposed to Ar gas at pressures varying between  $4$  and  $13\text{ MPa}$ , for times between  $5$  and  $340$  hours, then rapidly decompressed to  $0.1\text{ MPa}$ .

Figure 22 shows the flow response on a depressurization of  $11\text{ MPa}$  after  $5.7$  hours of exposure to Ar gas. On decompression, a small upward bulging of the sample surface was followed by detachment of the upper part of the sample from the tube walls. This is indicated by the change in wall refraction between Figs. 22c and 22d as the Silly Putty ‘unsticks’ from ‘wetting’ the tube wall and allows a thin film of gas to change the refractive index structure of the interface. The detachment possibly represents a solid-like radial contraction to axial extension (Poisson’s ratio) caused by elastic expansion of the bubble-free sample on decompression. Alternatively, the momentary appearance of doming of the sample surface suggests a significant degree of friction with the tube wall and the



Volcanic Eruptions, Explosive: Experimental Insights, Figure 22

An 11-MPa rapid decompression of viscoelastic 'Silly Putty', saturated with Ar gas (Taddeucci et al. [116]), resulted in an expanding degassing flow. The domed flow top indicates the presence of significant wall friction, despite optical indication that the "Silly Putty" did not wet the tube wall during expansion (compare images c and d). Constant velocity expansion then occurred (e to h) with a circumferential fracture forming near the flow front at g. Flow front oscillation is apparent between g and h, but no concurrent pressure data are reported. From h, the flow front velocity increases dramatically as pervasive brittle fractures appear in the flow margin (i and j) where strain rates are highest. Fragmentation did not occur because the fractures did not propagate through the flow center where strain rates were low and viscous behavior persisted. The fractures could then close and heal due to viscous flow once strain rate declined. These experiments give valuable insight into the process of margin fracture, which could provide an efficient means for gas escape, thus suppressing explosive behavior. They also demonstrate a mechanism of generating fluctuations in magma effusion rate. At volcanic scale, margin fracture may be expected to generate a seismic response, and field evidence of such processes exists (Tuffen and Dingwell [122]). Reprinted with kind permission from Elsevier

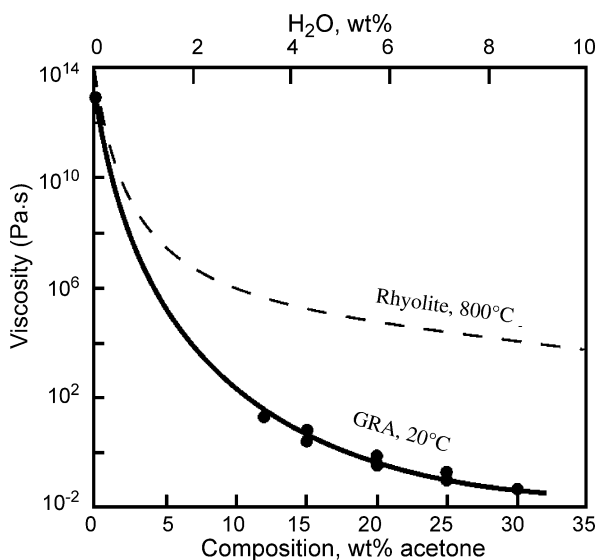
generation of high strain rate; the Silly Putty wetting of the tube wall breaks down, but wall friction remains significant. Bubbles nucleate and start to expand. The sample surface domes up, suggesting that despite detachment, wall friction is still appreciable (compare with Fig. 20). The sample then expands and develops fractures at the tube wall. Sharp fractures near the flow head were defined as 'rim fractures', with other more numerous and widespread fractures called 'pervasive fractures'. These fractures represent a solid-like response to the high strain rate imposed at the flow margin and give experimental support for a theoretical process proposed by Gonnermann and Manga [47] who postulated that fragmentation was occurring due to

viscous shear along flow edges. The fracturing process may also create seismic signals [122] and act as a trigger for seismically detectable resonant phenomena. The fractures only propagate as far as strain rates maintain brittle behavior, which acts to confine fractures within the outer section of the flow. Fragmentation will occur if the strain rate is sufficient for brittle behavior to persist right to the center of the flow. The fractures were also observed to close and heal because, once formed, lower strain rates caused reversion to viscous behavior. The development of such fracture networks has significant implication for mechanisms of magma degassing. The continual formation and healing of margin fractures gives rapid access to the conduit mar-

gin for gas deeper in the flow. This has consequences for transitions between effusive and explosive behavior.

Only a few Silly Putty experiments underwent explosive fragmentation. All these were triggered at a sudden tube widening, emphasizing the importance of conduit geometry on flow behavior in volcanic systems. In comparison, similar decompressions of pre-vesiculated samples [53] resulted in much more extensive fragmentation. This demonstrates that, although there may be more gas dissolved in an unvesiculated sample than contained in the pores of another non-degassing sample, the pre-vesiculated material is more explosive. The process of bubble nucleation and diffusive growth provides a regulating mechanism on the effective maximum rate of decompression and reduces initial explosive activity. However, the continued supply of volatile into bubbles means that degassing flows are likely to expand more given sufficient time, but in the absence of geometric structures, will they fragment in a brittle solid manner?

The viscosity of the liquid phase in magma increases by several orders of magnitude (Fig. 1) as water degasses, also decreasing the strain-rates over which viscous flow turns to brittle solid behavior. Water and melt interact to change the lengths of silicate molecules and ‘lubricate’ differential molecular motion, hence changing viscosity. As an analogue material, gum rosin derives from removing volatile compounds from tree sap, leaving 3-ring aromatic hydrocarbons typified by abietic acid ( $C_{19}H_{29}COOH$ ). It is a glassy organic solid at room temperature, but melts in the region of 80°C, and becomes increasingly less viscous as temperature rises. Warm gum rosin behaves in a viscoelastic fashion (Bagdassarov and Pinkerton [10]), being brittle when deformed rapidly but viscous under more gentle treatment. Such behavior is phenomenologically similar to a silicate melt. Phillips et al. [97] found that certain oxygen-bearing organic solvents, such as acetone, dissolved in gum rosin to yield a liquid at room temperature. At concentrations below about 20% w/w solvent these appear to be true solutions [71], with a partial pressure of acetone below that of the pure liquid suggesting some chemical interaction; another phenomenological similarity to hydrated magma. The viscosity of this liquid was found to be a non-linear function of the acetone content of the solution, again similar to a hydrated silicate melt (Fig. 23). Decompression of gum rosin-acetone solution below the saturated vapor pressure of acetone (about 20 kPa at room temperature) was found to result in volatile degassing and consequent increase in liquid viscosity. Slow decompression resulted in the establishment of a degassing interface that consumed the liquid solution. No measures were taken to remove nucleation sites from



Volcanic Eruptions, Explosive: Experimental Insights, Figure 23 The viscosity of hydrated silicate melts changes non-linearly with water content (Fig. 1). One experimental system uses gum rosin as an analogy for silicate melt, and organic solvents such as acetone and diethyl ether instead of water. The degassing of gum rosin acetone (GRA) results in increasing liquid viscosity (lower scale), in similarity with the degassing of hydrated magma (upper scale) (Mourtada-Bonnefoi and Mader [90]). The volatile content between natural and analogue systems can be compared by quantifying in terms of moles of volatile per unit volume of liquid; 4% water w/w in magma equates to about 20% w/w acetone in gum rosin. The scales here are not matched on this basis. Reprinted with kind permission from Elsevier

the liquid, suggesting that nucleation was sensitive to pressure and was confined to a surface layer in the liquid. This is similar behavior to boiling experiments [50,113], where the reaction force from expansion of degassing volatile and accompanying liquid was considered to elevate pressure in the underlying liquid and suppress bubble nucleation and growth, therefore acting to confine rapid degassing to a thin interface. This process may not be as prevalent in  $H_2O-CO_2$  and Silly Putty systems because absolute pressures were larger, in relation to hydrostatic, than in gum rosin-acetone or boiling refrigerant experiments.

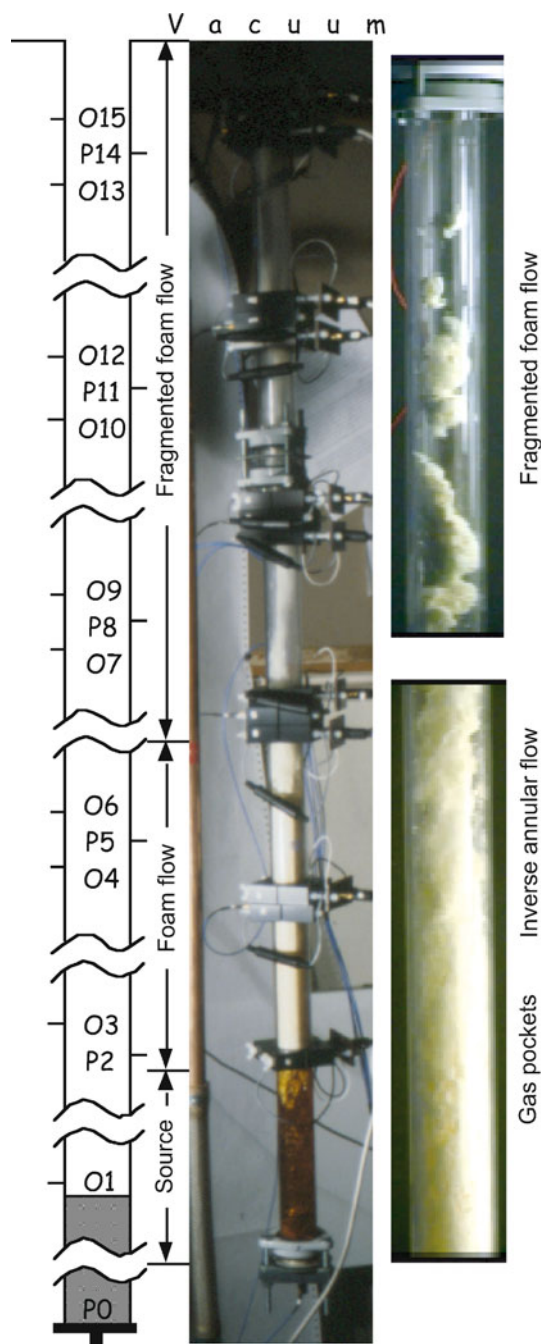
Rapid decompressions of gum rosin-acetone resulted in expanding foam growing from the liquid surface [97]. Flow-front acceleration was generally constant, except for large supersaturations with high acetone content where accelerations increased with time. These were reinterpreted by Mourtada-Bonnefoi and Mader [89] in the light of additional experiments that found bubble growth prior to fragmentation is diffusion-driven, but limited by the high-viscosity skins that develop around growing bubbles because of the reducing volatile concentration. This infers



that diffusion-limited degassing is unlikely before fragmentation removes viscous control as a physical mechanism. Accelerations and velocities using gum rosin-acetone were an order of magnitude less than those produced by CO<sub>2</sub>-H<sub>2</sub>O experiments, but this is consistent with the order of magnitude lower driving pressures. Experiments carried out with high supersaturations allowed extensive acetone degassing to take place. This resulted in the production of solid gum rosin foam with elongated bubbles, phenomenologically similar to woody or tube pumice.

The experiments of Phillips et al. [97] were carried out using small volumes of solution in un-instrumented and relatively short tubes that did not allow sufficient time or length for flow patterns to develop and stabilize. Lane et al. [70] carried out similar experiments using longer tubes, larger volumes of gum rosin-acetone and gum rosin-diethyl ether solutions, and high-speed pressure measurement. Instantaneous decompression from 10<sup>5</sup> to 10<sup>2</sup> Pa resulted in rapid degassing from the liquid surface. Flow became pseudo-steady after about 0.2 s, with foam expanding from the liquid interface, and continued explosive behavior for about 1.5 s. Pockets of gas developed at the tube wall, eventually merging to allow the expanding foam to detach from the wall. Removal of wall drag allowed the foam core to accelerate as an inverse annular, or detached flow, and break up into foam fragments that were ejected from the top of the shock tube (Fig. 24). The fragmentation of the foam core was precluded during expansion of a non-degassing system [93], suggesting that degassing systems have different flow patterns to those driven by gas expansion alone. The reason for this difference may lie with the smaller overpressure in bubbles of pseudo-steady degassing systems compared with those in transient expanding flows. The lower bubble pressures prevent bubble wall breakage and stabilize detached flow until fragmentation commences [93]. This suggests that flow behavior preceding fragmentation may differ between explosive volcanism triggered by rapid decompression of vesiculated magma, for example the May 18, 1980 Mt. St. Helens lateral blast triggered by a landslide, and sustained explosive volcanism such as the Plinian eruption which followed the blast. The fragmentation mechanism may also be different, with brittle spalling fragmentation before expansion (Figs. 9, 16 and 17) suggested for the lateral blast and fluid expansion before ‘plastic’ fragmentation (Figs. 19, 20 and 24) for the Plinian eruption.

The size of fragments of gum rosin foam was initially significantly smaller than tube diameter, but increased to near tube diameter and elongated with time, suggesting a declining flow energy. Interestingly, some of the flow features found by Cagnoli et al. [25] in expanding parti-



Volcanic Eruptions, Explosive: Experimental Insights, Figure 24 An 0.1-MPa rapid decompression of a solution of gum rosin 19% w/w diethyl ether generates a degassing flow (Lane et al. [70]). Bubble growth occurs predominantly at the upper liquid surface. The expanding flow develops gas pockets at the wall, which merge and allow a central foam core to detach from the tube wall giving inverse annular flow. The detached foam core is no longer constrained by wall drag and expands until it breaks up to give a fragmented flow. The fragments of gum rosin foam are ejected from the shock tube. Reprinted with kind permission of AGU

cle beds appear similar to those in expanding gum rosin foams, with bubbles appearing at the margins of both flows and being stretched out (Fig. 15). The sequence of flow pattern changes in the gum rosin flows was found to result in the generation of characteristic pressure fluctuations. Pressure fluctuations may be inherent in the unsteady nature of flows undergoing changes in behavior and such phenomena could be detectable by seismic means during volcanic flow.

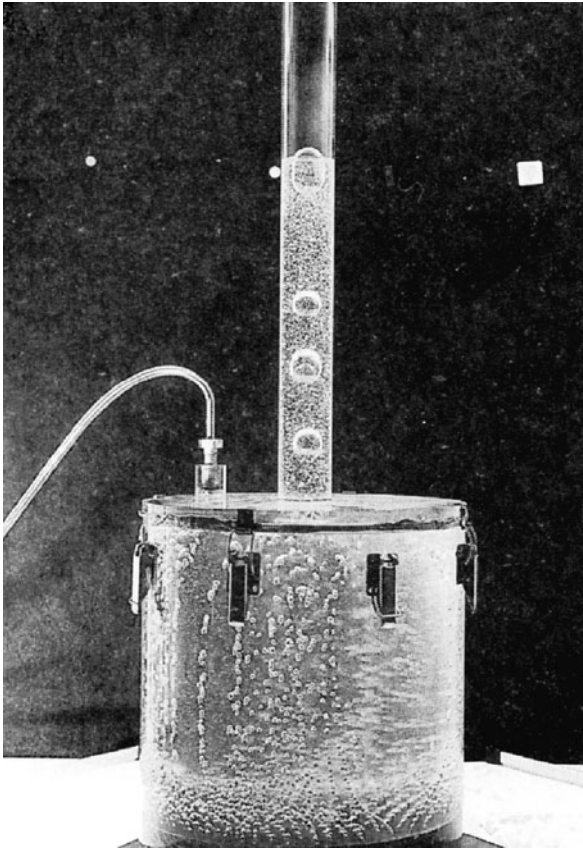
Magma can fragment in the presence of pre-existing bubbles alone without any degassing (e. g., Fig. 9), and the presence of crystals increases the number density of nucleation sites (e. g., Fig. 4), and hence the surface area over which degassing can occur. Magmas are decompressed with crystals and bubbles often present, an aspect investigated analogously by Mourtada-Bonnefoi and Mader [90] using gum rosin-acetone mixtures and a variety of particles and existing bubbles. It was found that fragmentation occurred at lower volatile contents and lower decompressions in the presence of particles and bubbles. It was also found that flow behavior depends on the distribution of particles in the liquid, with more rapid degassing where particle concentrations were high. This suggests that heterogeneity in the distribution of crystals and pre-existing bubbles will develop into sizable features within a degassing flow, at least before fragmentation. This raises the possibility that flow heterogeneity over a range of size scales may then play a role in the fragmentation process.

**Low VEI Events** High VEI events do not occur very frequently compared to anthropogenic timescales, and do not lend themselves readily to detailed field investigation of subterranean flow processes. Conversely, low VEI events, typified by Strombolian activity, occur on anthropogenically short timescales and form relatively safe natural laboratories to study magma flow in the volcanic conduit. Low VEI eruptions commonly occur at volcanoes where magma viscosity is low and degassing water is able to separate relatively easily from the magma; indeed, this separation is required to produce low VEI events. Analogue experiments of separating flows may, therefore, be compared with relatively well-studied natural processes in ways that high-VEI events cannot. Experimental conditions are also simplified in relation to those applied to high-VEI processes, with the assumptions that liquid rheology and temperature remain constant being easier to justify. Low-VEI processes may also be studied without shock tube or high-pressure apparatus, making observation and measurement more straightforward.

Strombolian eruptions are characterized by short-lived explosive events that erupt significantly higher water to sil-

icate mass ratio than that of the bulk magma [15,30]. These events are separated by longer periods of relative quiescence, suggesting the presence of a gas separation process that produces gas-rich regions that rise up the conduit between intervals of relatively gas-depleted magma. In two-phase flow terminology, one such regime is that of slug flow, where large ascending gas slugs can form from the coalescence of many much smaller bubbles. However, evidence from extensive research on water-gas systems indicates that for such a flow pattern to be stable in the conduit the overall gas volume fraction would probably need to exceed about 0.25 [34], and the time intervals between explosions would be similar to, or less than explosion duration. It is also very questionable that slug flow would form from numerous small bubbles in a relatively high viscosity liquid like basalt magma. This suggests that other mechanisms operate to promote formation of gas slugs. Jaupart and Vergnolle [63,64] were among the first to recognize that volcanic conduits are not straight vertical tubes, but exhibit significant geometrical heterogeneity that may play a major role in eruptive behavior. Experiments were carried out to examine the effect of a flat roof in trapping the numerous small bubbles rising through the magma (Fig. 25). Small air bubbles were injected at variable rate into the base of a tank containing liquids of various viscosities. The bubbles rose to the roof of the tank, which had a vertical small diameter outlet tube. Three modes of system behavior were observed. At lower liquid viscosity, a raft of bubbles formed at the roof. At some critical thickness the bubble raft collapsed on the scale of the tank roof and flowed up the outlet tube as a large gas bubble. This behavior was considered analogous to Hawaiian eruptions. At higher liquid viscosities, the bubble raft collapsed on a scale smaller than the tank roof and a series of bubbles that are much larger than those injected, but smaller than the low viscosity case, emerge up the outlet pipe. This gives insight into Strombolian eruptive activity. The final mode of behavior was the development of a relatively stable bubble raft and the escape of small bubbles up the outlet pipe to give bubbly flow at the surface.

The ascent of large gas bubbles, or slugs, up a tube can be controlled by surface tension, liquid viscosity or liquid inertia. In basaltic systems surface tension has negligible control, and inertia dominates with some component of viscous influence [106]. Conveniently, this allows the use of water and air at laboratory scale to give inertia-controlled slug flow. Seyfried and Freundt [106] carried out experiments applying the flow patterns that develop with increasing air flux through a column of water to basaltic volcanism. Single gas slug ascent was found to only occur through sudden incidents of limited volume



Volcanic Eruptions, Explosive: Experimental Insights, Figure 25 Numerous small nitrogen bubbles rising in a silicone oil of viscosity  $1 \text{ Pa s}$  (Jaupart and Vergnolle [63]) encounter a structural discontinuity in the form of the horizontal roof of the experimental tank. Bubbles trapped under the roof collide and coalesce to emerge up the narrow exit pipe as much larger bubbles. Such a mechanism can be envisaged for the production of large bubbles some  $1\text{--}3 \text{ km}$  beneath the vent at Stromboli volcano (Burton et al. [21]); however, the experimental geometry canonically demonstrates the process rather than simulating conduit structure. Large bubbles emerging from the trap then expand and become overpressured on approach to the surface (James et al. [61]). Reproduced with kind permission of Cambridge University Press

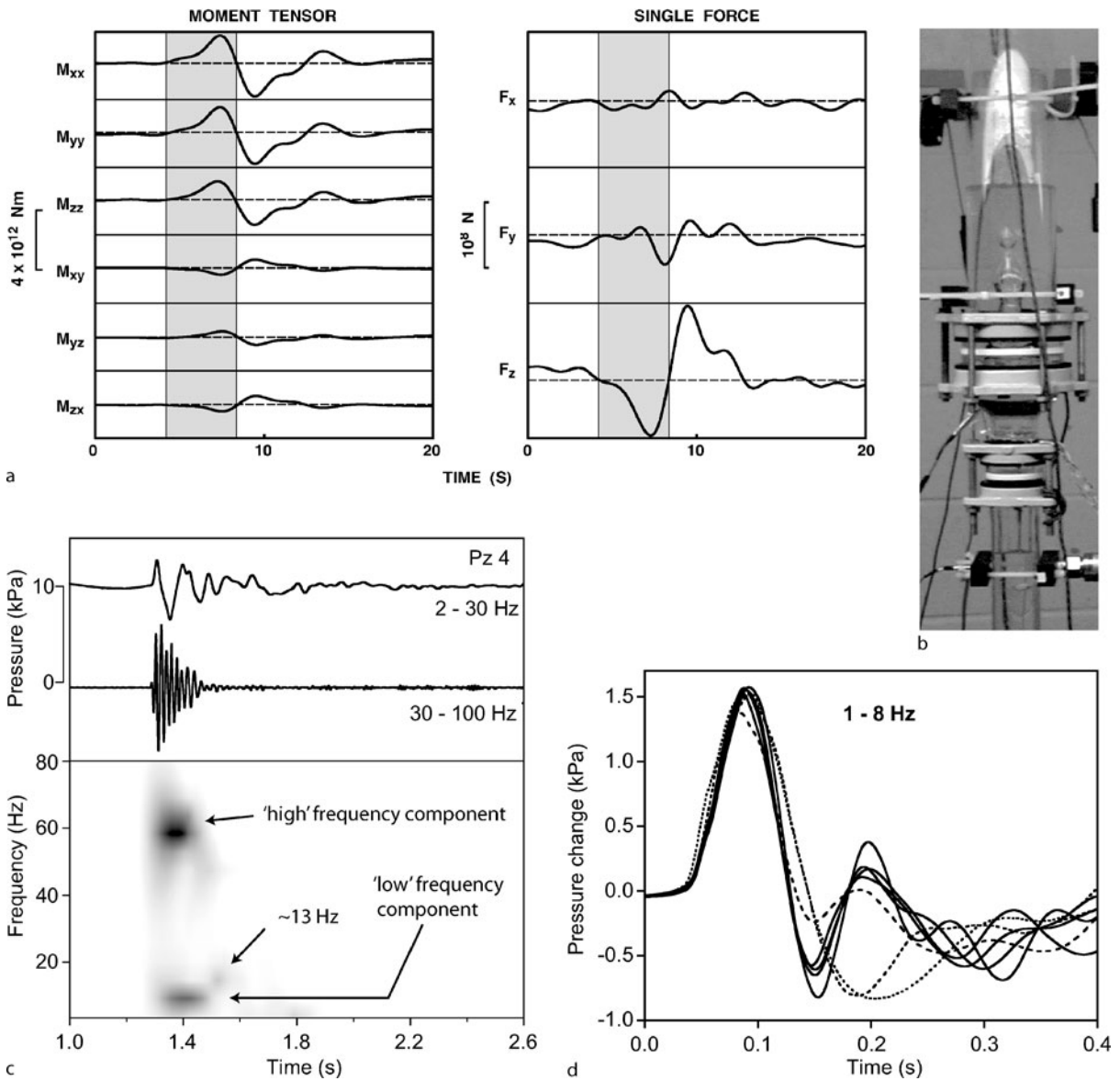
gas release, with one mechanism being that of foam collapse from a trap [63,64]. It was observed that slug burst in the tube was a gentle process, but that if the liquid level was in the reservoir atop the tube (akin to a lava lake) then burst was more vigorous and could be likened to Strombolian activity; another example of the impact of conduit geometry on flow. Experiments with gas traps in the water column found that the coalescence of trapped gas pockets resulted in significant pressure oscillations that were a potential source mechanism for volcanic tremor; linking

volcano-seismic signals to fluid flow could be a powerful method of experimental scaling inaccessible in high VEI events.

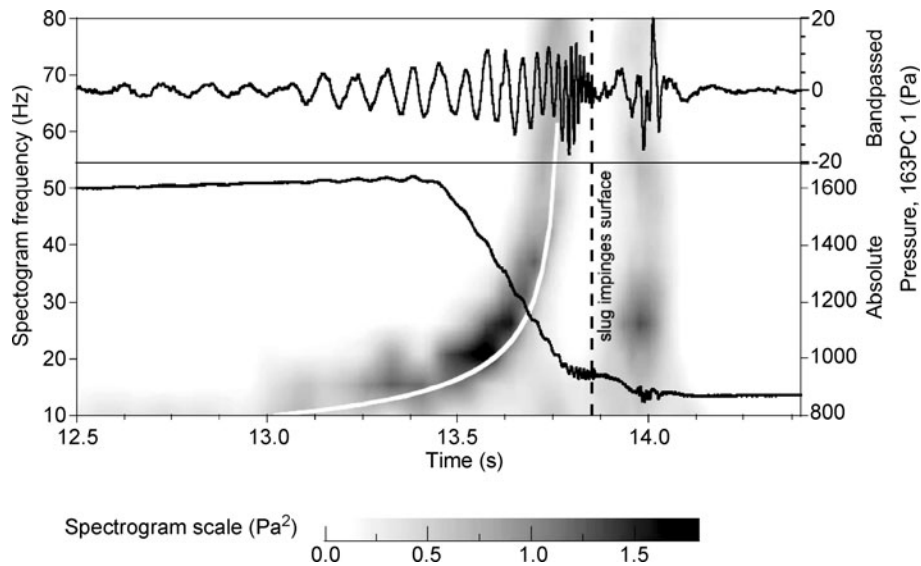
Seismic signals may result from the collapse of a foam layer trapped at a roof. Measurements made within apparatus similar to that of Jaupart and Vergnolle [63,64] reveal pressure fluctuations both within the fluid and in the air above the fluid surface in the outlet tube [102]. These could be linked to seismic and acoustic signals measured at Stromboli volcano and give insight into the role of fluid-flow on the generation of seismo-acoustic waves during explosive volcanism. During the build up of the foam layer, no pressure changes were detected. Collapse of the foam layer and motion of the gas slug up the outlet pipe produced pressure changes in the liquid tank of a few-hundred Pa, and of a few-tenths Pa in the air above the liquid level. Collapse of the foam layer generates upward gas flow in the outlet pipe thereby producing downward motion of the dense liquid phase in the outlet pipe, reducing pressure. Oscillation of liquid on the gas spring of the slug produces pressure oscillations within and above the liquid. Slug burst results in a sudden increase in oscillation frequency above the liquid, and a small pressure pulse in the liquid due to decay of liquid flow around the slug nose [59]. Ripepe et al. [102] postulate that seismic signals measured at Stromboli are generated during rapid gas expansion during foam collapse, and estimated that the slug be about  $4.5 \text{ m}$  long and induce a pressure drop of  $0.1 \text{ MPa}$  in the magma.

Inversion of highly repeatable seismic waveforms measured at Stromboli Volcano, Italy [31,33] has enabled imaging of the conduit as well as revealing the forces and pressures being created by magma flow during Strombolian eruptions. In summary, Stromboli's plumbing system appears to comprise a dyke dipping about  $72^\circ$  that runs from the vent system to an intersection with a dyke dipping about  $45^\circ$  and  $900 \text{ m}$  below the vent system. Stable and repeatable seismic sources are inferred at a depth of about  $220 \text{ m}$  below the vent system, and have the characteristics of a dyke segment intersecting the main conduit. Another, weaker, seismic source is located at the intersection of the two main dykes about  $900 \text{ m}$  below the vents. The dominant seismic source at  $220 \text{ m}$  below vent generates a downward vertical force of about  $10^8 \text{ N}$  followed by an upward force of similar magnitude (Fig. 26a). The source volume increases during the downward force and decreases with the upward force. These measurements provided a means of testing laboratory models of Strombolian eruptions.

James et al. [59] studied pressure oscillations resulting from the ascent of single gas slugs in a vertical tube.



Volcanic Eruptions, Explosive: Experimental Insights, Figure 26  
 The geometry of experimental tubes has been observed to change flow behavior in a number of instances. James et al. [60] investigated the dynamics of a gas slug ascending through a tube widening. Inversion of seismic data measured at Stromboli volcano (a, Chouet et al. [31]) indicated an initial downward force of about  $10^8$  N, accompanied by pressure increase expanding the conduit. This was followed by upward force and conduit contraction with a period of about 10 s. The experimental ascent of a gas slug through a tube widening (b) caused breakup of the slug and production of a transient gas jet. Spectral analysis of pressure (c) showed three resonant components. The low frequency component (d, about 6 Hz) showed one to two cycles of pressure increase followed by decrease, qualitatively consistent with pressure change in the conduit at Stromboli during explosive eruption. Measurement of apparatus displacement indicated a peak downward force of 30 N, which when naively scaled by  $100^3$  for physical dimension and 2.5 for density difference gives a force of  $7.5 \times 10^7$  N, encouragingly close to that estimated from seismic inversion. Reprinted with kind permission of AGU



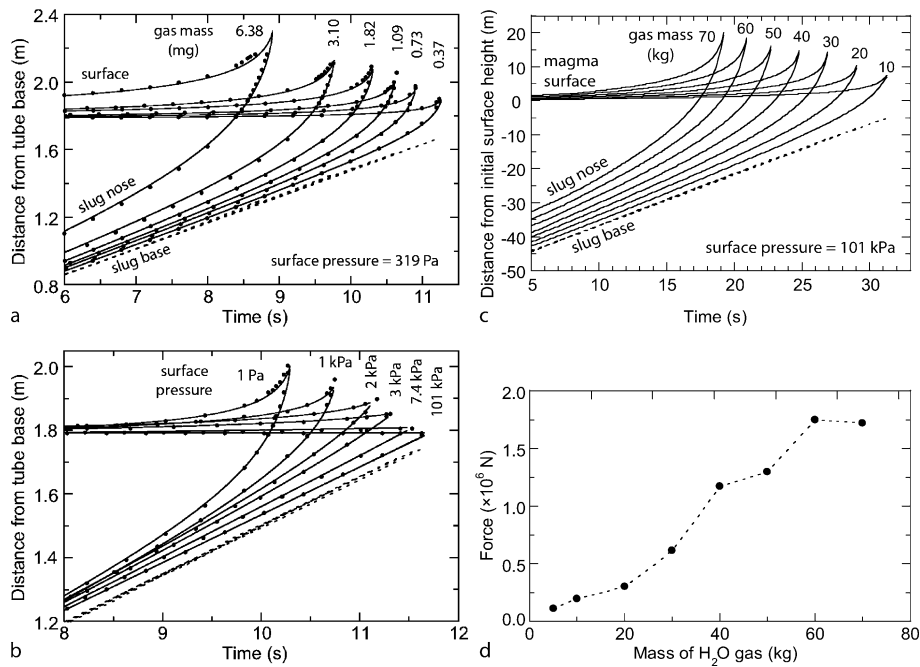
Volcanic Eruptions, Explosive: Experimental Insights, Figure 27

Strombolian eruptions are short-lived, periodic and have high volatile content. For these reasons they are associated with the formation, ascent and bursting of large bubbles of water vapor (often known as slugs), rather than the expansion of a vesicular magma (Fig. 21). The ascent of air slugs in water during small-scale experiments caused pressure oscillations (James et al. [59]). A spectrogram (gray shading) of band-passed (upper trace) pressure data (lower trace) was found to match well with calculations of the oscillation frequency (white trace) of the water above the gas slug bouncing resonantly on the gas slug. This oscillatory mechanism could not be used to account for the seismic signals used to image the conduit at Stromboli (Chouet et al. [31]). Reprinted with kind permission from Elsevier

The oscillations observed at the top of the liquid column (Fig. 27) showed an increasing amplitude and decreasing period as the slug ascended, with oscillations ceasing as the slug reached the surface. Burst of the meniscus produced further pressure fluctuation. Modeling of the oscillations revealed the source mechanism as bouncing of the liquid mass above the slug on the spring of gas within the slug. The observed pressure oscillations were not similar to the change in volume of the VLP seismic source at Stromboli (Fig. 26a), nor were oscillation amplitudes sufficient to account for forces of  $10^8$  N at volcanic scale. Bubble bursting was also quiescent rather than explosive.

Such experiments on Strombolian processes were not scaled for expansion of the gas slug as it rises, and gas chemistry suggests that slug source depths could be in the range 900–2700 m (Burton et al. [21]). A slug ascending from 1000 m in magma will expand about 200 times to reach atmospheric pressure, but in the experiments expansions were in the region of 1.2. In order to investigate more appropriate values, James et al. [61] used low vapor-pressure vacuum pump oil as an analogue liquid and varied the ambient pressure above the liquid surface between  $10^5$  and 10 Pa. The reduction in ambient pressure provided a proxy for slug ascent from depth and allowed potential

gas expansions of factors up to  $10^4$ . The positions of the slug base, slug nose and liquid surface for ascending slugs of various gas masses and ambient pressures are shown in Fig. 28. Regardless of ambient pressure and slug mass, the ascent velocity of the slug base remained essentially constant. However, the slug nose accelerated rapidly on approach to the surface when ambient pressure was low, and burst with a pressure higher than ambient. Gas slugs that underwent little expansion burst at near ambient pressure. Scaling to the volcanic case was undertaken using experimentally verified CFD modeling. This showed that slug burst pressure is a function of conduit radius and slug mass, with large slugs in narrow conduits producing gas overpressures in the region of 1 MPa at burst. This overpressure will fragment the upward-moving (tens of  $\text{m s}^{-1}$ ) meniscus of liquid existing just prior to burst. The effect of such decompression on the vesiculated basalt magma flowing down the conduit walls and slug base could be explosive fragmentation (Namiki and Manga [94]). This would eject pyroclasts, adding greatly to the pyroclast volume from the meniscus and explaining how explosive slug burst results in the ejection of significant masses of pyroclasts. There is also the possibility that the gas slug is not a single bubble, but more accurately described as a foam



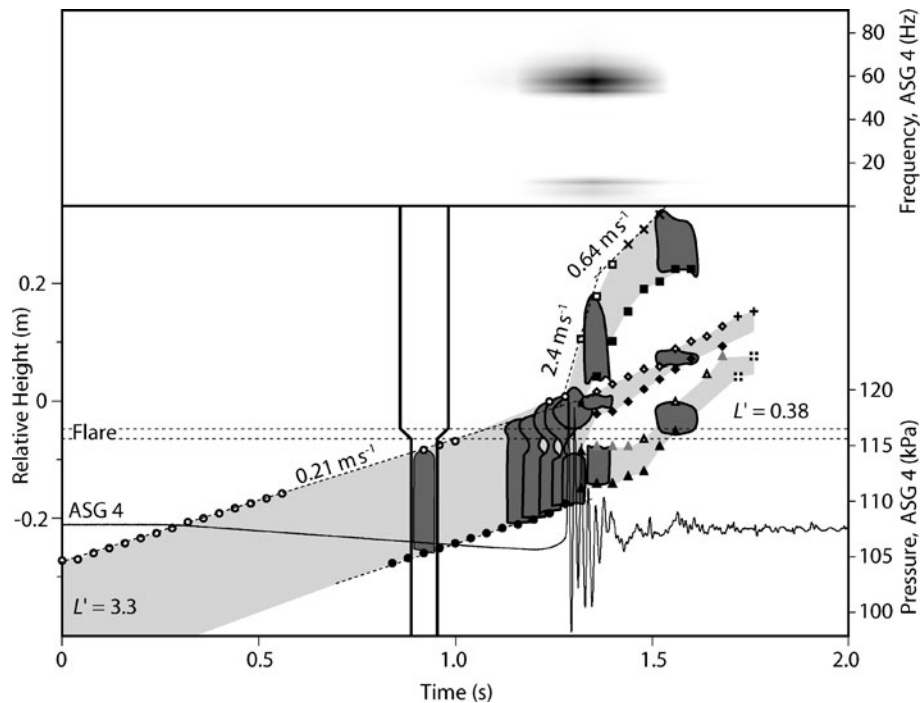
Volcanic Eruptions, Explosive: Experimental Insights, Figure 28

Experiments on slug flow are numerous in the engineering literature, but volcanic flows operate under conditions not investigated for anthropogenic purposes. One of these conditions is the large pressure change a gas bubble will experience in rising from some depth. James et al. [61] investigated the dynamics of rapid gas slug expansion as a result of this decompression using an air bubble and vacuum pump oil as analogue fluids. Using surface pressures below atmospheric allows the air bubble to undergo considerable decompression under laboratory conditions. The degree of slug expansion, and hence explosivity on burst, was found to depend on (a) the mass of gas in the slug as well as (b) the applied surface pressure. Computational fluid dynamic modeling was used to apply the experimental findings at volcanic scale (c). The process of slug expansion generated significant pressure change and vertical force (d), but the amplitudes are two orders of magnitude less than those calculated at Stromboli (Chouet et al. [31,33]). Reproduced with kind permission of the Geological Society of London

raft, although such a structure may not survive the degrees of expansion undergone in ascent from 1000 m depth with its bubble walls intact. The rapid 1 MPa decompression of a foam raft would also produce pyroclasts as the foam fragments.

Rapid slug expansion, which generates the explosivity of Strombolian eruptions, is a process operating within the top few tens of meters of the magma column (Fig. 28c). This, and the fact that the forces generated are of order  $10^6$  N upward, means that slug expansion is not the fluid dynamic source process responsible for the VLP seismic signals inverted by Chouet et al. [31]. Slug expansion probably plays little role in this particular seismic signal because the magma head is probably too large for the slug to be expanding at any appreciable rate. The stable position of the seismic source suggests geometric control of a flow process, a phenomena we have seen a number of times before. James et al. [60] measured pressure and force as a gas slug ascended from a narrow tube through a flare into

a wider tube. The gas slug segmented on passing through the flare and liquid flow during this process generated pressure oscillations (Fig. 29). Analysis of these oscillations (Fig. 26) revealed three components centered on 60, 13 and 6 Hz, with the low frequency component showing increasing pressure followed by a decrease, resembling the expansion-contraction waveform from inversion of VLP seismic data (Fig. 26a [31]). The vertical force generated during the experimental pressure increase was 30 N downward. Applying a simplistic scaling based on density and volume differences between laboratory and volcano gives a force in the region of  $10^8$  N, of similar magnitude to that emerging from inversion of VLP seismic data. It was, therefore, suggested that the source mechanism for VLP seismic signals measured at Stromboli by Chouet et al. [31] was the ascent of slugs sourced at depth [21] through a geometric widening of the feeder dyke inclined at  $72^\circ$ . The upward widening disrupts the downward flow of liquid around the ascending slug, allowing significant thicken-



Volcanic Eruptions, Explosive: Experimental Insights, Figure 29

The source mechanism suggested by laboratory experiment (Fig. 26) for seismic signals measured by Chouet et al. [31] at Stromboli Volcano, Italy is the creation and rapid deceleration of a downward-moving liquid piston, formed as a gas slug breaks up in a conduit widening (James et al. [60]). The liquid piston 'bounces' on the lower portion of the slug for a cycle or two before being consumed by the ascending remains of the original slug. The cartoon superimposed on the pressure trace correlates the position of gas bubbles with pressure measurement and spectral analysis of pressure fluctuations. The dramatic velocity increase of the slug nose does not appear to play an obvious role as a source mechanism and may be associated with the highly unstable bubble in the tube widening. Reprinted with kind permission of AGU

ing of the falling film. As this thickening falls through the downward narrowing, it pinches closed and segments the slug. The resulting liquid piston continues to descend because of its inertia, and increases pressure below it. The piston then oscillates for a cycle or two before being disrupted by the buoyancy of the gas beneath. Experimental insight suggests that the source mechanism for the VLP seismic signal is the deceleration of the liquid piston.

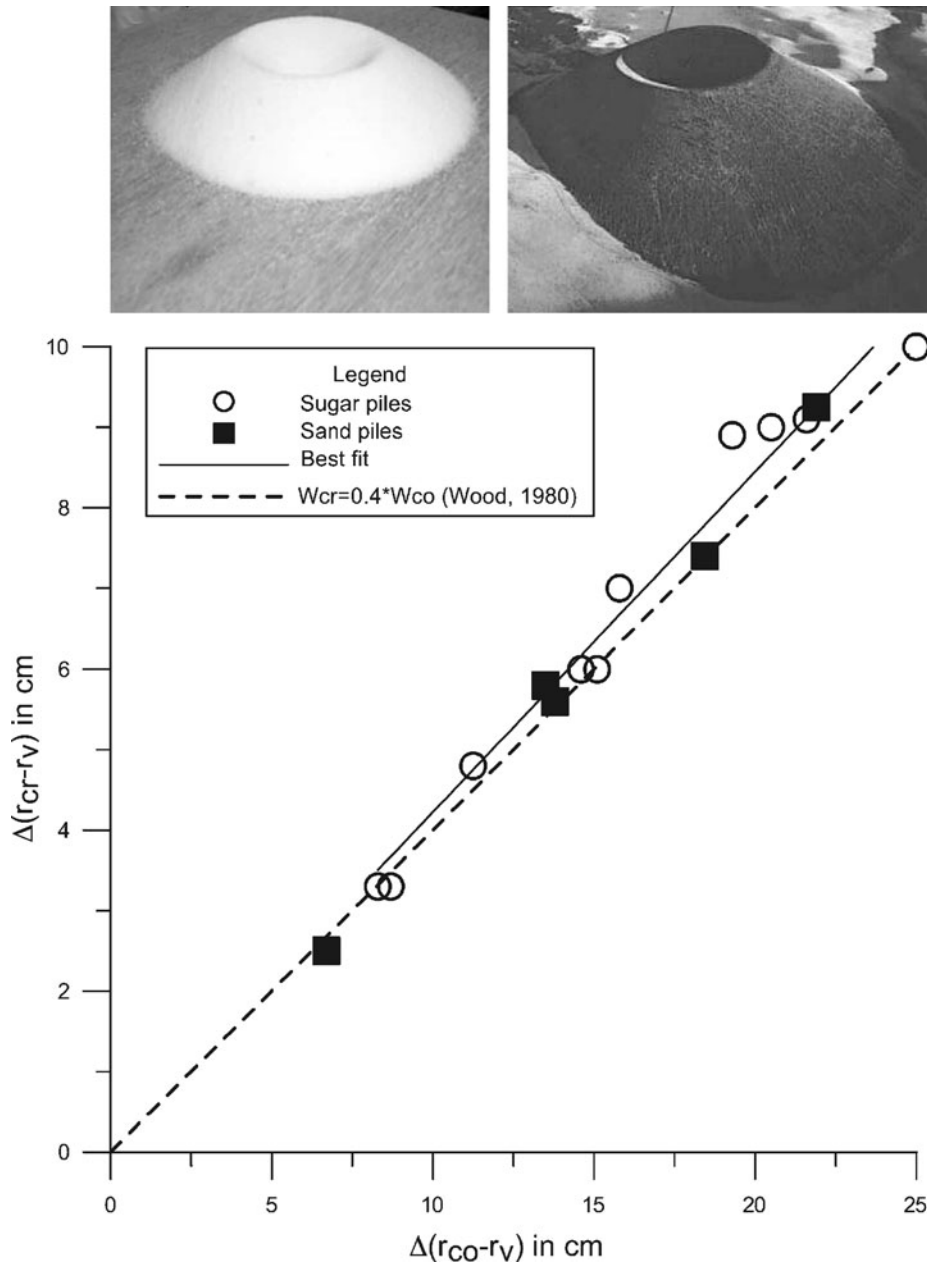
### Explosion Consequences

Various consequences of explosive degassing of magma have been studied by analogue means, and many of these are reviewed in Sparks et al. [109]. Here, we review a sample of more recent literature.

**Cinder Cones** Explosive volcanic eruptions produce pyroclastic constructs, the most common form being known as cinder or scoria cones. The basis for understanding

cinder cone growth was given by McGetchin et al. [85]. These structures, which are piles of generally unconsolidated magma fragments, are prone to gravitational collapse and rapid erosion. To test the hypothesis that cinder cone geometry is controlled by the behavior of unconsolidated fragments, Riedel et al. [100] carried out small-scale analogue experiments using grains of sugar and sand. These were built into cones, drained in the center to represent the crater. These grain piles showed similar geometry to natural cinder cones (Fig. 30). Quantification of the geometry showed that analogue and natural cones were identical in shape to within 5%. This demonstrates that the angles of repose exert fundamental control on cinder cone geometry, with surface grain flow maintaining angle of repose during cone growth. The angle of repose nature of cinder cones emphasizes their vulnerability to erosion and collapse.

**Pyroclastic Flows** Pyroclastic flows 'hug' the ground because their density is greater than that of the atmosphere.



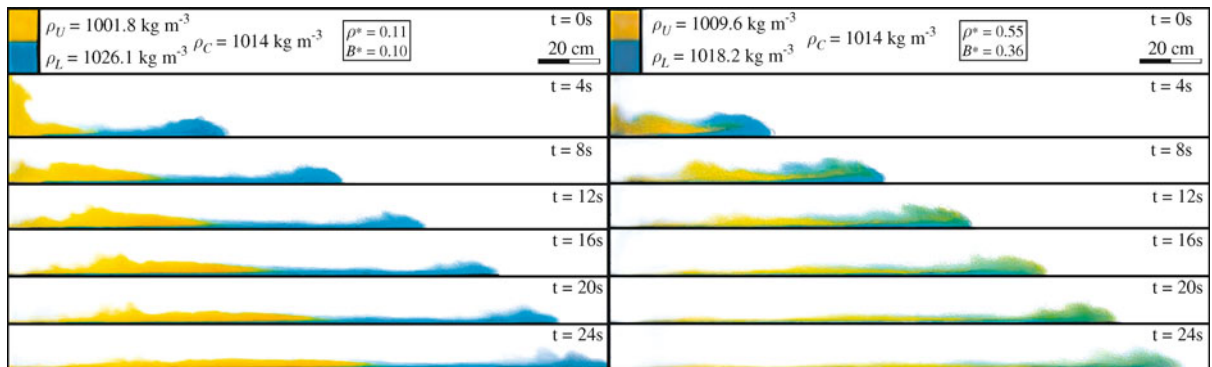
Volcanic Eruptions, Explosive: Experimental Insights, Figure 30

Cinder cones are the proximal expression of fallout from explosive volcanic eruptions. Riedel et al. [100] found that small piles of sugar (left image) or sand look similar in shape to much large piles of cinder (right image). Comparison of characteristic dimensions of laboratory cones ( $r_{cr}$  is distance from central hole to crater rim,  $r_{co}$  is distance from central hole to pile extremity and  $r_v$  is vent radius) with established measurements of natural cinder cones confirms the similarity. These experiments showed that cinder cones grow by the surface transport of grains under gravity to leave the slopes at angle of repose. Reprinted with kind permission from Elsevier

Gravity then drives the pyroclastic material in a lateral fashion; hence pyroclastic flows are part of the family of gravity currents, and the sub-family of particle-driven flows that includes turbidity currents. The flow behav-

ior of pyroclastic flows, which result from explosive volcanism or dome collapse, has been investigated at laboratory scales using analogue materials for a number of years [109]. Gladstone et al. [46] experimentally examined





Volcanic Eruptions, Explosive: Experimental Insights, Figure 31

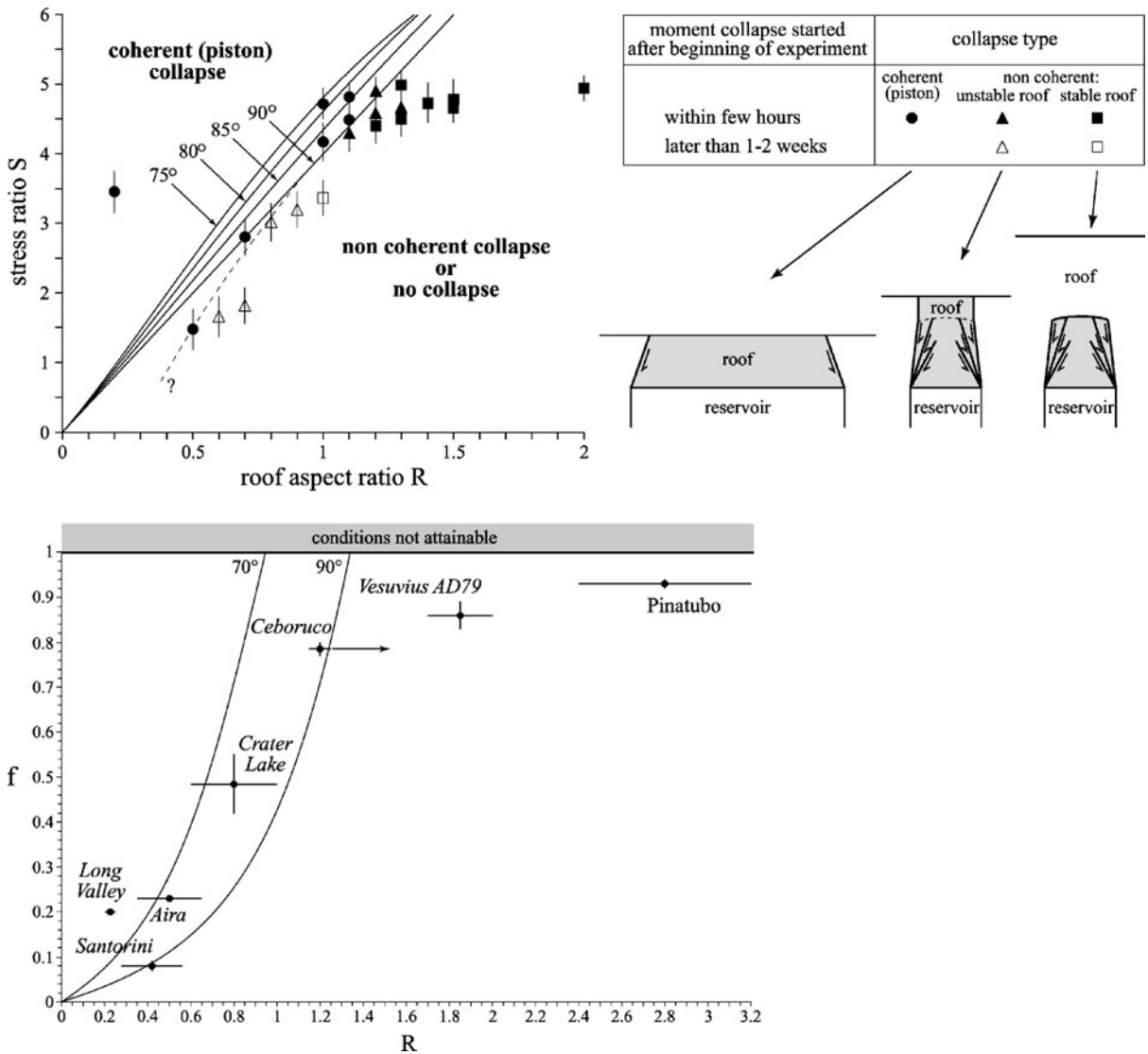
The first order behavior of gravity currents, which include pyroclastic flows, can be studied by analogue experiment using dyed brine solutions of different density propagating in a water-filled flume tank (Gladstone et al. [46]). Density stratification in the flow source material may be preserved under certain conditions of buoyancy and density (*left panel*). Pyroclastic deposits from such flows are likely to show vertical and horizontal stratification. Under other buoyancy and density conditions (*right panel*) the initial stratification is removed by turbulent mixing and resultant pyroclastic flow deposits are expected to be unstratified. Experiments that reduce complexity, as here by not including suspended particles, often give greater insight into primary controlling parameters. Complexity can then be added to understand the smaller scale processes that lead to diagnostic features identifiable in volcanic deposits. Reprinted with kind permission from Blackwell Publishing Ltd

the dynamics of gravity currents that were initially stratified in density using a 3 m long flume. A range of brine solutions of different density were layered into a lock at one end of the flume and released into tap water. The solutions were used singly, in two or three layers to investigate how stratification can develop or be homogenized in pyroclastic flows and turbidity currents. Flows were visualized by coloring the solutions allowing detection of mixing and stratification.

Flow behavior was found to depend on two dimensionless parameters. Firstly, the relative values of buoyancy between brine solutions,  $B^*$ , where most salt is in the lower layer for  $0 < B^* < 0.5$ , and in the upper layer for  $0.5 < B^* < 1$ . Secondly, the density ratio of the solutions,  $\rho^*$ , which is small for strong stratification, approaches 1 for weak stratification and equals 1 for unstratified conditions. When the lower layer has a greater proportion of the driving buoyancy ( $B^* < 0.5$ ), it can run ahead leading to flow stratification ( $\rho^*$  approaches 0, Fig. 31), or the layers can mix to produce a homogeneous current ( $\rho^*$  approaches 1, Fig. 31). If the upper layer is more buoyant ( $B^* > 0.5$ ), it travels to the nose of the current by mixing to produce a homogeneous flow ( $\rho^*$  approaches 1) or overtaking, leading to flow stratification ( $\rho^*$  approaches 0). This suggests that buoyancy-contrast controls separation, whilst density-contrast controls mixing. These observations give insights into the origin of stratification and grain size contrast in pyroclastic flow deposits and allow constraint of flow conditions from field observation.

**Caldera Collapse** The highest VEI eruptions result from the ejection of large volumes of magma. As magma emerges, country rocks overlying the magma storage area subside, resulting in the formation of a caldera. The caldera roof may fail in a number of modes. Roche and Druitt [103] carried out a force balance calculation for a caldera roof to fail as a coherent single piston, predicting that  $\Delta P/\tau_c \geq 4R$ , where  $\Delta P$  is the pressure below lithostatic at which failure occurs,  $\tau_c$  is the rupture shear stress and  $R$  is the ratio of roof thickness to diameter. Experiments were carried out using sand or a slightly cohesive sand-flour mixture as an analogue to country rock, and a draining reservoir of silicone fluid to simulate removal of magma. Comparison of experiment and theory (Fig. 32) shows that coherent roof collapse did occur in line with theoretical prediction. The experiments also showed two other more complex modes of collapse that occur at roof aspect ratios larger than the coherent collapse field, and that these collapse mechanisms may take considerable time to complete. Comparison of the experimental data with known parameters for seven calderas suggested that four had collapsed as coherent pistons, with the other collapses being non-coherent.

Experiments have helped to constrain possible volumes of caldera forming eruptions from pre-caldera conditions [44], showing a potentially powerful route of predicting the time and size of caldera-forming eruptions. Experiments were carried out using a bed of quartz sand as a rock analogue and a deflating water-filled latex bal-

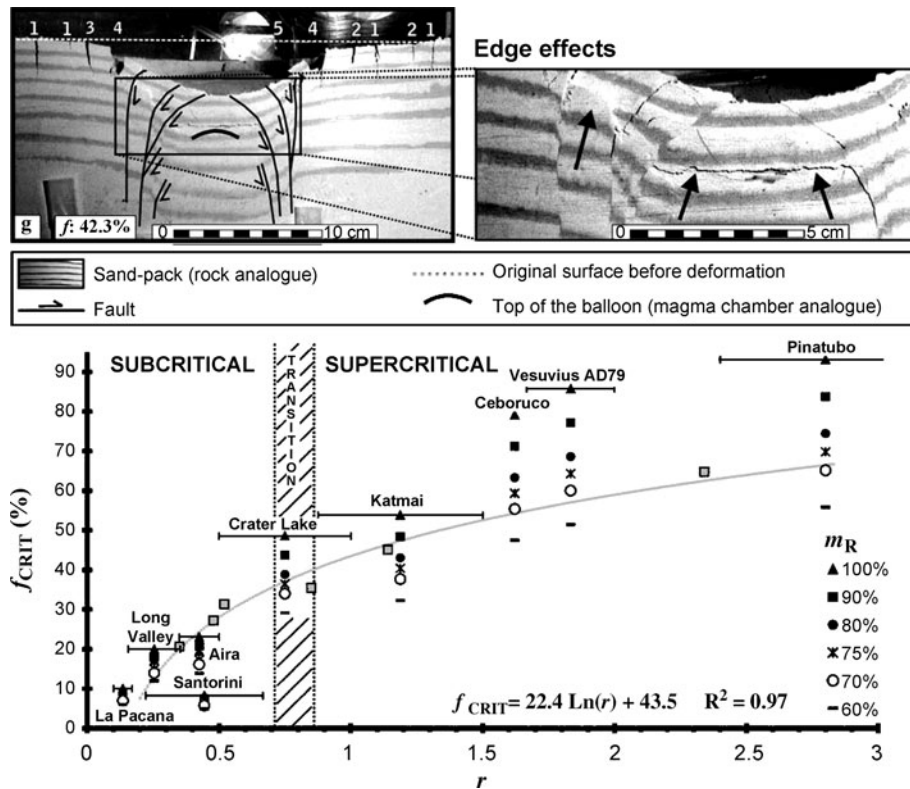


Volcanic Eruptions, Explosive: Experimental Insights, Figure 32

Large VEI eruptions may develop into caldera collapse as the country rock overlying the magma body subsides with pressure drop in the magma. Roche and Druitt [103] experimentally verified a model of coherent, piston-like formation of a caldera (upper left panel). Different modes of collapse were also found experimentally, some of which required significant time to elapse (upper right panel). This suggests that the roof of a deep, narrow magma body may not collapse simultaneously with eruption, but some time afterwards. Comparison of experimentally verified theory and estimates for proportion of magma erupted ( $f$ ) and roof aspect ratio ( $R$ ) in natural caldera collapses (bottom panel) indicate that coherent collapse does occur. Reprinted with kind permission from Elsevier

loon to represent the magma body. Extensive arguments are presented to justify application of the experimental results at volcanic scales. Coherent piston collapse was observed at low roof aspect ratios (roof thickness/diameter), becoming more non-coherent as aspect ratio increased. These experiments showed different fault structures and demonstrated the complex 3D nature of the faults that result from caldera collapse (Fig. 33). Fault development

was monitored as the balloon was drained to establish the sequence of events prior to, and during, caldera collapse. The relationship between the proportion of water drained from the balloon to make collapse inevitable ( $f_{crit}$ ) and the aspect ratio of the roof ( $r$ ) is empirically given as  $f_{crit} = 22.4 \ln(r) + 43.5$  (Fig. 33). Data from caldera-forming eruptions were compared to the experimental trend and found to be in reasonable agreement.



Volcanic Eruptions, Explosive: Experimental Insights, Figure 33

Geyer et al. [44] experimentally ascertained (*top panel*) the critical proportion of a magma chamber that must be erupted for caldera collapse to occur ( $f_{CRIT}$ ). This was related to the aspect ratio of the roof ( $r$ ) by an empirical equation. Experimental data (*gray squares*) were compared with values estimated for caldera-forming eruptions (*bottom panel*) for both subcritical (coherent or piston) collapse and supercritical (non-coherent) collapse. The estimated  $f_{CRIT}$  values are a function of the percentage of magma actually removable ( $m_R$ ). This work raises the possibility of being able to forecast the timing and volume of caldera collapse on the basis of previous eruption history and estimates of chamber depth and volume. Reprinted with kind permission from Elsevier

This suggests that the small-scale experiments mimic the first-order behavior of much larger-scale caldera collapse, but that there are issues of field measurement and second order process that require further investigation. Nevertheless, this work demonstrates the potential of being able to predict the onset and volume of caldera collapse from combining field observation and laboratory experiment.

### Future Directions

Experiments investigating processes that could occur both during, and as a consequence of, explosive volcanic eruptions are many and varied. All the experiments provide insight into volcanic processes because they are all analogous in certain aspects, whether in size in the case of experiments with natural materials, or with analogue materials as well. Curiosity into the fluid dynamics of degassing and

expanding multi-phase flows with non-Newtonian and viscoelastic materials will continue to expand knowledge over a wide parameter range of which volcanic processes form a small but motivating part. Experiments will form a key part of such research and will continue to inspire numerical models that can provide a data density unavailable experimentally. One future direction could be the development of experiments that use much greater volumes of fluid and run for longer time periods. This would be especially attractive using natural materials, but the apparatus and instrumentation challenge is considerable.

The richness of phenomena that emerge from this complex system suggests that volcanoes are individuals in their eruptive behavior, an observation common to field measurement of volcanoes. Eruption models will increasingly be tested by their ability to account for observed eruption phenomena, and linking of detailed field measurement, volcano-specific experimentation and a unified

modeling approach is one of the major current drivers of experimental volcanism. Highly instrumented volcanoes where measurement, in time and space, of such parameters as wide-spectrum ground and atmosphere motion, thermal outputs, chemistry and physics of magma components, electrical effects, flow rates and flow features are required to drive the understanding and forecasting of volcanic events; a task that, in general, becomes more challenging as VEI increases.

The role of experimental models is in providing a real system that can be repeated and measured in some detail. Computer modeling, verified by many experiments, may then be applied at volcanic scale. Quantitative comparison to field observation then tests the model system. Such combined tactics are likely to play a major role in the advancement of volcanology, and it is possible to envisage near real-time interpretation of field measurement in terms of fluid-dynamic, elasto-dynamic, atmospheric and chemical models built on the results of field, numerical and experimental volcanism. Such an autonomous system would be a major tool in quantifying fluid behavior in volcanic conduits and provide rapid assessment of changes taking place.

## Acknowledgments

We thank Bernard Chouet for a highly constructive review of this chapter, William H. K. Lee for editing this section and the staff at Springer including Julia Koerting and Kerstin Kindler.

## Bibliography

### Primary Literature

- Alidibirov M, Dingwell DB (1996) Magma fragmentation by rapid decompression. *Nature* 380:146–149
- Alidibirov M, Dingwell DB (1996) An experimental facility for the investigation of magma fragmentation by rapid decompression. *Bull Volcanol* 58:411–416
- Alidibirov M, Dingwell DB (2000) Three fragmentation mechanisms for highly viscous magma under rapid decompression. *J Volcanol Geotherm Res* 100:413–421
- Alidibirov M, Panov V (1998) Magma fragmentation dynamics: experiments with analogue porous low-strength material. *Bull Volcanol* 59(7):481–489
- Alidibirov M, Dingwell DB, Stevenson RJ, Hess K-U, Webb SL, Zinke J (1997) Physical properties of the 1980 Mount St. Helens cryptodome magma. *Bull Volcanol* 59:103–111
- Anilkumar AV, Sparks RSJ, Sturtevant B (1993) Geological implications and applications of high-velocity two-phase flow experiments. *J Volcanol Geotherm Res* 56(1–2):145–160
- Auger E, D'Auria L, Martini M, Chouet B, Dawson P (2006) Real-time monitoring and massive inversion of source parameters of very long period seismic signals: an application to Stromboli Volcano, Italy. *Geophys Res Lett* 33(4):L04301
- Bagdassarov NS, Dingwell DB (1992) A rheological investigation of vesicular rhyolite. *J Volcanol Geophys Res* 50:307–322
- Bagdassarov NS, Dingwell DB (1993) Frequency dependent rheology of vesicular rhyolite. *J Geophys Res* 98:6477–6487
- Bagdassarov N, Pinkerton H (2004) Transient phenomena in vesicular lava flows based on laboratory experiments with analogue materials. *J Volcanol Geotherm Res* 132:115–136
- Barenblatt GI (2003) *Scaling*. Cambridge University Press, Cambridge. ISBN 0-521-53394-5
- Behrens H, Gaillard F (2006) Geochemical aspects of melts: volatiles and redox behavior. *Elements* 2(5):275–280. ISSN 1811-5209
- Berthoud G (2000) Vapor explosions. *Ann Rev Fluid Mech* 32:573–611
- Birch F, Dane EB (1942) Viscosity. In: Birch F, Schairer JF, Spicer HC (eds) *Handbook of physical constants*, vol 97. Geological Society of America Special Paper, vol 36, pp 97–173
- Blackburn EA, Wilson L, Sparks RSJ (1976) Mechanisms and dynamics of Strombolian activity. *J Geol Soc Lond* 132:429–440
- Blower JD (2001) Factors controlling permeability-porosity relationships in magma. *Bull Volcanol* 63:497–504
- Blower JD, Keating JP, Mader HM, Phillips JC (2001) Inferring volcanic degassing processes from vesicle size distributions. *Geophys Res Lett* 28(2):347–350
- Bottinga Y, Weill DF (1972) The viscosity of magmatic silicate liquids: a model for calculation. *Am J Sci* 272:438–475
- Burgisser A, Gardner JE (2005) Experimental constraints on degassing and permeability in volcanic conduit flow. *Bull Volcanol* 67:42–56
- Burgisser A, Bergantz GW, Breidenthal RE (2005) Addressing complexity in laboratory experiments: the scaling of dilute multiphase flows in magmatic systems. *J Volcanol Geotherm Res* 141:245–265
- Burton M, Allard P, Muré F, La Spina A (2007) Magmatic gas composition reveals the source depth of slug-driven Strombolian explosive activity. *Science* 317:227–230
- Büttner R, Dellino P, Zimanowski B (1999) Identifying magma-water interaction from the surface features of ash particles. *Nature* 401(6754):688–690
- Büttner R, Roder H, Zimanowski B (1997) Electrical effects generated by experimental volcanic explosions. *Appl Phys Lett* 70(14):1903–1905
- Büttner R, Zimanowski B, Roder H (2000) Short-time electrical effects during volcanic eruption: Experiments and field measurements. *J Geophys Res* 105(B2):2819–2827
- Cagnoli B, Barmin A, Melnik O, Sparks RSJ (2002) Depressurization of fine powders in a shock tube and dynamics of fragmented magma in volcanic conduits. *Earth Planet Sci Lett* 204(1–2):101–113
- Cambridge Polymer Group (2002) The Cambridge Polymer Group Silly Putty™ “Egg”. <http://www.campoly.com/documents/appnotes/sillyputty.pdf>, accessed 10/10/08
- Carey SN, Sigurdsson H (1982) Influence of particle aggregation on deposition of distal tephra from the May 18, 1980, eruption of Mount St-Helens volcano. *J Geophys Res* 87(NB8):7061–7072
- Carroll MR, Holloway JR (eds) (1994) *Volatiles in magmas*. Rev Miner 30:517. ISBN 0-939950-36-7

29. Chojnicki K, Clarke AB, Phillips JC (2006) A shock-tube investigation of the dynamics of gas-particle mixtures: Implications for explosive volcanic eruptions. *Geophys Res Lett* 33:L15309. doi:10.1029/2006GL026414
30. Chouet BA, Hamisevicz B, McGetchin TR (1974) Photoballistics of volcanic jet activity at Stromboli, Italy. *J Geophys Res* 79:4961–4976
31. Chouet B, Dawson P, Ohminato T, Martini M, Saccorotti G, Giudicepietro F, De Luca G, Milana G, Scarpa R (2003) Source mechanisms of explosions at Stromboli Volcano, Italy, determined from moment-tensor inversions of very-long-period data. *J Geophys Res* 108(B1):2019. doi:10.1029/2002JB001919
32. Chouet B, Dawson P, Nakano M (2006) Dynamics of diffusive bubble growth and pressure recovery in a bubbly rhyolitic melt embedded in an elastic solid. *J Geophys Res* 111: B07310
33. Chouet B, Dawson P, Martini M (2008) Shallow-conduit dynamics at Stromboli Volcano, Italy, imaged from waveform inversions. In: Lane SJ, Gilbert JS (eds) *Fluid motion in volcanic conduits: a source of seismic and acoustic signals*. Geological Society, London. Special Publication 307:57–84. doi:10.1144/SP307.5
34. Clift R, Grace JR, Weber ME (1978) *Bubbles, drops and particles*. Academic Press, London, p 380. ISBN 012176950X
35. Costa A (2006) Permeability-porosity relationship: a reexamination of the Kozeny–Carman equation based on a fractal pore-space geometry assumption. *Geophys Res Lett* 33:L02318
36. Dellino P, Zimanowski B, Büttner R, La Volpe L, Mele D, Sulpizio R (2007) Large-scale experiments on the mechanics of pyroclastic flows: design, engineering, and first results. *J Geophys Res* 112:B04202
37. Dickinson JT, Langford SC, Jensen LC, McVay GL, Kelso JF, Pantano CG (1988) Fractoemission from fused-silica and sodium-silicate glasses. *J Vac Sci Technol A* 6(3):1084–1089
38. Dingwell DB (1996) Volcanic dilemma: flow or blow? *Science* 273(5278):1054–1055
39. Dingwell DB, Webb SL (1989) Structural relaxation in silicate melts and non-Newtonian melt rheology in geologic processes. *Phys Chem Miner* 16:508–516
40. Donaldson EE, Dickinson JT, Bhattacharya SK (1988) Production and properties of ejecta released by fracture of materials. *J Adhes* 25(4):281–302
41. Druitt TH, Avaré G, Bruni G, Lettieri P, Maez F (2007) Gas retention in fine-grained pyroclastic flow materials at high temperatures. *Bull Volcanol* 69(8):881–901
42. Eichelberger JC, Carrigan CR, Westrich HR, Price RH (1986) Non-explosive silicic volcanism. *Nature* 323:598–602
43. Gardner JE (2007) Heterogeneous bubble nucleation in highly viscous silicate melts during instantaneous decompression from high pressure. *Chem Geol* 236:1–12
44. Geyer A, Folch A, Marti J (2006) Relationship between caldera collapse and magma chamber withdrawal: an experimental approach. *J Volcanol Geotherm Res* 157(4):375–386
45. Gilbert JS, Lane SJ (1994) The origin of accretionary lapilli. *Bull Volcanol* 56(5):398–411
46. Gladstone C, Ritchie LJ, Sparks RSJ, Woods AW (2004) An experimental investigation of density-stratified inertial gravity currents. *Sedimentology* 51(4):767–789
47. Gonnermann HM, Manga M (2003) Explosive volcanism may not be the inevitable consequence of magma fragmentation. *Nature* 424:432–435
48. Grunewald U, Zimanowski B, Büttner R, Phillips LF, Heide K, Büchel G (2007) MFCI experiments on the influence of NaCl-saturated water on phreatomagmatic explosions. *J Volcanol Geotherm Res* 159:126–137
49. Hess K-U, Dingwell DB (1996) Viscosities of hydrous leucogranitic melts: a non-Arrhenian model. *Am Miner* 81:1297–1300
50. Hill LG, Sturtevant B (1990) An experimental study of evaporation waves in a superheated liquid. In: Meier GEA, Thompson PA (eds) *Adiabatic waves in liquid-vapor systems*. IUTAM Symposium Göttingen, Germany. Springer, Berlin, pp 25–37. ISBN 3-540-50203-3
51. Hoover SR, Cashman KV, Manga M (2001) The yield strength of subliquidus basalts – experimental results. *J Volcanol Geotherm Res* 107(1–3):1–18
52. Hurwitz S, Navon O (1994) Bubble nucleation in rhyolitic melts: experiments at high pressure, temperature, and water content. *Earth Planet Sci Lett* 122(3–4):267–280
53. Ichihara M, Rittel D, Sturtevant B (2002) Fragmentation of a porous viscoelastic material: implications to magma fragmentation. *J Geophys Res* 107(B10):2229. doi:10.1029/2001JB000591
54. Ida Y (2007) Driving force of lateral permeable gas flow in magma and the criterion of explosive and effusive eruptions. *J Volcanol Geotherm Res* 162(3–4):172–184
55. Ishibashi H, Sato H (2007) Viscosity measurements of subliquidus magmas: alkali olivine basalt from the Higashi-Matsuura district, Southwest Japan. *J Volcanol Geotherm Res* 160(3–4):223–238
56. James MR, Lane SJ, Gilbert JS (2000) Volcanic plume electrification: experimental investigation of a fracture-charging mechanism. *J Geophys Res* 105(B7):16641–16649
57. James MR, Gilbert JS, Lane SJ (2002) Experimental investigation of volcanic particle aggregation in the absence of a liquid phase. *J Geophys Res* 107(B9):2191
58. James MR, Lane SJ, Gilbert JS (2003) Density, construction, and drag coefficient of electrostatic volcanic ash aggregates. *J Geophys Res* 108(B9):2435
59. James MR, Lane SJ, Chouet B, Gilbert JS (2004) Pressure changes associated with the ascent and bursting of gas slugs in liquid-filled vertical and inclined conduits. *J Volcanol Geotherm Res* 129(1–3):61–82
60. James MR, Lane SJ, Chouet BA (2006) Gas slug ascent through changes in conduit diameter: laboratory insights into a volcano-seismic source process in low-viscosity magmas. *J Geophys Res* 111:B05201. doi:10.1029/2005JB003718
61. James MR, Lane SJ, Corder SB (2008) Modelling the rapid near-surface expansion of gas slugs in low-viscosity magmas. In: Lane SJ, Gilbert JS (eds) *Fluid motion in volcanic conduits: a source of seismic and acoustic signals*. Geological Society, London. Special Publication 307:147–167. doi:10.1144/SP307.9
62. Jaupart C, Allegre CJ (1991) Gas content, eruption rate and instabilities of eruption regime in silicic volcanoes. *Earth Planet Sci Lett* 102:413–429
63. Jaupart C, Vergnolle S (1988) Laboratory models of Hawaiian and Strombolian eruptions. *Nature* 331(6151):58–60

64. Jaupart C, Vergnolle S (1989) The generation and collapse of a foam layer at the roof of a basaltic magma chamber. *J Fluid Mech* 203:347–380
65. Kalinichev AG (2001) Molecular simulations of liquid and supercritical water: thermodynamics, structure, and hydrogen bonding. *Rev Miner Geochem* 42:83–129
66. Kaminski E, Jaupart C (1998) The size distribution of pyroclasts and the fragmentation sequence in explosive volcanic eruptions. *J Geophys Res* 103(B12):29759–29779
67. Klug C, Cashman KV (1996) Permeability development in vesiculating magmas: implications for fragmentation. *Bull Volcanol* 58:87–100
68. Kouchi A, Tsuchiyama A, Sunagawa I (1986) Effect of stirring on crystallisation kinetics of basalt: texture and element partitioning. *Contrib Miner Petrol* 93:429–438
69. Kueppers U, Perugini D, Dingwell DB (2006) “Explosive energy” during volcanic eruptions from fractal analysis of pyroclasts. *Earth Planet Sci Lett* 248(3–4):800–807
70. Lane SJ, Chouet BA, Phillips JC, Dawson P, Ryan GA, Hurst E (2001) Experimental observations of pressure oscillations and flow regimes in an analogue volcanic system. *J Geophys Res* 106(B4):6461–6476
71. Lane SJ, Phillips JC, Ryan GA (2008) Dome-building eruptions: insights from analogue experiments. In: Lane SJ, Gilbert JS (eds) *Fluid motion in volcanic conduits: a source of seismic and acoustic signals*. Geological Society, London
72. Lautze NC, Houghton BF (2007) Linking variable explosion style and magma textures during 2002 at Stromboli volcano, Italy. *Bull Volcanol* 69(4):445–460
73. Liu Y, Zhang Y, Behrens H (2005) Solubility of H<sub>2</sub>O in rhyolitic melts at low pressures and a new empirical model for mixed H<sub>2</sub>O–CO<sub>2</sub> solubility in rhyolitic melts. *J Volcanol Geotherm Res* 143:219–235
74. Llewellyn EW, Manga M (2005) Bubble suspension rheology and implications for conduit flow. *J Volcanol Geotherm Res* 143:205–217
75. Llewellyn EW, Mader HM, Wilson SDR (2002) The rheology of a bubbly liquid. *Proc R Soc Lond A* 458:987–1016
76. Lyakhovskiy V, Hurwitz S, Navon O (1996) Bubble growth in rhyolitic melts: experimental and numerical investigation. *Bull Volcanol* 58:19–32
77. Mader HM (1998) Conduit flow and fragmentation. *Geol Soc London* 145:51–71
78. Mader HM, Zhang Y, Phillips JC, Sparks RSJ, Sturtevant B, Stolper E (1994) Experimental simulations of explosive degassing of magma. *Nature* 372(6501):85–88
79. Mader HM, Phillips JC, Sparks RSJ, Sturtevant B (1996) Dynamics of explosive degassing of magma: Observations of fragmenting two-phase flows. *J Geophys Res Solid Earth* 101(B3):5547–5560
80. Mader HM, Brodsky EE, Howard D, Sturtevant B (1997) Laboratory simulations of sustained volcanic eruptions. *Nature* 388(6641):462–464
81. Marshall JR, Sauke TB, Cuzzi JN (2005) Microgravity studies of aggregation in particulate clouds. *Geophys Res Lett* 32(11):L11202
82. Martel C, Dingwell DB, Spieler O, Pichavant M, Wilke M (2000) Fragmentation of foamed silicic melts: an experimental study. *Earth Planet Sci Lett* 178:47–58
83. Martel C, Dingwell DB, Spieler O, Pichavant M, Wilke M (2001) Experimental fragmentation of crystal- and vesicle-bearing silicic melts. *Bull Volcanol* 63:398–405
84. Mason RM, Starostin AB, Melnik OE, Sparks RSJ (2006) From Vulcanian explosions to sustained explosive eruptions: the role of diffusive mass transfer in conduit flow dynamics. *J Volcanol Geotherm Res* 153(1–2):148–165
85. McGetchin TR, Settle M, Chouet B (1974) Cinder cone growth modeled after Northeast Crater, Mount Etna, Sicily. *J Geophys Res* 79:3257–3272
86. McMillan PF (1994) Water solubility and speciation models. In: Carroll MR, Holloway JR (eds) *Volatiles in magmas*.
87. Melnik O, Sparks RSJ (2002) Dynamics of magma ascent and lava extrusion at Soufrière Hills Volcano, Montserrat. In: Druitt T, Kokelaar P (eds) *The eruption of Soufrière Hills Volcano, Montserrat, from 1995 to 1999*. Geological Society, London, pp 153–171
88. Moore JG, Peck DL (1962) Accretionary lapilli in volcanic rocks of the western continental United-States. *J Geol* 70(2): 182–193
89. Mourtada-Bonnefoi CC, Mader HM (2001) On the development of highly-viscous skins of liquid around bubbles during magmatic degassing. *Geophys Res Lett* 28(8):1647–1650
90. Mourtada-Bonnefoi CC, Mader HM (2004) Experimental observations of the effects of crystals and pre-existing bubbles on the dynamics and fragmentation of vesiculating flows. *J Volcanol Geotherm Res* 129:83–97. doi:10.1016/S0377-0273(03)00233-6
91. Murase T (1962) Viscosity and related properties of volcanic rocks at 800° to 1400°C. *Hokkaido Univ Fac Sci Jour Ser VII* 1:487–584
92. Murase T, McBirney AR (1973) Properties of some common igneous rocks and their melts at high-temperatures. *Geol Soc Am Bull* 84(11):3563–3592
93. Namiki A, Manga M (2005) Response of a bubble bearing viscoelastic fluid to rapid decompression: implications for explosive volcanic eruptions. *Earth Planet Sci Lett* 236:269–284
94. Namiki A, Manga M (2006) Influence of decompression rate on the expansion velocity and expansion style of bubbly fluids. *J Geophys Res* 111:B11208
95. Newhall CG, Self S (1982) The explosivity index (VEI) – an estimate of explosive magnitude for historical volcanism. *J Geophys Res* 87(NC2):1231–1238
96. Parfitt EA (2004) A discussion of the mechanisms of explosive basaltic eruptions. *J Volcanol Geotherm Res* 134:77–107
97. Phillips JC, Lane SJ, Lejeune A-M, Hilton M (1995) Gum rosin – acetone system as an analogue to the degassing behaviour of hydrated magmas. *Bull Volcanol* 57:263–268
98. Pinkerton H, Norton G (1995) Rheological properties of basaltic lavas at sub-liquidus temperatures – laboratory and field-measurements on lavas from Mount Etna. *J Volcanol Geotherm Res* 68(4):307–323
99. Pyle DM (2000) Sizes of volcanic eruptions. In: Sigurdsson H, Houghton B, McNutt SR, Rymer H, Stix J (eds) *Encyclopaedia of volcanoes*. Academic Press, pp 263–269. ISBN 0-12-643140-X
100. Riedel C, Ernst GGJ, Riley M (2003) Controls on the growth and geometry of pyroclastic constructs. *J Volcanol Geotherm Res* 127(1–2):121–152
101. Riley CM, Rose WI, Bluth GJS (2003) Quantitative shape measurements of distal volcanic ash. *J Geophys Res Solid Earth* 108(B10):2504

102. Ripepe M, Ciliberto S, Della Schiava M (2001) Time constraints for modeling source dynamics of volcanic explosions at Stromboli. *J Geophys Res Solid Earth* 106(B5):8713–8727
103. Roche O, Druitt TH (2001) Onset of caldera collapse during ignimbrite eruptions. *Earth Planet Sci Lett* 191(3–4):191–202
104. Saar MO, Manga M (1999) Permeability-porosity relationship in vesicular basalts. *Geophys Res Lett* 26:111–114
105. Schumacher R, Schmincke HU (1995) Models for the origin of accretionary lapilli. *Bull Volcanol* 56(8):626–639
106. Seyfried R, Freundt A (2000) Experiments on conduit flow and eruption behavior of basaltic volcanic eruptions. *J Geophys Res Solid Earth* 105:23727–23740
107. Shaw HR (1972) Viscosities of magmatic silicate liquids: an empirical method of prediction. *Am Jour Sci* 272:870–893
108. Sparks RSJ (1997) Causes and consequences of pressurisation in lava dome eruptions. *Earth Planet Sci Lett* 150:177–189
109. Sparks RSJ, Bursik MI, Carey SN, Gilbert JS, Glaze LS, Sigurdson H, Woods AW (1997) *Volcanic plumes*. Wiley, Chichester. ISBN 0-471-93901-3
110. Spera FJ, Borgia A, Strimple J, Feigenson M (1988) Rheology of melts and magmatic suspensions I. Design and calibration of a concentric cylinder viscometer with application to rhyolitic magma. *J Geophys Res* 93:10273–10294
111. Spieler O, Alidibirov M, Dingwell DB (2003) Grain-size characteristics of experimental pyroclasts of 1980 Mount St. Helens cryptodome dacite: effects of pressure drop and temperature. *Bull Volcanol* 65:90–104
112. Spieler O, Kennedy B, Kueppers U, Dingwell DB, Scheu B, Taddeucci J (2004) The fragmentation threshold of pyroclastic rocks. *Earth Planet Sci Lett* 226:139–148
113. Sugioka I, Bursik M (1995) Explosive fragmentation of erupting magma. *Nature* 373(6516):689–692
114. Suzuki T (1983) A theoretical model for dispersion of tephra. In: Shimozuru D, Yokoyama I (eds) *Arc volcanism, physics and tectonics*. Terra Scientific publishing company Terrapub, Tokyo, pp 95–113
115. Taddeucci J, Spieler O, Kennedy B, Pompilio M, Dingwell DB, Scarlato P (2004) Experimental and analytical modeling of basaltic ash explosions at Mount Etna, Italy, 2001. *J Geophys Res* 109:B08203. doi:10.1029/2003JB002952
116. Taddeucci J, Spieler O, Ichihara M, Dingwell DB, Scarlato P (2006) Flow and fracturing of viscoelastic media under diffusion-driven bubble growth: an analogue experiment for eruptive volcanic conduits. *Earth Planet Sci Lett* 243:771–785
117. Taddeucci J, Scarlato P, Andronico D, Cristaldi A, Büttner R, Zimanowski B, Küppers U (2007) Advances in the study of volcanic ash. *Eos Trans AGU* 88(24):253
118. Takeuchi S, Nakashima S, Tomiya A (2008) Permeability measurements of natural and experimental volcanic materials with a simple permeameter: Toward an understanding of magmatic degassing process. *J Volcanol Geotherm Res* doi:10.1016/j.jvolgeores.2008.05.010
119. Takeuchi S, Nakashima S, Tomiya A, Shinohara H (2005) Experimental constraints on the low gas permeability of vesicular magma during decompression. *Geophys Res Lett* 32:L10312
120. Textor C, Graf HF, Longo A, Neri A, Ongaro TE, Papale P, Timmerck C, Ernst GG (2005) Numerical simulation of explosive volcanic eruptions from the conduit flow to global atmospheric scales. *Ann Geophys* 48(4–5):817–842
121. Trigila R, Battaglia M, Manga M (2007) An experimental facility for investigating hydromagmatic eruptions at high-pressure and high-temperature with application to the importance of magma porosity for magma-water interaction. *Bull Volcano* 69:365–372
122. Tuffen H, Dingwell D (2005) Fault textures in volcanic conduits: evidence for seismic trigger mechanisms during silicic eruptions. *Bull Volcanol* 67:370–387
123. Walker GPL, Wilson L, Howell ELG (1971) Explosive volcanic eruptions. 1. Rate of fall of pyroclasts. *Geophys J Royal Astron Soc* 22(4):377–383
124. Watson EB (1994) Diffusion in volatile-bearing magmas. In: Carroll MR, Holloway JR (eds) *Volatiles in magmas*.
125. Wilson L, Huang TC (1979) Influence of shape on the atmospheric settling velocity of volcanic ash particles. *Earth Planet Sci Lett* 44(2):311–324
126. Wohletz KH (1983) Mechanisms of hydrovolcanic pyroclast formation: grain-size, scanning electron microscopy, and experimental studies. *J Volcanol Geotherm Res* 17(1–4):31–63
127. Woolley AR, Church AA (2005) Extrusive carbonatites: a brief review. *Lithos* 85(1–4):1–14
128. Yokoyama I (2005) Growth rates of lava domes with respect to viscosity of magmas. *Ann Geophys* 48(6):957–971
129. Zhang YX (1998) Experimental simulations of gas-driven eruptions: kinetics of bubble growth and effect of geometry. *Bull Volcanol* 59(4):281–290
130. Zhang Y (1999) A criterion for the fragmentation of bubbly magma based on brittle failure theory. *Nature* 402:648–650
131. Zhang Y, Behrens H (2000) H<sub>2</sub>O diffusion in rhyolitic melts and glasses. *Chem Geol* 169:243–26
132. Zhang YX, Sturtevant B, Stolper EM (1997) Dynamics of gas-driven eruptions: experimental simulations using CO<sub>2</sub>-H<sub>2</sub>O-polymer system. *J Geophys Res Solid Earth* 102(B2):3077–3096
133. Zhang Y, Xu Z, Liu Y (2003) Viscosity of hydrous rhyolitic melts inferred from kinetic experiments: a new viscosity model. *Am Miner* 88:1741–1752
134. Zimanowski B, Lorenz V, Frohlich G (1986) Experiments on phreatomagmatic explosions with silicate and carbonatitic melts. *J Volcanol Geotherm Res* 30(1–2):149–153
135. Zimanowski B, Büttner R, Lorenz V, Häfele H-G (1997) Fragmentation of basaltic melt in the course of explosive volcanism. *J Geophys Res* 102(B1):803–814

## Books and Reviews

- Barnes HA (1997) Thixotropy – a review. *J Non-Newtonian Fluid Mech* 70:1–33
- Barnes HA (1999) The yield strength – a review or ‘πανταρει’ everything flows? *J Non-Newtonian Fluid Mech* 81:133–178
- Chhabra RP (2006) *Bubbles, drops, and particles in non-Newtonian fluids*. CRC Press, London, p 771. ISBN 0824723295
- Chouet B (2003) *Volcano seismology*. Pure and Applied Geophysics 160:739–788
- De Kee D, Chhabra RP (2002) *Transport processes in bubbles, drops, and particles*. Taylor and Francis, p 352. ISBN 1560329068
- Fan L-S, Zhu C (1998) *Principles of gas-solid flows*. Cambridge University Press, Cambridge, p 575. ISBN 0521581486
- Gilbert JS, Sparks RSJ (1998) *The physics of explosive volcanic eruptions*. The Geological Society, London. ISBN 1-86239-020-7

- Gonnermann HM, Manga M (2007) The fluid mechanics inside a volcano. *Ann Rev Fluid Mech* 39:321–356
- Henderson G, Calas G, Stebbins J (eds) (2006) Glasses and melts: linking geochemistry and materials science. *Elements* 2(5):257–320. ISSN 1811–5209
- Kaminski E, Tait S, Carazzo G (2005) Turbulent entrainment in jets with arbitrary buoyancy. *J Fluid Mech* 526:361–376
- Mather TA, Harrison RG (2006) Electrification of volcanic plumes surveys. *Geophysics* 77(4):387–432
- Moretti R, Richet P, Stebbins JF (2006) Physics, chemistry and rheology of silicate melts and glasses. *Chem Geol* 229(1–3):1–226
- Parfitt L, Wilson L (2008) *Fundamentals of physical volcanology*. Blackwell, p 256. ISBN 0632054433
- Sigurdsson H, Houghton B, McNutt SR, Rymer H, Stix J (eds) (2000) *Encyclopaedia of volcanoes*. Academic Press. ISBN 0-12-643140-X
- Tong L-S, Tang YS (1997) *Boiling heat transfer and two-phase flow*. Taylor and Francis, p 542. ISBN 1560324856
- Webb S (1997) Silicate melts: relaxation, rheology, and the glass transition. *Rev Geophys* 35(2):191–218
- Wood CA (1980) Morphometric evolution of cinder cones. *J Volcanol Geotherm Res* 7:387–413



## Volcanic Eruptions: Cyclicity During Lava Dome Growth

OLEG MELNIK<sup>1,2</sup>, R. STEPHEN J. SPARKS<sup>2</sup>,  
ANTONIO COSTA<sup>2,3</sup>, ALEXEI A. BARMIN<sup>1</sup>

<sup>1</sup> Institute of Mechanics, Moscow State University,  
Moscow, Russia

<sup>2</sup> Earth Science Department, University of Bristol,  
Bristol, UK

<sup>3</sup> Istituto Nazionale di Geofisica e Vulcanologia,  
Naples, Italy

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Dynamics of Magma Ascent During Extrusive Eruptions](#)

[Short-Term Cycles](#)

[Long-Term Cycles](#)

[Future Directions](#)

[Acknowledgments](#)

[Bibliography](#)

### Glossary

**Andesite** Magma or volcanic rock is characterized by intermediate SiO<sub>2</sub> concentration. Andesite magmas have rheological properties that are intermediate between basalt and rhyolite magmas. Silica content in andesites ranges from approximately 52 to 66 weight percent. Common minerals in andesite include plagioclase, amphibole and pyroxene. Andesite is typically erupted at temperatures between 800 to 1000°C. Andesite is particularly common in subduction zones, where tectonic plates converge and water is introduced into the mantle.

**Basalt** Magma or volcanic rock contains not more than about 52% SiO<sub>2</sub> by weight. Basaltic magmas have a low viscosity. Volcanic gases can escape easily without generating high eruption columns. Basalt is typically erupted at temperatures between 1100 to 1250°C. Basalt flows cover about 70% of the Earth's surface and huge areas of the terrestrial planets and so are the most important of all crustal igneous rocks.

**Bingham liquid** is a fluid that does not flow in response to an applied stress until a critical yield stress is reached. Above the critical yield stress, strain rate is proportional to the applied stress, as in a Newtonian fluid.

**Bubbly flow** A multi-phase flow regime, in which the gas phase appears as bubbles suspended in a continuous liquid phase.

**Conduit** A channel, through which magma flows towards the Earth's surface. Volcanic conduits can commonly be approximately cylindrical and typically a few 10's meters across or bounded by near parallel sides in a magma-filled fracture. Conduits can be vertical or inclined.

**Crystallization** Conversion, partial or total, of a silicate melt into crystals during solidification of magma.

**Degassing n. (degas v.)** The process by which volatiles that are dissolved in silicate melts come out of solution in the form of bubbles. Open- and closed-system degassing can be distinguished. In the former, volatiles can be lost or gained by the system. In the latter, the total amount of volatiles in the bubbles and in solution in the magma is conserved.

**Differentiation** The process of changing the chemical composition of magma by processes of crystallization accompanied by separation melts from crystals.

**Dome** A steep-sided, commonly bulbous extrusion of lava or shallow intrusion (cryptodome). Domes are commonly, but not exclusively, composed of SiO<sub>2</sub>-rich magmas. In dome-forming eruptions the erupted magma is so viscous, or the discharge rate so slow, that lava accumulates very close to the vent region, rather than flowing away. Pyroclastic flows can be generated by collapse of lava domes. Recent eruptions producing lava domes include the 1995–2006 eruption of the Soufrière Hills volcano, Montserrat, and the 2004–2006 eruption of Mount St. Helens, USA.

**Dyke** A sheet-like igneous intrusion, commonly vertical or near vertical, that cuts across pre-existing, older, geological structures. During magmatism, dykes transport magma toward the surface or laterally in fracture-like conduits. In the geologic record, dykes are preserved as sheet-like bodies of igneous rocks.

**Explosive eruption** A volcanic eruption in which gas expansion tears the magma into numerous fragments with a wide range of sizes. The mixture of gas and entrained fragments flows upward and outward from volcanic vents at high speed into the atmosphere. Depending on the volume of erupted material, eruption intensity and sustainability, explosive eruptions are classified as Strombolian, Vulcanian, sub-Plinian, Plinian or Mega-Plinian; this order is approximately in the order of increasing intensity. Strombolian and Vulcanian eruptions involve very short-lived explosions.

**Extrusive flow or eruption** A non-explosive (non-pyro-

clastic) magma flow from a volcanic conduit during a lava dome-building eruption or lava flow.

**Mafic** Magma, lava, or tephra with silica concentrations of approximately  $\text{SiO}_2 < 55\%$ .

**Magma** Molten rock that consists of up three components: liquid silicate melt, suspended crystalline solids, and gas bubbles. It is the raw material of all volcanic processes. Silicate magmas are the most common magma type and consist of long, polymeric chains and rings of Si–O tetrahedra, between which are located cations (e. g.  $\text{Ca}^{2+}$ ,  $\text{Mg}^{2+}$ ,  $\text{Fe}^{2+}$ , and  $\text{Na}^+$ ). Anions (e. g.  $\text{OH}^-$ ,  $\text{F}^-$ ,  $\text{Cl}^-$ , and  $\text{S}^-$ ) can substitute for the oxygen in the silicate framework. The greater the silica ( $\text{SiO}_2$ ) content of the magma, the more chains and rings of silicate tetrahedra there are to impede each other and hence the viscosity of the magma increases. The pressure regime and composition of the magma control the minerals that nucleate and crystallize from a magma when it cools or degasses.

**Magma chamber** A subsurface volume within which magma accumulates, differentiates and crystallizes. Igneous intrusions can constrain the form and size of some magma chambers, but in general the shape and volume of magma chambers beneath active volcanoes are poorly known. Magma reservoir is an equivalent term.

**Melt** Liquid part of magma. Melts (usually silicate) contain variable amounts of dissolved volatiles. The primary volatiles are usually water and carbon dioxide.

**Newtonian liquid** A liquid for which the strain rate is proportional to the applied stress. The proportionality coefficient is called the viscosity.

**Microlite** Crystal with dimensions less than 100  $\mu\text{m}$ . Usually microlites crystallize at shallow levels of magmatic system.

**Phenocryst** Crystal with dimensions larger than 100  $\mu\text{m}$ . Usually phenocrysts grow in magmatic reservoirs prior to an eruption and or are entrained by magma in the chamber.

**Pyroclastic flow or surge** A gas-particle flow of pyroclasts suspended in a mixture of hot air, magmatic gas, and fine ash. The flow originates by the gravitational collapse of a dense, turbulent explosive eruption column at the source vent, or by dome collapse, and moves down-slope as a coherent flow. Pyroclastic flows and surges are distinguished by particle concentration in the flow, surges being more dilute. Variations in particle concentration result in differences in the deposits left by flows and surges.

**Silicic** Magma, lava, or tephra with silica concentrations of approximately  $\text{SiO}_2 > 55\%$ . The magmas are com-

monly rich in Al, Na- and K- bearing minerals. Silicic magmas are typically very viscous and can have high volatile contents. Rhyolite is an example of a silicic magma.

**Volatile** A component in a magmatic melt which can be partitioned in the gas phase in significant amounts during some stage of magma history. The most common volatile in magmas is water vapor  $\text{H}_2\text{O}$ , but there are commonly also significant quantities of  $\text{CO}_2$ ,  $\text{SO}_2$  and halogens.

### Definition of the Subject

We consider the process of slow extrusion of very viscous magma that forms lava domes. Dome-building eruptions are commonly associated with hazardous phenomena, including pyroclastic flows generated by dome collapses, explosive eruptions and volcanic blasts. These eruptions commonly display fairly regular alternations between periods of high and low or no activity with time scales from hours to years. Usually hazardous phenomena are associated with periods of high magma discharge rate, thus, understanding the causes of pulsatory activity during extrusive eruptions is an important step towards forecasting volcanic behavior, especially the transition to explosive activity when magma discharge rate increases by a few orders of magnitude. In recent years the risks have increased because the population density in the vicinity of many active volcanoes has increased.

### Introduction

Many volcanic eruptions involve the formation of lava domes, which are extrusions of very viscous, degassed magmas. The magma is so viscous that it accumulates close to the vent. Extrusion of lava domes is a slow and long-lived process, and can continue for many years or even decades [71,83,85]. Typical horizontal dimensions of lava domes are several hundred meters, heights are of an order of tens to several hundred meters, and volumes several million to hundreds of million cubic meters. Typical magma discharge rates (measured as the increase of dome volume with time in dense rock equivalent (DRE)) can reach up to 20–40  $\text{m}^3/\text{s}$ , but are usually below 10  $\text{m}^3/\text{s}$  [83].

Dome-building eruptions are commonly associated with hazardous phenomena, including pyroclastic flows and tsunamis generated by dome collapses, explosive eruptions and volcanic blasts. Dome-building eruptions can also contribute to edifice instability and sector collapse, as occurred on Montserrat on 26 December 1997 [87]. Lava dome activity can sometimes precede or follow major explosive eruptions; the eruption of Pinatu-

bo, Philippines (1991) is an example of the former [37], and the eruption of Mount St. Helens, USA (1980–1986) is an example of the latter [89].

Several lava dome eruptions have been documented in detail and show quite complex behaviors. Substantial fluctuations in magma discharge rate have been documented. In some cases these fluctuations can be quite regular (nearly periodic), as in the extrusion of lava in 1980–1982 on Mount St. Helens [89] and in the 1922–2002 activity of the Santiaguito lava dome, Guatemala [35]. In these cases, periods of high magma discharge rate alternate with longer periods of low magma discharge rate or no extrusion. In some volcanoes, such as Shiveluch, Kamchatka, the intervals of no extrusion are so long compared with the periods of dome growth that the episodes of dome growth have been described as separate eruptions of the volcano rather than episodes of the same eruption. Other dome-building activity can be nearly continuous and relatively steady, as observed at Mount St. Helens in 1983 [89] and at the Soufrière Hills Volcano, Montserrat between November 1999 and July 2003. In yet other cases the behavior can be more complex with quite sudden changes in magma discharge rate, which cannot be related to any well-defined regularity or pattern (e.g. Lascar volcano, Chile, [57]).

Pauses during lava dome-building eruptions are quite common. For example, at Mount St. Helens there were 9 pulses of dome growth with a period of  $\sim 74$  days, a duration of 1–7 days and no growth in between [89]. The Soufrière Hills Volcano Montserrat experienced a long (20 months) pause in extrusion after the first episode of growth [72]. On Shiveluch volcano in Kamchatka episodes of dome growth occurred in 1980, 1993 and 2000, following a major explosion in 1964 [28]. Each episode of dome growth began with magma discharge rate increasing over the first few weeks to a peak of 8–15 m<sup>3</sup>/s, with a gradual decline in magma discharge rate over the following year. In between the episodes very minimal activity was recorded.

Fluctuations in magma discharge rate have been documented on a variety of time-scales from both qualitative and quantitative observations. Several lava dome eruptions are characterized by extrusion of multiple lobes and flow units [68,94]. In the case of the Soufrière Hills Volcano, extrusion of shear lobes can be related to spurts in discharge rate and is associated with other geophysical changes, such as onset of seismic swarms and marked changes in temporal patterns of ground tilt [90,91,94]. These spurts in discharge rate have been fairly regular for substantial periods, occurring every 6 to 7 weeks over a 7 month period in 1997 [21,87,91]. These spurts are com-

monly associated with large dome collapses and pyroclastic flows and, in some cases, with the onset of periods of repetitive Vulcanian explosions [14,26]. Consequently the recognition of this pattern has become significant for forecasting activity for hazard assessment purposes. In the Soufrière Hills Volcano and Mount Pinatubo much shorter fluctuations in magma discharge rate have been recognized from cyclic variations in seismicity, ground tilt, gas fluxes and rock-fall activity [23,91,93]. This cyclic activity has typical periods in the range of 4 to 36 hours. Cyclic activity has been attributed to cycles of gas pressurization and depressurization with surges in dome growth related to degassing, rheological stiffening and stick-slip behavior [23,49,61,91,98].

Dome eruptions can show transitions to explosive activity, which sometimes can be linked to spurts in magma discharge rate. For example, in 1980, periodic episodes of lava dome extrusion on Mount St. Helens were initiated by explosive eruptions, which partly destroyed the dome that had been extruded in each previous extrusion episode [89]. At Unzen Volcano, Japan a single Vulcanian explosive eruption occurred in June 1991 when the magma discharge rate was at its highest [68]. At the Soufrière Hills Volcano, repetitive series of Vulcanian explosions have occurred following large dome collapses in periods when magma discharge rates were the highest of the eruption [26,88]. In the case of Lascar Volcano, Chile, an intense Plinian explosive eruption occurred on 18 and 19 April, 1993, after nine years of dome extrusion and occasional short-lived Vulcanian explosions [57].

Lava dome eruptions require magma with special physical properties. In order to produce a lava dome rather than a lava flow, the viscosity of the magma must be extremely high so that the lava cannot flow easily from the vent. High viscosity is a consequence of factors such as relatively low temperature (typically 750–900°C), melt compositions rich in network-forming components (principally Si and Al) efficient gas loss during magma decompression, and crystallization as a response to cooling and degassing. Viscosities of silica-rich magmas, such as rhyolites and some andesites, are increased by several orders of magnitude by the loss of dissolved water during decompression. Many, but not all, domes also have high crystal content (up to 60 to 95 vol%), with crystallization being triggered mostly by degassing [10,86]. In order to avoid fragmentation that leads to an explosive eruption, magma must have lost gas during ascent. Consider, for example, a magma at 150 MPa containing 5wt% of dissolved water decompressed to atmospheric pressure. Without gas loss the volume fraction of bubbles, will be more than 99%. Typical dome rock contains less than 20 vol% of bub-

bles, although there is evidence that magma at depth can be more bubble-rich (e. g., [13,74]). On the other hand, very commonly there is no change in temperature or bulk magma composition in the products of explosive and extrusive eruptions for a particular volcano. This suggests that the properties of magma that are conducive to the formation of lava domes are controlled by physico-chemical transformations that occurred during magma ascent to the surface.

Two other important factors that influence whether lava domes or flows form are topography and discharge rate. The same magma can form a dome if the discharge rate is low, and a lava flow if the rate is high [29,92]. The discharge rate is controlled by overall conduit resistance that is a function of viscosity, conduit size and shape, and driving pressure (the difference between chamber pressure and atmospheric pressure). Additionally the same magma can form a dome on low slopes, such as a flat crater (e. g. the mafic andesite dome of the Soufrière Volcano, St. Vincent; [40]) and a lava flow on steep slopes.

Prior to an eruption, magma is usually stored in a shallow crustal reservoir called a magma chamber. For several volcanoes magma chambers can be detected and characterized by earthquake locations, seismic tomography, petrology or interpretation of ground deformation data [55]. Typical depths of magma chambers range from a few kilometers to tens of kilometers. Volumes range from less than one to several thousand km<sup>3</sup> [55], but are usually less than a hundred km<sup>3</sup>. Magma chambers are connected to the surface by magma pathways called conduits. There is evidence that the conduits that feed lava dome eruptions can be both dykes or cylindrical. Dykes of a few meters width are commonly observed in the interior of eroded andesite volcanoes. Dyke feeders to lava domes have been intersected by drilling at Inyo crater, California, USA [56] and at Mount Unzen [67]. Geophysical studies point to dyke feeders; for example fault-plane solutions of shallow volcano-tectonic earthquakes indicate pressure fluctuations in dykes [75,76]. Deformation data at Unzen, combined with structural analysis, indicate that the 1991–1995 dome was fed by a dyke [68]. Dykes are also the only viable mechanism of developing a pathway through brittle crust from a deep magma chamber to the surface in the initial stages of an eruption [50,77].

Cylindrical conduits commonly develop during lava dome eruptions. The early stages of lava dome eruptions frequently involve phreatic and phreatomagmatic explosions that create near surface craters and cylindrical conduits [12,73,87,89,96,99]. These explosions are usually attributed to interaction of magma rising along a dyke with ground water. Cylindrical conduits formed by explosions

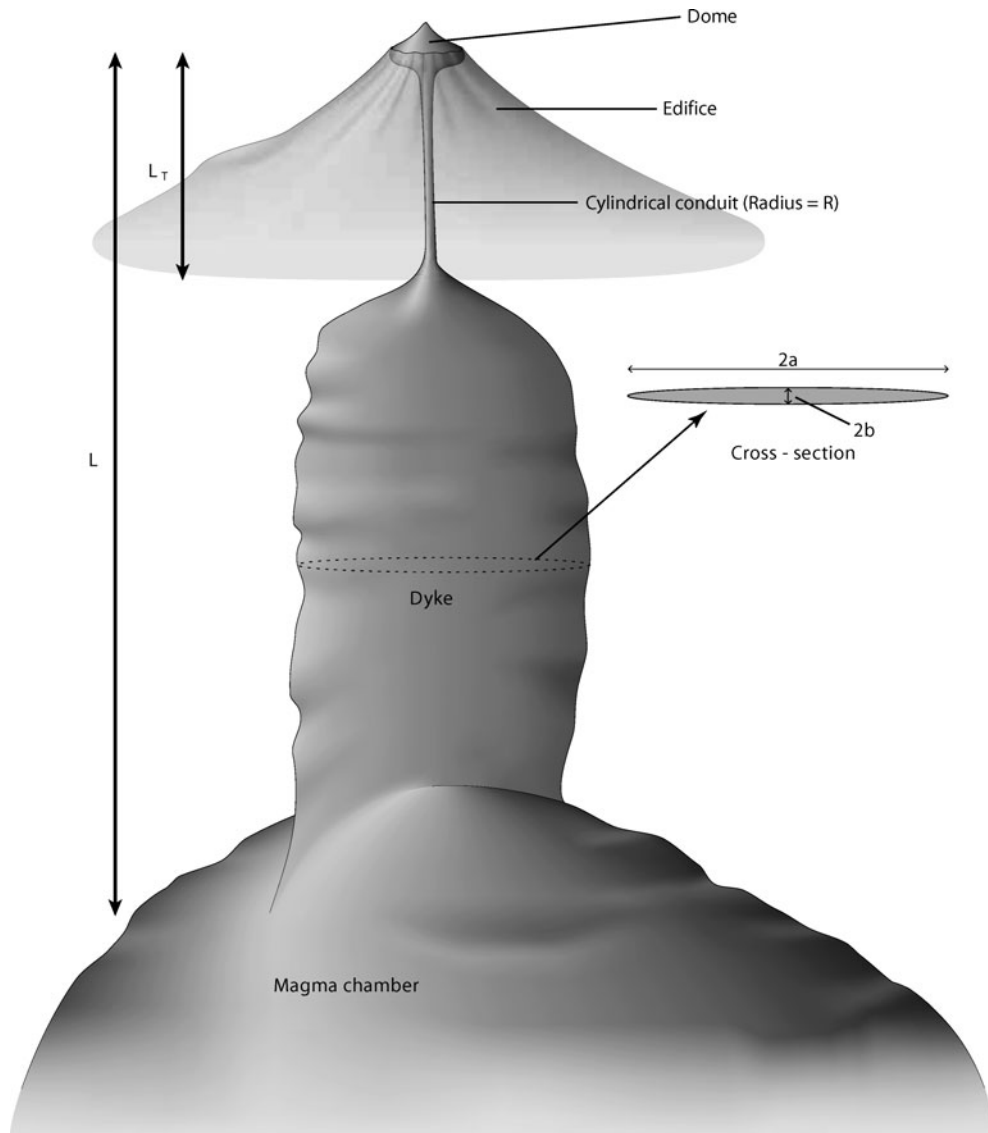
are confined to relatively shallow parts of the crust, probably of order hundreds of meters depth and < 1 km, as indicated by mineralogical studies [73]. Examples of such initial conduit forming activity include Mount Usu (Japan) Mt. St. Helens, and Soufrière Hills Volcano [12,87,99]. Many lava dome eruptions are also characterized by Vulcanian, sub-Plinian and even Plinian explosive eruptions. Examples include Mount Unzen, Mount St. Helens, Santiaguito and Soufrière Hills Volcano [68,87,89,96]. Here the fragmentation front may reach to depths of several kilometers [54] with the possibility of cylindrical conduit development due to severe underpressurization and mechanical disruption of conduit wallrocks. Subsequently domes can be preferentially fed along the cylindrical conduits created by earlier explosive activity. On the Soufrière Hills Volcano, early dome growth was characterized by extrusion of spines with nearly cylindrical shape [87].

Observations of magma discharge rate variations on a variety of time-scales highlight the need to understand the underlying dynamic controls. Research has increasingly focused on modeling studies of conduit flow dynamics during lava-dome eruptions. We will restrict our discussions here to mechanisms that lead to cyclic and quasi-periodic fluctuations in magma discharge rate on various timescales, mainly focusing on long-term cycles. Issues concerning the transition between explosive and extrusive activity are discussed in detail in [45,80,81,97] and in the special volume of *Journal of Volcanology and Geothermal Research* dedicated to modeling of explosive eruptions [79]. Several papers consider the processes that occur at the surface and relate dome morphology and dimensions with controlling parameters.

Combined theoretical, experimental and geological studies identify four main types of dome: spiny, lobate, platy, and axisymmetric [5,29,30]. These types of dome reflect different regimes which are controlled by discharge rates, cooling rates and yield strength, and the viscosity of the dome-forming material. In recent years mathematical modeling has been used to semi-quantitatively describe spreading of lava domes, including models based on the thin layer approximation [1,2] and fully 2D simulations of lava dome growth which account for visco-elastic and -plastic rheologies [32,33].

### Dynamics of Magma Ascent During Extrusive Eruptions

In order to understand the causes of cyclic behavior during extrusive eruptions, first we need to consider the underlying dynamics of volcanic systems, and discuss physical and chemical transitions during magma ascent.



Volcanic Eruptions: Cyclicity During Lava Dome Growth, Figure 1

Schematic view of the volcanic system. In the upper part the conduit is cylindrical with a radius  $R$ . A transition from the cylinder to a dyke occurs at depth  $L_T$ . The length scale for the transition from cylinder to dyke is  $w_T$ . The dyke has an elliptical cross-section with semi-axis lengths  $a_0$  and  $b_0$ . The chamber is located a depth  $L$ . In the text we used also the following auxiliary variables:  $D = 2R$  for conduit diameter,  $L_d = L - L_T$  for the dyke vertical length,  $W_d = 2a_0$  for the dyke width, and  $H_d = 2b_0$  for the dyke thickness. After [20]

The physical framework for the model of a volcanic system is shown in Fig. 1. Magma is stored in a chamber at depth  $L$ , with a chamber pressure  $P_{ch}$  that is higher than hydrostatic pressure of magma column and drives magma ascent. Magma contains silicate melt, crystals and dissolved and possibly exsolved volatiles. During ascent of the magma up the conduit the pressure decreases and volatiles exsolve forming bubbles. As the bubble concentration be-

comes substantial, bubble coalescence take place and permeability develops [27,45], allowing gas to escape from ascending magma, both in vertical (through the magma) and horizontal (to conduit wallrocks) directions. If magma ascends slowly, gas escape results in a significant reduction of the volume fraction of bubbles; such a process is termed open system degassing. In comparison closed system degassing is characterized by a negligible gas escape.

Reduction in bubble content, combined with relatively low gas pressures and efficient decoupling of the gas and melt phases, prevents magma fragmentation and, thus, development of explosive eruption [59,60,82].

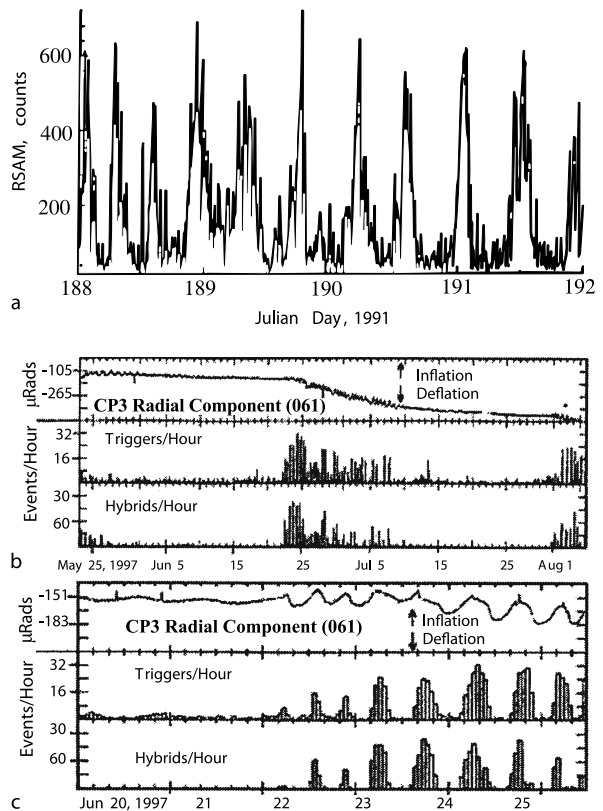
Due to typical low ascent velocities (from millimeters to a few centimeters per second) magma ascent times to the Earth's surface from the magma chamber range from a several hours to many weeks. These ascent times are often comparable with the times that are required for crystals to grow significantly and for heat exchange between magma and wallrocks. The main driving force for crystallization is related to exsolution of volatiles from the magma, leading to increase in the liquidus temperature  $T_L$ , and development of magma undercooling  $\Delta T = T_L - T$  [10]. Crystallization leads to release of latent heat, and magma temperature can increase with respect to the initial temperature [6]. As a consequence of increasing crystal content, magma viscosity increases by several orders of magnitude [16,20,21] and magma becomes a non-Newtonian fluid [78]. As will be shown later, crystallization induced by degassing can become a key process in causing variable flow rates.

Due to the long duration of extrusive eruptions, the magma chamber can be replenished with significant amounts of magma from underlying sources [41,65]. Replenishment can lead to pressure build-up in the magma chamber, volatile and heat exchange between host and new magmas. The composition of the magma can also change over time, due to differentiation, crustal rock assimilation, or magma mixing. Thus, any model that explains magma ascent dynamics needs to deal with many complexities. Of course there is no single model that can take into account all physical processes in a volcanic system. Additional complications arise from the fact that the physical properties of magma at high crystal contents, such as rheology or crystal growth kinetics and geometry of volcanic systems, are typically poorly constrained. Several issues regarding the dynamics of multiphase systems have not been resolved theoretically, especially for cases where the volume fractions of components of the multiphase system are comparable.

Below we will present a review of existing models that treat cyclic behavior during extrusive eruptions on different timescales.

### Short-Term Cycles

Cyclic patterns of seismicity, ground deformation and volcanic activity (Fig. 2 from [23]) have been documented at Mount Pinatubo, Philippines, in 1991 [37] and Soufrière Hills volcano, Montserrat, British West Indies, in 1996–



**Volcanic Eruptions: Cyclicity During Lava Dome Growth, Figure 2** Examples of cyclic behavior in a lava dome eruption. RSAM records from Mt. Pinatubo, Philippines following its climatic eruption in June 1991 (a, after [23]). Radial tilt, triggered earthquakes, and hybrid earthquakes for 17 May to 6 August 1997, at Soufrière Hills Volcano (b, after [21,90]). Parts of three 5–7 week cycles are shown, with each cycle showing high amplitude tilt and seismicity pulsations with short time scale that last for several weeks after the start of the cycle. The onset of a cycle is rapid, as detailed in c for the cycle initiating on 22 June 1997

1997 [90,91]. At Soufrière Hills, periodicity in seismicity and tilt ranged from  $\sim 4$  to 30 h, and the oscillations in both records continued for weeks. Cyclic behavior was first observed in the seismicity (RSAM) records beginning in July 1996, when the record of dome growth constrained the average supply rate to between 2 and 3  $\text{m}^3/\text{s}$  [88]. The oscillations in the RSAM records initially had low amplitudes, and no tilt-measurement station was close enough to the vent to detect any pressure oscillations in the conduit. By August 1996, RSAM records showed strong oscillatory seismicity at dome-growth rates between 3 and 4  $\text{m}^3/\text{s}$ . Tilt data, taken close enough to the vent (i.e., Chances Peak [90]) to be sensitive to conduit pressure oscillations, are only available for February 1997 and May–August 1997. In the latter period dome growth rate in-

crease from  $\sim 5 \text{ m}^3/\text{s}$  in May to between 6 and  $10 \text{ m}^3/\text{s}$  in August. Both near-vent tilt and RSAM displayed oscillatory behavior during this period and were strongly correlated in time. Similar RSAM oscillations having periods of 7 to 10 h were observed at Mount Pinatubo following the climactic eruption in 1991 [37]. At both volcanoes, oscillation periods were observed that do not fit any multiple of Earth or ocean tides.

The cyclic activity at both Pinatubo and Soufrière Hills Volcano are strongly correlated with eruptive behavior and other geophysical phenomena. In the Pinatubo case and on Soufrière Hills Volcano in August, September and October 1997, the cycles were linked to short-lived volcanic explosions. In the case of the Soufrière Hills Volcano, explosions in August 1997 occurred at the peak in the tilt. The peak in tilt also marked the onset of episodes of increased rock falls [7,8] and [91] is attributed to increased magma discharge rates.  $\text{SO}_2$  flux data show that the cycles are linked to surges in gas release, which reach a peak about an hour after the tilt peak [93]. Green et al. [31] have shown that several families of near identical long period earthquakes occur during the tilt cycle, starting at the inflexion point on the up-cycle and finishing before the inflexion point on the down-cycle.

Several models [23,49,70,98,98] have been proposed to explain the observed cyclicity. In these models [23,98] the conduit is divided into two parts. Magma is assumed to be forced into the lower part of volcanic conduit at a constant rate. In Denlinger and Hoblitt [23] magma in the lower part of the conduit is assumed to be compressible. In Wylie et al. [98] the magma is incompressible but the cylindrical conduit is allowed to expand elastically. In both models the lower part of the conduit, therefore, acts like a capacitor that allows magma to be stored temporally in order to release it during the intense phase of the eruption. In the upper part of the conduit friction is dependent on magma discharge rate, with a decrease in friction resulting in an increase in discharge rate, over a certain range of discharge rates. In Denlinger and Hoblitt [23], when magma discharge rate reaches a critical value, magma detaches from the conduit walls and a stick-slip transition occurs. Rapid motion of magma leads to depressurization of the conduit and a consequent decrease in discharge rate, until at another critical value the magma again sticks to the walls. Pressure starts to increase again due to influx of new magma into the conduit. On a pressure-discharge diagram the path of eruption is represented by a hysteresis loop. In Wylie et al. [98] the friction is controlled by volatile-dependent viscosity. Volatile exsolution delay is controlled by diffusion. When magma ascends rapidly, volatiles have no time to exsolve and viscosity remains low. Depressur-

ization of the upper part of the conduit leads to a decrease in magma discharge rate and an increase in viscosity due to more intense volatile exsolution.

In Neuberger et al. [70] a steady 2D conduit flow model was developed. The full set of Navier–Stokes equations for a compressible fluid with variable viscosity was solved by means of a finite element code. Below some critical depth the flow was considered to be viscous and Newtonian, with a no-slip boundary condition at the wall. Above this depth a plug develops, with a wall boundary condition of frictional slip. The slip criterion was based on the assumption that the shear stress inside the magma overcomes some critical value. Simulations reveal that slip occurs in the shallow part of the conduit, in good agreement with locations of long-period volcanic earthquakes for the Soufrière Hills Volcano, [31]. However, some parameters used in the simulations (low crystallinity, less than 30% and high discharge rate, more than  $100 \text{ m}^3/\text{s}$ ) are inconsistent with observations.

Lensky et al. [49] developed the stick-slip model by incorporating degassing from supersaturated magma together with a sticking plug. Gas diffuses into the magma, which cannot expand due to the presence of a sticking plug, resulting in a build up of pressure. Eventually the pressure exceeds the strength of the plug, which fails in stick-slip motion and the pressure is relieved. The magma sticks again when the pressure falls below the dynamic friction value. In this model the time scale of the cycles is controlled by gas diffusion. The influence of permeable gas loss, crystallization and elastic expansion of the conduit on the period of pulsations was studied.

A shorter timescale of order of minutes was investigated in Iverson et al. [43] in relation to repetitive seismic events during the 2004–2006 eruption of Mount St. Helens. The flow dynamics is controlled by the presence of a solid plug that is pushed by a Newtonian liquid, with the possibility of a stick-slip transition. Inertia of the plug becomes important on such short timescales.

Models to explain the occurrence of Vulcanian explosions have also been developed by Connor et al. [15], Jaquet et al. [44] and Clarke et al. [13]. A statistical model of repose periods between explosions by Connor et al. [15] shows that data fit a log-logistic distribution, consistent with the interaction of two competing processes that decrease and increase gas pressure respectively. Jaquet et al. [44] show the explosion repose period data have a memory. The petrological observations of Clarke et al. [13] on clasts from Vulcanian explosions associated with short-term cycles support a model where pressure builds up beneath a plug by gas diffusion, but is opposed by gas leakage through a permeable magma foam. Although

these models include some of the key processes and have promising explanatory power, they do not consider the development of the magma plug explicitly. This process is considered in [24], where a model of magma ascent with gas escape is proposed.

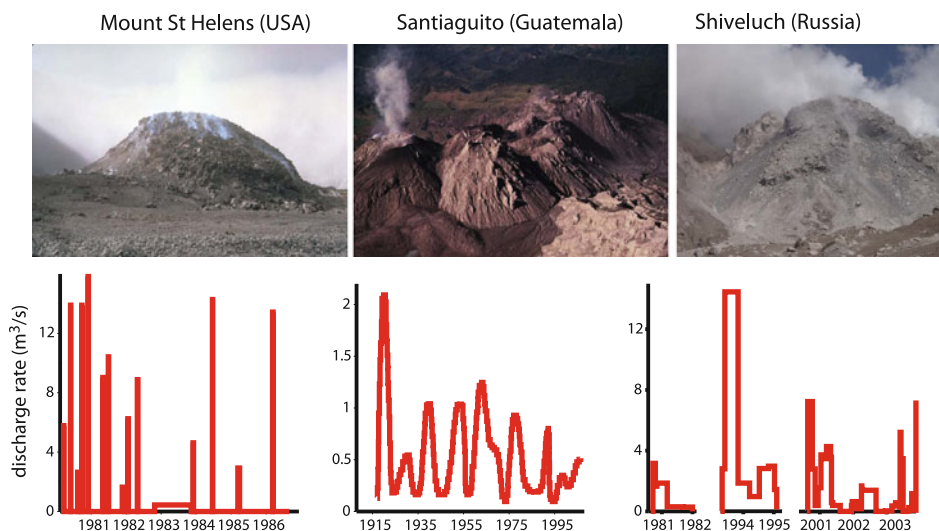
### Long-Term Cycles

Figure 3 shows views of three lava domes (Mount St. Helens, USA, Santiaguito, Guatemala and Shiveluch, Russia). Measurements of magma discharge rate variations with time are presented below [25,35,89]. Behavior at the first two volcanoes is rather regular, whereas at Shiveluch, long repose periods are followed by an initial rapid increase in eruptive activity with subsequent decrease and complete stop of the eruption. Growth of the lava dome at Unzen volcano, Japan, 1991–1995 was similar to Shiveluch [25,68].

There are three types of conceptual models that attempt to explain long term variation in magma discharge rate. Maeda [53], after [42], considers a simple system that contains a spherical magma chamber located in elastic rocks with a cylindrical conduit located in visco-elastic rocks. Magma viscosity is assumed to be constant. The magma chamber is replenished with a time dependent influx rate. The model reproduces discharge rate variation at Unzen volcano by assuming a bell-shaped form of influx rate dependence on time. There are two controversial assumptions in the model. First, the assumption that

the conduit wallrocks are visco-elastic, while the magma chamber wallrocks are purely elastic cannot be justified because near the chamber rock temperature is the highest and it is more reasonable to expect viscous properties for chamber wallrocks rather than for the conduit. The equation that links conduit diameter with magmatic overpressure assumes viscous rock properties up to infinity. If the chamber is located in visco-elastic rocks, oscillations in discharge rate are not possible. Second, in order to obtain reasonable timescales, the rock viscosity must be rather small, of order of  $10^{13}$  Pa s, which is only slightly higher than the typical viscosities of the magma.

Another set of models attribute cyclic behavior to heat exchange between ascending magma and wallrocks, which accounts for temperature dependent viscosity [17,95]. The idea of both models is that magma cools down as it ascends, and heat flux is proportional to the difference between the average temperature of the magma and the temperature of the wallrocks. If magma ascends quickly than heat loss is small in comparison with heat advection. Magma viscosity remains low as a consequence and allows high magma discharge rates. In contrast, when magma ascends slowly it can cool substantially and viscosity increases significantly. Both models suggest that, for a fixed chamber pressure, there can be up to three steady state solutions with markedly different discharge rates. Transition between these steady-state solutions leads to cyclic variations in discharge rate. Whitehead and Helfrich [95] demonstrated the existence of cyclic regimes in exper-



Volcanic Eruptions: Cyclicity During Lava Dome Growth, Figure 3

Observed discharge rate versus time for (a) Mount St. Helens dome growth and (b) Santiaguito volcano and Shiveluch (c). Photos: Mount St. Helens by Lyn Topinka (1985), Santiaguito by Gregg Bluth (2002), Shiveluch by Pavel Plechov (2001)



iments using corn syrup. In application to magma ascent in a volcanic conduit, these models have strong limitations, because a constant wall-rock temperature is assumed. However, as an eruption progresses the wallrocks heat up and heat flux decreases, a condition that makes periodic behavior impossible for long-lived eruptions. For such a long-lived eruption like Santiaguito (started in 1922) wallrocks are expected to be nearly equilibrated in temperature with the magma, and heat losses from magma became small. It is possible that this decrease in heat flux contributes to a slow progressive increase in temperature that is observed on timescales longer than the period of pulsations. For example, magma at Santiaguito becomes progressively less viscous, resulting in a transition from mainly lava dome to lava flow activity.

Models, developed by authors of this manuscript, consider that degassing-induced crystallization is a major controlling process for the long-term cyclicity during lava dome building eruptions. There is increasing evidence that there is a good correlation between magma discharge rate and crystallinity of the magma [10,68]. An increase in crystal content leads to an increase in magma viscosity [16,20,21] and, thus, influences magma ascent dynamics. First, we consider a simplified model of magma ascent in a volcanic conduit that accounts for crystallization and rheological stiffening.

### A Simplified Model

In Barmin et al. [3] the following simplifying assumptions have been made in order to develop a semi-analytical approach to magma ascent dynamics.

1. Magma is incompressible. The density change due to bubble formation and melt crystallization is neglected.
2. Magma is a viscous Newtonian fluid. Viscosity is a step function of crystal content. When the concentration of crystals  $\beta$  reaches a critical value  $\beta_*$  the viscosity of magma increases from value  $\mu_1$  to a higher value  $\mu_2$ . Later on we will consider magma rheology in more detail (see Sect. "Rheology of Crystal-Bearing Magma and Conduit Resistance"), but a sharp increase in viscosity over a narrow range of crystal content has been confirmed experimentally (e. g., [9,48]).
3. Crystal growth rate is constant and no nucleation occurs in the conduit. The model neglects the fact that magma is a complicated multi-component system and its crystallization is controlled by the degree of undercooling (defined as the difference between actual temperature of the magma and its liquidus temperature). Later in the paper a more elaborate model for magma crystallization will be considered.
4. The conduit is a vertical cylindrical pipe. Elastic deformation of the wallrocks is not included in the model. This assumption is valid for a cylindrical shape of the conduit at typical magmatic overpressures, but is violated when the conduit has a fracture shape. Real geometries of volcanic conduits and their inclination can vary significantly with depth.
5. The magma chamber is located in elastic rocks and is fed from below, with a constant influx rate. For some volcanoes, like Santiaguito or Mount St. Helens, average magma discharge rate remained approximately constant during several periods of pulsation. Thus the assumption of constant influx rate is valid. For volcanoes like Mount Unzen or Shiveluch, there is an evidence of pulse-like magma recharge [25,53].

With above simplification the system of equations for unsteady 1D flow is as follows:

$$\frac{\partial}{\partial t} \rho + \frac{\partial}{\partial x} \rho u = 0; \quad \frac{\partial}{\partial t} n + \frac{\partial}{\partial x} n u = 0 \quad (1a)$$

$$\frac{\partial p}{\partial x} = -\rho g - \frac{32\mu u}{\delta^2}; \quad \mu = \begin{cases} \mu_1, & \beta < \beta_* \\ \mu_2, & \beta \geq \beta_* \end{cases} \quad (1b)$$

$$\frac{\partial}{\partial t} \beta + u \frac{\partial \beta}{\partial x} = 4\pi n r^2 \chi = (36\pi n)^{\frac{1}{3}} \beta^{\frac{2}{3}} \chi \quad (1c)$$

Here  $\rho$  is the density of magma,  $u$  is the vertical cross-section averaged ascent velocity,  $n$  is the number density of crystals per unit volume,  $p$  is the pressure,  $g$  is the acceleration due to gravity,  $\delta$  is the conduit diameter,  $\beta$  is the volume concentration of crystals,  $\beta_*$  is a critical concentration of crystals above which the viscosity changes from  $\mu_1$  to  $\mu_2$ ,  $r$  is the crystal radii,  $\chi$  is the linear crystal growth rate, and  $x$  is the vertical coordinate. The first two Eqs. (1a) represent the conservation of mass and the number density of crystals, the second (1b) is the momentum equation with negligible inertia, and the third (1c) is the crystal growth equations with  $\chi = \text{constant}$ . We assume the following boundary conditions for the system (1):

$$x = 0: \quad \frac{dp_{\text{ch}}}{dt} = \frac{\gamma}{V_{\text{ch}}} (Q_{\text{in}} - Q_{\text{out}}); \quad \beta = \beta_{\text{ch}}; \quad n = n_{\text{ch}}$$

$$x = l: \quad p = 0$$

Here  $\gamma$  is the rigidity of the wall-rock of the magma chamber,  $V_{\text{ch}}$  is the chamber volume,  $\beta_{\text{ch}}$  and  $n_{\text{ch}}$  are the crystal concentration and number density of crystals per unit volume in the chamber,  $p_{\text{ch}}$  is the pressure in the chamber,  $l$  is the conduit length,  $Q_{\text{in}}$  is the flux into the chamber and  $Q_{\text{out}} = \pi \delta^2 u / 4$  is the flux out of the chamber into the conduit. Both  $\beta_{\text{ch}}$  and  $n_{\text{ch}}$  are assumed constant. We neglect the influence of variations of the height of the lava dome

on the pressure at the top of the conduit and assume that the pressure there is constant. As the magma is assumed to be incompressible, the pressure at the top of the conduit can be set to zero because the atmospheric pressure is much smaller than magma chamber pressure.

From the mass conservation equation for the case of constant magma density,  $u = u(t)$  and  $n = n_{ch}$  everywhere. Equations (1) can be integrated and transformed from partial differential equations to a set of ordinary differential equations with state-dependent delay representing a “memory” effect on crystal concentration (see [3] for details).

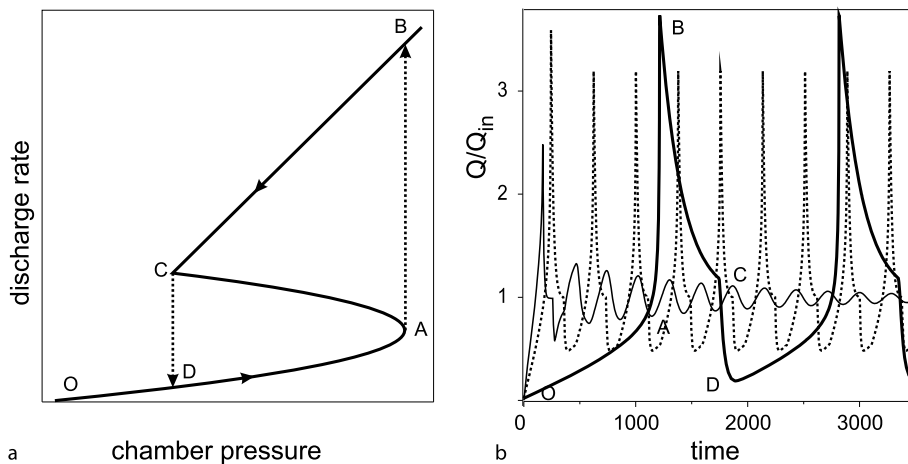
**Results and Applications**

The general steady-state solution for magma ascent velocity variations with chamber pressure is shown in Fig. 4a. Solutions at high magma ascent velocities, when the critical concentration of crystals is not reached inside the conduit, result in a straight line (CB), which is the same as for the classical Poiseuille solution for a fluid with constant viscosity. At low ascent velocities there is a quadratic relationship (OAC) between chamber pressure and ascent velocity (see [3] for derivation of the equation). A key feature of the steady-state solution is that, for a fixed chamber pressure, it is possible to have three different magma ascent velocities. We note that for  $\mu_2/\mu_1 = 1$  only the branch CB exists and for  $1 < \mu_2/\mu_1 \leq 2$  there is a smooth transition between the lower branch (ODA) and the upper branch (CB) and multiple steady-state regimes do not exist.

We first consider the case where chamber pressure changes quasi-statically and the value of  $Q_{in}$  is between  $Q_A$  and  $Q_C$ . Starting at point O the chamber pressure increases, because the influx into the chamber is higher than the outflux. At point A, a further increase in pressure is not possible along the same branch of the steady-state solution and the system must change to point B, where the outflux of magma is larger than the influx. The chamber pressure and ascent velocity decrease along BC until the point C is reached and the system must change to point D. The cycle DABC then repeats itself. Provided the chamber continues to be supplied at the same constant rate repetition of this cycle results in periodic behavior. The transitions AB and CD in a cycle must involve unsteady flow.

Oscillations in magma discharge rate involve large variations in magma crystal content. This relation is observed on many volcanoes. For example, pumice and samples of the Soufrière Hills dome that were erupted during periods of high discharge [88] have high glass contents (25 to 35%) and few microlites [65], whereas samples derived from parts of the dome that were extruded more slowly (days to weeks typically) have much lower glass contents (5 to 15%) and high contents of groundmass microlites. These and other observations [34,68] suggest that micro-lite crystallization can take place on similar time scales to the ascent time of the magma.

Of more general interest is to consider unsteady flow behaviour. We assume that the initial distribution of parameters in the conduit corresponds to the steady-state solution of system (1) with the initial magma discharge rate,  $Q_0$ , being in the lowest regime. The behaviour of an erup-



Volcanic Eruptions: Cyclicity During Lava Dome Growth, Figure 4  
 The general steady-state solution and possible quasi-static evolution of an eruption. After [3]. Different curves correspond to different values of dimensionless parameter  $\kappa = (\pi \delta^2 \gamma)/(4V_{ch} \rho g)$  (0.12 – thin line, 0.05 – dotted line, and 0.005 – bold line)

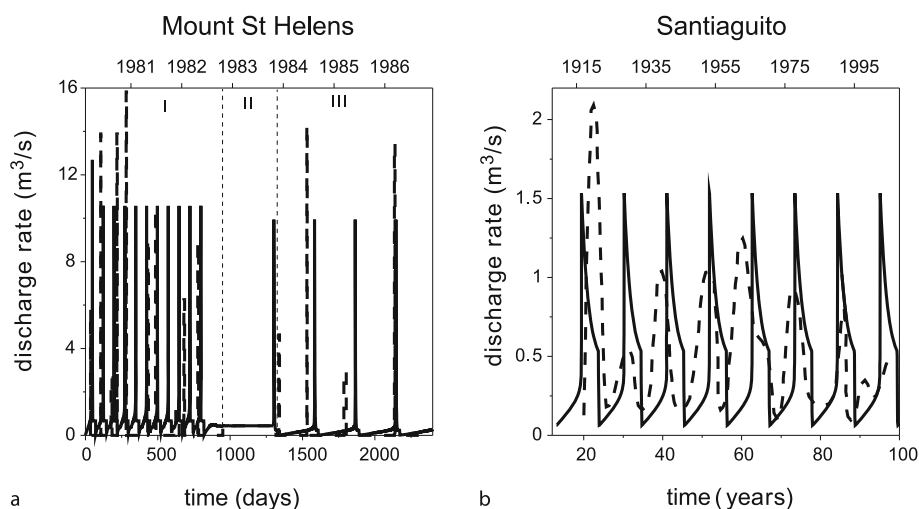
tion with time depends strongly on the value of  $Q_{in}$ . If  $Q_{in}$  corresponds to the upper or the lower branch of the steady state solution the eruption stabilizes with time with  $Q = Q_{in}$  and  $dp_{ch}/dt = 0$ . However, if  $Q_{in}$  corresponds to the intermediate branch of the steady-state solution, with  $Q_{in}$  between  $Q_A$  and  $Q_C$ , periodic behaviour is possible. Figure 4b shows three eruption scenarios for different values of the magma chamber volume  $V_{ch}$ . When  $V_{ch}$  is small the eruption stabilizes with time. In contrast, undamped periodic oscillations occur for values of  $V_{ch}$  larger than some critical value. For very large magma chamber volumes the transient solution almost exactly follows the steady-state solution, with unsteady transitions between the regimes. The time that the system spends in unsteady transitions in this case is much shorter than the period of pulsations. The maximum discharge rate during the cycle is close to  $Q_B$ , the minimum is close to  $Q_D$  and the average is equal to  $Q_{in}$ . The period of pulsations increases as the volume of magma chamber increases.

Now we apply the model to two well-documented eruptions: the growth of lava domes on Mount St. Helens (1980–1986) and on Santiaguito (1922–present). Our objective here is to establish that the model can reproduce the periodic behaviors observed at these two volcanoes. Estimates can be obtained for most of the system parameters. However, magma chamber size is not well-constrained and so the model can be used to make qualitative inferences on relative chamber size. Given the uncertainties in the parameter values and the simplifications in the model development, the approach can be characterized as

mimicry. Adjustments in some parameters were made to achieve best fits with observations, but the particular best-fits are not unique.

For Mount St. Helens our model is based on data presented by [89]. Three periods of activity can be distinguished during the period of dome growth. The first period consists of 9 pulses of activity with average peak magma discharge rates  $\sim 15 \text{ m}^3/\text{s}$  during 1981–1982. Each pulse lasted from 2 to 7 days (with a mean value of 4 days) with the average period between the pulses being 74 days and the average discharge rate during the period being  $Q_1 \sim 0.67 \text{ m}^3/\text{s}$ . The second period is represented by continuous dome growth and lasted more than a year (368 days) with a mean magma discharge rate of  $Q_2 \sim 0.48 \text{ m}^3/\text{s}$ . During the last period there were 5 episodes of dome growth with peak magma discharge rates up to  $\sim 15 \text{ m}^3/\text{s}$ , an average period of pulsation of  $\sim 230$  days and a mean discharge rate of  $Q_3 \sim 0.23 \text{ m}^3/\text{s}$ . We assume that the intensity of influx into the magma chamber,  $Q_{in}$ , is equal to the average magma discharge rate over the corresponding periods. There might have been a progressive decrease in the intensity of influx during the eruption, but, due to limitations of the model, we assume that  $Q_{in}$  changes as a step function between the periods.

The best-fit model for the eruption is presented in Fig. 5a, with the parameters used for the simulation summarized in Table 2 in [3]. During the first period of the eruption  $Q_{in} = Q_1$ . This corresponds to the intermediate branch of the steady-state solution and cyclic behaviour occurs. In the second period  $Q_{in} = Q_2$  the system moves



Volcanic Eruptions: Cyclicity During Lava Dome Growth, Figure 5

Discharge rate versus time for a Mount St. Helens dome growth and b Santiaguito volcano. Dotted lines represent the observed values of discharge rates and the solid lines are the best fit simulations. After [3]

to the lower regime and the eruption stabilizes with time. For the third period of the eruption the parameters of the system have been changed, so that  $Q_{in} = Q_3 < Q_2$ , corresponding to the intermediate regime once again and periodic behaviour occurs. This condition can be satisfied by a decrease in the diameter of the conduit, or a decrease in crystal growth rate, or the number density of crystals. All these mechanisms are possible: decrease in the diameter could be a consequence of magma crystallization on the conduit walls, while a decrease in either crystal growth rate or number density of crystals could be reflect observed changes in magma composition [89]. The influence of conduit diameter is the strongest because ascent velocity, for the same discharge rate, depends on the square of the diameter. The required change in diameter is from 18 to 12 m, but this change can be smaller if we assume a simultaneous decrease in crystal growth rate.

Since 1922, lava extrusion at Santiaguito has been cyclic [35]. Each cycle begins with a 3–6 year long high (0.5–2.1 m<sup>3</sup>/s) magma discharge rate phase, followed by a longer (3–11 years) low ( $\sim 0.2$  m<sup>3</sup>/s) discharge rate phase. The time-averaged magma discharge rate was almost constant at  $\sim 0.44$  m<sup>3</sup>/s between 1922 and 2000. The first peak in discharge rate had a value  $>2$  m<sup>3</sup>/s, whereas the second peak had a much smaller value. The value for the second peak is underestimated as it is calculated based on the dome volume only, but does not include the volume of dome collapse pyroclastic flows. Later peaks show an increase in magma discharge rate until 1960 (Fig. 5b, dashed line). Post-1960, the duration of the low discharge rate phase increased, the peak discharge and the time-averaged discharge rates for each cycle decreased, and the difference between discharge rates during the high and low discharge rate phases of each cycle decreased. Our best-fit model is shown in Fig. 5b and the parameters estimates are listed in Table 2 in [3]. The model reproduces the main features of the eruption, including the period of pulsations, the ratio between low and high magma discharge rates, and the range of observed discharge rates. We cannot, however, reproduce the decrease in the amplitude of pulsations within the framework of the model using fixed parameter values.

The theory provides a potential method to estimate magma chamber volumes. For Mount St. Helens our estimate of the chamber size ( $\sim 0.6$  km<sup>3</sup>) is comparable with the total erupted volume in the entire 1980–1986 eruption and is consistent with the fact that geophysical imaging did not identify a large magma body. Santiaguito volcano erupted more than 10 km<sup>3</sup> in the 1902 explosive eruption [96] and more than 1 km<sup>3</sup> of lava domes since 1922. The best-fit model estimate of a large (64 km<sup>3</sup>) chamber is consistent with much larger eruption volumes, long peri-

ods, and longevity of the eruption in comparison to Mount St. Helens. One limitation of the model is that the supply of deep magma from depth to the chamber is assumed to be constant.

### Model Development

In this section we further develop models to examine new effects and relax some of the simplifications of earlier models. We investigate a number of effects that were not fully explained or considered in previous studies [3,61,62]. The new model incorporates a more advanced treatment of crystallization kinetics based on the theoretical concepts developed in [38,46], and is calibrated by experimental studies in andesitic systems [22,34]. In particular, we distinguish growth of phenocrysts formed in the magma chamber from crystallization of microlites during magma ascent. Previous models have assumed that magma is always Newtonian, so we study models of conduit flow assuming non-Newtonian rheology, with rheological properties being related to crystal content. Latent heat is released during the crystallization of ascending magma due to degassing and we show that this can have an important influence on the dynamics. Elastic deformation of conduit walls leads to coupling between magma ascend and volcano deformations.

**System of Equations** We model magma ascent in a dyke-shaped conduit with elliptical cross-section using a set of 1D transient equations written for horizontally averaged variables [20,21]:

$$\frac{1}{S} \frac{\partial}{\partial t} (S \rho_m) + \frac{1}{S} \frac{\partial}{\partial x} (S \rho_m V) = -G_{mc} - G_{ph} \quad (2)$$

$$\frac{1}{S} \frac{\partial}{\partial t} (S \rho_{mc}) + \frac{1}{S} \frac{\partial}{\partial x} (S \rho_{mc} V) = G_{mc} \quad (3a)$$

$$\frac{1}{S} \frac{\partial}{\partial t} (S \rho_{ph}) + \frac{1}{S} \frac{\partial}{\partial x} (S \rho_{ph} V) = G_{ph} \quad (3b)$$

$$\frac{1}{S} \frac{\partial}{\partial t} (S \rho_d) + \frac{1}{S} \frac{\partial}{\partial x} (S \rho_d V) = -J \quad (4a)$$

$$\frac{1}{S} \frac{\partial}{\partial t} (S \rho_g) + \frac{1}{S} \frac{\partial}{\partial x} (S \rho_g V_g) = J \quad (4b)$$

Here  $t$  denotes time,  $x$  the vertical coordinate,  $\rho_m$ ,  $\rho_{ph}$ ,  $\rho_{mc}$ ,  $\rho_d$  and  $\rho_g$  are the densities of melt, phenocrysts, microlites, dissolved gas and exsolved gas respectively, and  $V$  and  $V_g$  are the velocities of magma and gas, respectively.  $G_{ph}$ ,  $G_{mc}$  represent the mass transfer rate due to crystallization of phenocrysts and microlites, respectively, and  $J$  the mass transfer rate due to gas exsolution,  $S$  is the cross-section area of the conduit. Equation (2) represents the mass con-

servation for the melt phase, Eqs. (3a) and (3b) are the conservation equations for microlites and phenocrysts respectively, Eqs. (4a) and (4b) represent the conservation of the dissolved gas and of the exsolved gas respectively.

$$\frac{\partial p}{\partial x} = -\rho g - F_c \quad (5)$$

$$V_g - V = -\frac{k}{\mu_g} \frac{\partial p}{\partial x} \quad (6)$$

Here  $p$  is the pressure,  $\rho$  the bulk density of magma,  $g$  the acceleration due to gravity,  $\mu$  is the magma viscosity,  $k$  is the magma permeability and  $\mu_g$  is the gas viscosity. Eq. (5) represents the equation of momentum for the mixture as a whole, in which the pressure drops due to gravity and conduit resistance are calculated for laminar flow in an elliptic pipe. Equation (6) is the Darcy law for the exsolved gas flux through the magma.

$$\begin{aligned} \frac{1}{S} \frac{\partial}{\partial t} (S\rho C_m T) + \frac{1}{S} \frac{\partial}{\partial x} (S\rho C_m VT) \\ = L_* (G_{mc} + G_{ph}) - C_m TJ - Q_{cl} + Q_{vh} \end{aligned} \quad (7)$$

Here  $C_m$  is the bulk specific heat of magma,  $T$  is the bulk flow-averaged temperature,  $L_*$  is latent heat of crystallization,  $Q_{cl}$  denotes the total heat loss by conduction to the conduit walls, and  $Q_{vh}$  denotes the total heat generation due to viscous dissipation. Here we consider the case of the latent heat release only. This assumption is valid when both  $Q_{cl} \approx 0$  and  $Q_{vh} \approx 0$  or when  $Q_{cl} + Q_{vh} \approx 0$ . The study of the effects of both heat loss and viscous heating, which are intrinsically two-dimensional [18,19], and their parametrization is the subject of ongoing research.

$$\rho_m = \rho_m^0 (1 - \alpha)(1 - \beta)(1 - c); \quad \rho_c = \rho_c^0 (1 - \alpha)\beta \quad (8a)$$

$$\rho_d = \rho_m^0 (1 - \alpha)(1 - \beta)c; \quad \rho_g = \rho_g^0 \alpha \quad (8b)$$

$$\rho = \rho_m + \rho_c + \rho_d + \rho_g \quad (8c)$$

$$\alpha = \frac{4}{3} \pi r_b^3 n; \quad \frac{\partial}{\partial t} (Sn) + \frac{\partial}{\partial x} (SnV) = 0; \quad p = \rho_g^0 RT \quad (9)$$

Here  $\alpha$  is the volume concentration of bubble,  $\beta$  is the volume concentration of crystals in the condensed phase (melt plus crystals), and  $c$  is mass concentration of dissolved gas (equal to volume concentration as we assume that the density of dissolved volatiles is the same as the density of the melt),  $\rho_m^m$  denotes the mean density of the pure melt phase,  $\rho_c^c$  is density of the pure crystal phase (with  $\rho_c = \rho_{ph} + \rho_{mc}$ ,  $\beta = \beta_{ph} + \beta_{mc}$ ),  $r_b$  is the bubble radius, and  $n$  the number density of bubble per unit

volume. Concerning the parametrization of mass transfer rate functions, we use:

$$J = 4\pi r_b n D \rho_m^0 (c - C_f \sqrt{p}) \quad (10)$$

$$\begin{aligned} G_{mc} = 4\pi \rho_c^0 (1 - \beta)(1 - \alpha) \\ \times U(t) \int_0^t I(\omega) \left( \int_0^t U(\eta) d\eta \right)^2 d\omega \end{aligned} \quad (11a)$$

$$G_{ph} = 3\gamma_s \left( \frac{4\pi N_{ph} \beta_{ph}^2}{3} \right)^{\frac{1}{3}} \rho_c^0 (1 - \beta)(1 - \alpha) U(t) \quad (11b)$$

Here  $J$  is parametrized using the analytical solution described in [69],  $U$  is the linear crystal growth rate ( $\text{m s}^{-1}$ ),  $I$  is the nucleation rate ( $\text{m}^{-3} \text{s}^{-1}$ ), which defines the number of newly nucleated crystal per cubic meter, and  $\gamma_s$  is a shape factor of the order of unity,  $D$  and  $C_f$  are the diffusion and the solubility coefficients, respectively. Concerning the mass transfer due to crystallization  $G_{mc}$ , we adapt a model similar to that described in [38]. Assuming spherical crystals, the Avrami–Johnson–Mehl–Kolmogorov equation in the form adopted by [46], for the crystal volume increase rate, is:

$$\frac{d\beta}{dt} = 4\pi Y_t U(t) \int_0^t I(\omega) \left( \int_0^t U(\eta) d\eta \right)^2 d\omega$$

where  $Y_t = (1 - \beta)(1 - \alpha)$  is the volume fraction of melt remaining uncrystallized at the time  $t$ . Therefore, we have  $G_{mc} = \rho_{mc} d\beta/dt$ . For the phenocryst growth rate  $G_{ph}()$  we assume that it is proportional to the phenocryst volume increase rate  $d\beta_{ph}/dt = 4\pi R_{ph}^2 N_{ph} U(t)$  times the crystal density  $\rho_c^0$  times the volume fraction of melt remaining uncrystallized at the time  $t$ . A detailed description of the parametrization used for the different terms is reported in [63].

For parametrizations of magma permeability  $k$  and magma viscosity  $\mu$  we use:

$$k = k(\alpha) = k_0 \alpha^j \quad (12)$$

$$\mu = \mu_m(c, T) \theta(\beta) \eta(\alpha, Ca) \quad (13)$$

where  $k$  is assumed to depend only on bubble volume fraction  $\alpha$ . Magma viscosity  $\mu$  depends on water content, temperature, crystal content, bubble fraction and capillary number as described in detail in the next section.

Regarding equations for semi-axes,  $a$  and  $b$ , we assume that the elliptical shape is maintained and that pressure change gradually in respect with vertical coordinate and time so that the plain strain analytical solution for an ellipse subjected to a constant internal overpressure [64,66],

remains valid:

$$a = a_0 + \frac{\Delta P}{2G} [-(1-2\nu)a_0 + 2(1-\nu)b_0] \quad (14a)$$

$$b = b_0 + \frac{\Delta P}{2G} [2(1-\nu)a_0 - (1-2\nu)b_0] \quad (14b)$$

where  $\Delta P$  is the overpressure, i. e. the difference between conduit pressure and far field pressure (here assumed to be lithostatic for a sake of simplicity),  $a_0$  and  $b_0$  are the initial values of the semi-axes,  $\nu$  is the host rock Poisson ratio, and  $G$  is the host rock rigidity.

Equations (2)–(14) are solved between the top of the magma chamber and the bottom of the lava dome that provides some constant load by using the numerical method described in [63]. The effects of dome height and morphology changes are not considered in this paper. We consider three different kinds of boundary conditions at the inlet of the dyke: constant pressure, constant influx rate and the presence of a magma chamber located in elastic rocks. The case of constant pressure is applicable when a dyke starts from either a large magma chamber or unspecified source, so that pressure variations in the source region remain small. An estimate of the volume of magma stored in the source region that allows pressure to be approximated as constant depends on wall-rock elasticity, magma compressibility (volatile content), and the total volume of the erupted material. If the magma flow at depth is controlled by regional tectonics, the case of constant influx rate into the dyke may be applicable if total variations in supply rate are relatively small on the timescale of the eruption.

For the case where magma is stored in a shallow magma chamber prior to eruption, and significant chamber replenishment occurs, the flow inside the conduit must be coupled with the model for the magma chamber. In this case, as explained in detail in [63], we assume that the relationship between the pressure at the top of the magma chamber  $p_{ch}$  and the intensity of influx  $Q_{in}$  and outflux  $Q_{out}$  of magma to and from the chamber is given by:

$$\frac{dp_{ch}}{dt} = \frac{4G \langle K \rangle}{\langle \rho \rangle V_{ch} (3 \langle K \rangle + 4G)} (Q_{in} - Q_{out}) \quad (15)$$

where  $V_{ch}$  is the magma chamber volume,  $\langle \rho \rangle$  and  $\langle K \rangle$  are the average magma density and magma bulk modulus, respectively, and  $G$  is the rigidity of rocks surrounding the chamber.

Cases of constant influx and of constant source pressure are the limit cases of Eq. (15) in the case of infinitely small and infinitely large magma chamber volume. We assume that the volume concentration of bubbles and phenocrysts are determined by equilibrium conditions and

that the temperature of the magma is constant. The effect of temperature change on eruption dynamics, due to interaction between silicic and basaltic magma, was studied in [63].

We use a steady-state distribution of parameters along the conduit as an initial condition for the transient simulation. The values are calculated for a low magma discharge rate, but the particular value of this parameter is not important because the system deviates from initial conditions to a cyclic or stabilized state, which does not depend on the initial conditions.

### Rheology of Crystal-Bearing Magma and Conduit Resistance

Magma viscosity is modeled as a product of melt viscosity  $\mu_m(c, T)$ , the relative viscosity due to crystal content  $\theta(\beta) = \Theta(\beta)\varphi(\beta)$ , and the relative viscosity due to the presence of bubbles  $\eta(\alpha, Ca)$ . The viscosity of the pure melt  $\mu_m(c, T)$  is calculated according to [36]. Viscosity increase due to the presence of the crystals is described through the function  $\Theta(\beta)$  [16,20,21]. As crystallization proceeds, the remaining melt becomes enriched in silica and melt viscosity increases. The parametrization of this effect is described by the function  $\varphi(\beta)$  in [21,25]. Effects of the solid fraction are parametrized as described in [21].

Effects due to the presence of bubbles are accounted for by adopting a generalization of [51] for an elliptical conduit [20,21].

In the case of Newtonian magma rheology, the friction force in an elliptical conduit can be obtained from a classical Poiseuille solution for low Reynolds number flow  $F_c = 4\mu(a^2 + b^2)/(a^2b^2)V$  [47]. High crystal or bubble content magmas may show non-Newtonian rheology. One possible non-Newtonian rheology is that of a Bingham material characterized by a yield strength  $\tau_b$  [4]. The stress-strain relation for this material is given by:

$$\begin{aligned} \tau_{ij} &= \left( \mu + \frac{\tau_b}{\gamma_b} \right) \gamma_{ij} \Leftrightarrow \tau > \tau_b \\ \gamma_{ij} &= 0 \Leftrightarrow \tau \leq \tau_b \end{aligned} \quad (16)$$

Here  $\tau_{ij}$  and  $\gamma_{ij}$  are the stress and strain rate tensors,  $\tau$  and  $\gamma_b$  are second invariants of corresponding tensors. According to this rheological law, the material behaves linearly when the applied stress is higher than a yield strength. No motion occurs if the stress is lower than a yield strength. In the case of a cylindrical conduit the average velocity can be calculated in terms of the stress on the conduit wall  $\tau_w$  [52]:

$$V = \frac{1}{12} \frac{r}{\tau_w^3 \mu} (\tau_b^4 + 3\tau_w^4 - 4\tau_b \tau_w^3) \quad (17)$$

Here  $r$  is the conduit radii. This form of equation gives an implicit relation between ascent velocity and pressure drop, and is not convenient to use. By introducing dimensionless variables  $\Pi = \mu V/\tau_b r$  and  $\Omega = \tau_w/\tau_b \geq 1$  relation (17) can be transformed into:

$$\Omega^4 - \frac{1}{6} (8 + 3\Pi) \Omega^3 + \frac{1}{3} = 0 \tag{18}$$

Following [63] a semi-analytical solution can be used for (18) and the conduit friction force can be expressed finally as:

$$F_c = \frac{2\tau_w}{r} = \frac{2\tau_b \Omega (\Pi)}{r}.$$

We note that a finite pressure gradient is necessary to initiate the flow in the case of Bingham liquid, in contrast to a Newtonian liquid.

**Results and Applications**

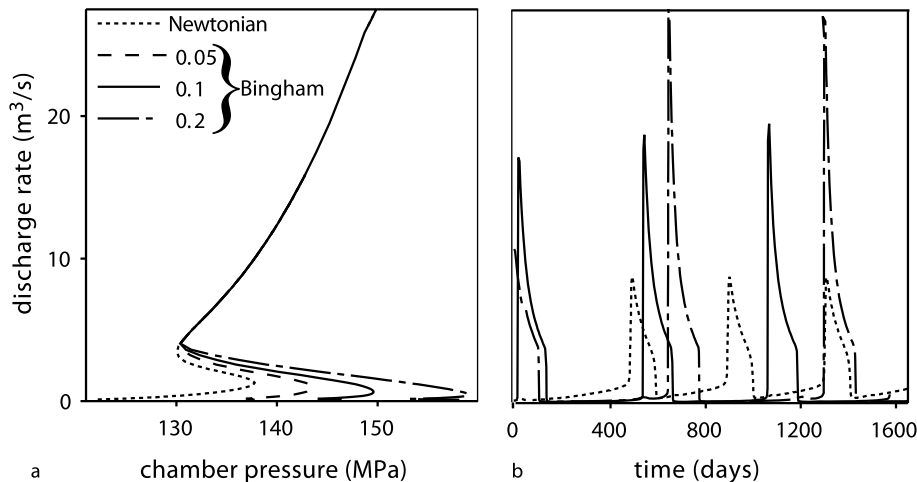
*Influence of non-Newtonian Properties on Eruption Behaviour* Now we compare the dynamics of magma extrusion in the cases of Newtonian and Bingham rheology. We will assume that yield strength is reached when the concentration of crystals reaches a critical value:

$$\tau = \begin{cases} \tau_b & \text{for } \beta > \beta_{cr} \\ 0 & \text{for } \beta \leq \beta_{cr} \end{cases} \tag{19}$$

Figure 6a shows a set of steady-state solutions for different values of  $\tau_b$ . Values of  $\tau_b$  and  $\beta_{cr}$  depend on crystal shape, crystal size distribution, magma temperature and other properties, but here are assumed to be constant. To illustrate the influence of Bingham rheology, the value of  $\beta_{cr} = 0.65$  was chosen so that, for discharge rate larger then  $\sim 5 \text{ m}^3/\text{s}$ , the magma has Newtonian rheology (see Fig. 6a). A more detailed study would require measurements of the rheological properties of magma for a wide range of crystal content and crystal size distributions. As the value of  $\tau_b$  the chamber pressure that is necessary to start the eruption increases.

Figure 6b shows the influence of these two rheological models on the dynamics of magma extrusion. In the case of Bingham rheology, magma discharge rate between the two pulses is zero until a critical chamber overpressure is reached. Then the magma discharge rate increases rapidly with decrease in crystal content, leading to a significant reduction of both magma viscosity and the length of the part of the conduit that is occupied by the Bingham liquid where  $\beta_c > \beta_{cr}$ . There is a transition in the system to the uppermost flow regime and the pressure then decreases quickly. Because the pressure at the onset of the pulse was significantly larger than in the case of a Newtonian liquid, the resulting discharge rate in the case of Bingham rheology is also significantly higher.

*Modeling of Conduit Flow during Dome Extrusion on Shiveluch Volcano* The maximum intensity of extrusion was



Volcanic Eruptions: Cyclicity During Lava Dome Growth, Figure 6  
**a** Steady-state solutions and dependence of discharge rate on time for Newtonian and Bingham rheology of the magma. Yield strength is a parameter marked on the curves (values in MPa). For Bingham rheology discharge rate remains zero between the pulses of activity. Bingham rheology results in much higher chamber pressures prior to the onset of activity and, therefore, much higher discharge rates in comparison with Newtonian rheology. **b** Comparison of the period of pulsation in discharge rate for Newtonian and Bingham rheologies. After [63]

reached at an early stage in all three eruptions (Fig. 3). We therefore suggest that dome extrusion was initiated by high overpressure in the magma chamber with respect to the lithostatic pressure. Depressurization of the magma chamber occurred as a result of extrusion. Without magma chamber replenishment, depressurization results in a decrease in magma discharge rate. In open system chambers replenishment of the chamber during eruption can lead to pulsatory behaviour [3]. The following account is derived from [25]. For the 1980–1981 eruption the monotonic decrease in discharge rate indicates that there was little or no replenishment of the magma chamber. During the 1993–1995 and 2001–2004 episodes, however, the magma discharge rate fluctuated markedly, suggesting that replenishment was occurring. The influx of new magma causes an increase in magma chamber pressure, and a subsequent increase in magma discharge rate. During the 2001–2004 eruption there were at least three peaks in discharge rate. Replenishment of the magma chamber with new hot magma can explain the transition from lava dome extrusion to viscous lava flow that occurred on Shiveluch after 10 May 2004, and which continues at the time of writing (2007).

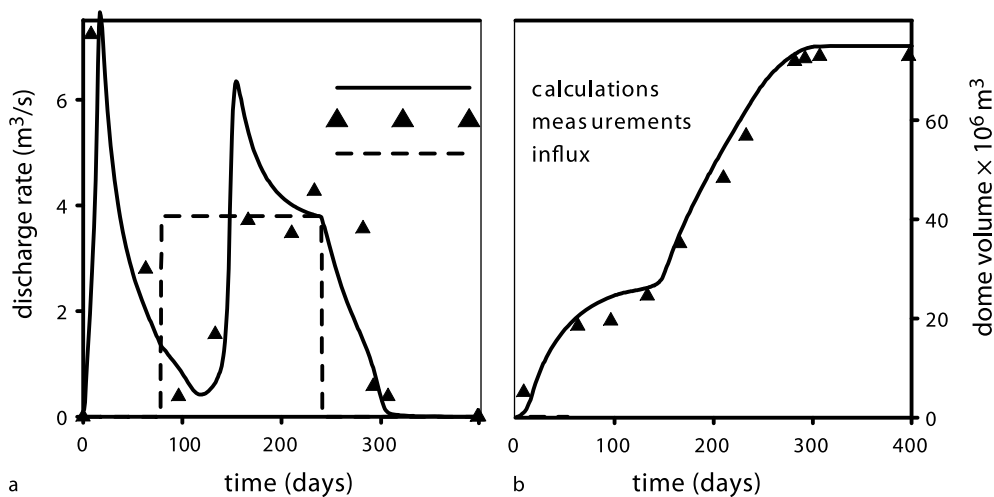
We simulated dome growth during the 2001–2002, because this dataset is the most complete and is supported by petrological investigations [39]. We assume the shape of the influx curve:

$$Q_{\text{in}} = \begin{cases} 0, & t < t_s \\ Q_0, & t_s \leq t \leq t_f \\ 0, & t > t_f \end{cases} \quad (20)$$

Influx occurs with constant intensity  $Q_0$  between times  $t_s$  and  $t_f$ . We have examined many combinations of values of these parameters within the constraints provided by observations. The best simulation results use the following values of parameters:  $Q_0 = 3.8 \text{ m}^3/\text{s}$ ,  $t_s = 77$  and  $t_f = 240$  days. A more continuous influx, dependent on time, is plausible, but there is no geophysical evidence that allows us to constrain the intensity of the influx, because ground deformation data are absent for Shiveluch volcano. The output of the model gives a magma chamber volume of  $12 \text{ km}^3$ , assuming a spherical chamber. Figure 7a shows the time-dependence of magma discharge rate, and Fig. 7b shows the increase in the volume of erupted material with time after 6th June 2001. The timing of magma influx is in good agreement with the residence time of basaltic magma in the system, as calculated from the olivine reaction rims. For further details see [25].

*5 to 7 Weeks Cycles on the Soufriere Hills Volcano: Evidence for a Dyke?* An approximately 5 to 7 week cyclic pattern of activity was recognized at the Soufrière Hills Volcano (SHV) [87,91] between April 1997 and March 1998 from peaks in the intensity of eruptive activity and geophysical data, including tilt and seismicity (Fig. 2).

In models discussed above, the time-scale of pulsations depends principally on the volume of the magma chamber, magma rheology and the cross-sectional area of the conduit. These models might provide an explanation for the 2–3 year cycles of dome extrusion observed at SHV, where deformation data indicate that the magma chamber regulates the cycles. However, the models cannot simulta-



Volcanic Eruptions: Cyclicity During Lava Dome Growth, Figure 7

**a** Comparison of calculated and measured discharge rates (**a**) and volumes of the dome (**b**) for the episode of the dome growth in 2001–2002. Influx into the magma chamber is shown by a dashed line in **a**. Time in days begins on 6th June, 2001. After [25]



neously explain the 5–7 week cycles. Thus another mechanism is needed.

The evidence for a dyke feeder at SHV includes GPS data [58], distribution of active vents, and seismic data [76]. We have assumed that, at depth, the conduit has an elliptical shape that transforms to a cylinder at shallow level. In order to get a smooth transition from the dyke at depth to a cylindrical conduit (see Fig. 1) the value of  $a_0$  in Eqs. (14) is parametrized as:

$$a_0(x) = A_1 \arctan\left(\frac{x - L_T}{w_T}\right) + A_2 \quad (21)$$

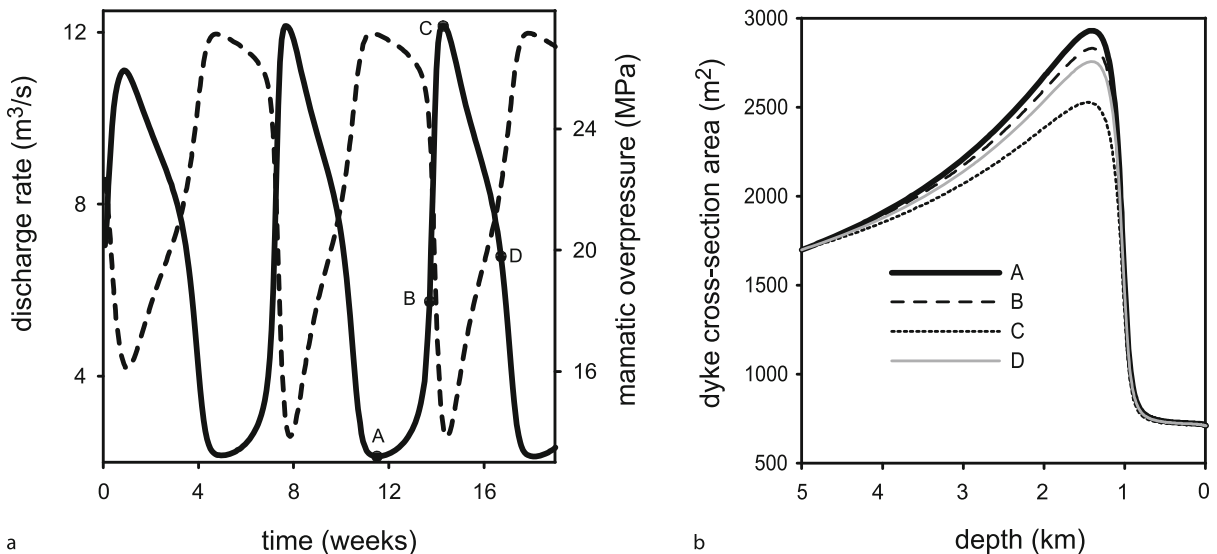
Here  $L_T$  and  $w_T$  are the position and the vertical extent of the transition zone between the ellipse and the cylinder and constants  $A_1$  and  $A_2$  are calculated to satisfy conditions  $a_0(L) = R$  and  $a_0(0) = a_0$ , where  $R$  is the radius of the cylindrical part of the conduit and  $a_0$  is the length of semi-major axis at the inlet of the dyke. The value of  $b_0$  is calculated in order to conserve the cross-section area of the unpressurized dyke, although it can also be specified independently.

In order to de-couple the influence of the dyke geometry from the oscillations caused by magma chamber pressure variations, we have assumed a fixed chamber pressure

as a boundary condition for the entrance to the conduit. This assumption is valid because the timescale of chamber pressure variations are much longer than the period of the cycle (2–3 years in comparison with 5–7 weeks).

Results presented in Figure 8 show that, even with a fixed chamber pressure, there are magma discharge rate oscillations. At the beginning of a cycle the magma discharge rate is at a minimum, while the overpressure (here presented for 1 km depth by a dashed line) and dyke width are at a maximum. At point A in Fig. 8a the crystal content and viscosity have reached their maximum values. Beyond this threshold condition, an increase in magma discharge rate results in decreasing pressure and dyke width. However, crystal content and viscosity also decrease and this effect decreases friction, resulting in flow rate increase and pressure decrease. At C the system reaches minimum viscosity and crystal content, which cannot decline further. Thereafter the magma discharge rate decreases, while the pressure and dyke width increase. The dyke acts like a capacitor, storing volume during this part of the cycle.

The period of oscillation depends on several parameters such as influx rate and dyke aspect ratio  $a/R$ . Typically the period decreases with increasing aspect ratio. The range of calculated periods varies between 38 and 51 days



Volcanic Eruptions: Cyclicity During Lava Dome Growth, Figure 8

**a** Dependence of magma discharge rate (solid line) and magmatic overpressure at depth of 1 km (dashed line) on time, for  $a = 240$  m and  $b = 2.25$  m at the inlet of the dyke. The period of cycle is 46 days, average discharge rate is  $6.2 \text{ m}^3/\text{s}$ , with peak rate about  $12 \text{ m}^3/\text{s}$ . **b** Profiles of cross-section areas of the conduit during one cycle. Curve A corresponds to the beginning of the cycle, B – to a point on the curve of ascending discharge rate, C – to maximum discharge, and D to the middle of descending discharge curve. At the beginning of the cycle, due to large viscosity of magma (at low discharge rate crystal content is high) large magmatic overpressure develops, reaching a maximum near the transition between the dyke and cylindrical conduit; the dyke inflates providing temporary magma storage. Minimum dyke volume corresponds to maximum discharge rate (curve C). After [21]

for semi major-axis lengths,  $a$ , from 175 to 250 m and semi minor-axes,  $b$ , from 2 to 4 m. These results match observed cyclicity at SHV. The start of a cycle is quite sharp (Fig. 2), with the onset of shallow hybrid-type (impulsive, low-frequency coda) earthquakes. The change to shorter period and higher amplitude tilt pulsations indicate a marked increase in average magma discharge rate [91,98]. The model cycles also have rapid onsets. The high-amplitude tilt pulsations lasted for several weeks [91], consistent with the duration of higher magma discharge rates early in each 5–7 week cycle. Tilt data (Fig. 2) are consistent with the model in that the episode of high magma discharge is associated with a marked deflation that lasts several weeks (see dashed curve at Fig. 8, representing magmatic overpressure at 1 km depth). The magma pressure builds up in the swelling dyke and then reaches a threshold, whereupon a surge of partly crystallized magma occurs, accompanied by elevated seismicity.

The models presented above have certain general features that are necessary to show cyclic behaviour. First of all, the resistance of the conduit must depend on magma discharge rate in a way that resistance decreases when discharge rate increases in some range of discharge rate. This dependence is reproduced by a sigmoidal curve. Resistance is a product of viscosity and velocity, and is linearly proportional to discharge rate. This means that magma viscosity must decrease as discharge rate strongly increases. There may be many reasons for this behaviour, including crystallization, temperature variation or gas diffusion. The second condition is that there must be some capacitor in the system that can store magma in a period of low discharge rate and release it in a period of high discharge rate. The role of this capacitor can be played by a magma chamber or dyke-shaped conduit located in elastic rocks, or by compressibility of the magma itself. The volumes of these capacitors are different and thus will cause pulsations with different periods. Currently there is no single model that can account for pulsations with multiple timescales.

### Future Directions

Our models indicate that magmatic systems in lava dome eruptions can be very sensitive to small changes in parameters. This sensitivity is most marked when the system is close to the cusps of steady-state solutions. If magma discharge rate becomes so high that gas cannot escape efficiently during magma ascent, then conditions for magma fragmentation and explosive eruption can arise. Empirical evidence suggests that conditions for explosive eruption arise when magma discharge rates reach approxi-

mately  $10 \text{ m}^3/\text{s}$  or more in dome eruptions [45,83]. Calculations show the possibility of such high discharge rates for the system parameters typical of lava dome-building eruptions.

We have illustrated model sensitivity of results by varying only one parameter at a time on plots of chamber pressure and discharge rate. However, magmatic systems have many controlling parameters that may vary simultaneously. Furthermore, some controlling parameters are likely to be interdependent (such as temperature, volatile content and phenocryst content, for example) and others may be independent (such as magma temperature and conduit dimensions). An eruption can be expected to move through  $n$ -parameter space, making simulation and its parameter depiction difficult. Our results are simplified, so system sensitivity and behaviour in the real world may be yet more complex. A volcanic system may be quite predictable when it is within a stable regime, but may become inherently unpredictable [84,85] when variations in the parameters move the system towards transition points and flow regime boundaries.

As in all complex systems there are many controlling parameters. Our models capture some of the key dynamics, but are still simplified in many respects, so do not fully capture the real variations. Our models do not, for example, consider variations in dome height, gas escape to surrounding rocks, strain-rate dependent rheological effects or time dependent changes in conduit diameter. The model for porosity is based on interpretation of measurements of porosity of erupted magma. The role of post-eruptive alterations of pore structure, for example, formation of cooling cracks, cannot be easily estimated. The model of bubble coalescence and permeability formation is important for understanding gas escape mechanisms and will provide constraints on transitions between extrusive and explosive activity. Because the model remains 1D, lateral distribution of parameters cannot be studied. These include: lateral pressure gradients, magma crystallization on the conduit walls, wallrocks melting or erosion, formation of shear zones and shear heating, heat flux to surrounding rocks. The models also make the simplifying assumption that influx into the chamber from a deep source is a constant, or given as a function of time. The dynamics of the magma chamber itself are oversimplified in all existing conduit flow models. Changes in magma properties in magma chamber can affect the long-term evolution of eruptions. We have considered water as the only volatile and the addition of other gas species (e. g.  $\text{CO}_2$  and  $\text{SO}_2$ ) would add further variability.

There are large uncertainties in some parameters, which are likely to be very strong controls, such as the rhe-

ological properties of high crystalline magmas and crystal growth kinetic parameters, notably at low pressures ( $< 30$  MPa) where experiments are very difficult to do (e.g. [22]). More experiments are necessary to understand the rheology of multiphase systems containing melt, crystals and bubbles. The effects of crystal shape, crystal size distribution and strain rate remain largely unclear. Some parameters, such as conduit geometry variation with depth, are highly uncertain. With so many parameters, good fits can be achieved by selecting plausible values for real systems. Barmin et al. [3], for example, were able to reproduce the patterns of discharge rate at Mount St. Helens and Santiaguito quite accurately. However, such models are not unique, partly because the actual values of some parameters may be quite different to the assumed values and partly because of the model simplifications.

Results obtained from waveform inversions of very-long-period seismic data over the past few years point to the predominance of a crack-like geometry for volcanic conduits [11], and it is becoming increasingly evident that the details of this geometry, such as conduit inclination, a sudden change in conduit direction, a conduit bifurcation, or a sudden increase in cross section, are all critically important in controlling magma flow dynamics. Future modeling attempts will need to be closely tied to information on conduit geometry derived from seismology to provide a more realistic view of volcanic processes.

The full simulation of any particular volcanic eruption in such a non-linear and sensitive system may appear a hopeless task. However, some reduction in uncertainties will certainly help to make the models more realistic. Further experimental studies of crystallization kinetics and the rheological properties of magma at high crystallinities are among the most obvious topics for future research. Advances in understanding the controls on magma input into an open-system chamber would be beneficial, since the delicate balance between input and output is a prime control on periodic behaviour.

Further model development includes 2D effects, elastic deformation effects in dyke-fed domes and coupling between magma chamber and conduit flow dynamics. Even with such improvements, large parameter uncertainties and modeling difficulties will remain. In such circumstances the logical approach is to start quantifying the uncertainties and sampling from them to produce probabilistic outputs based on ensemble models where numerical models of the kind discussed here can be run many times. A future challenge for numerical models will also be to produce simulated outputs which compare in detail with observations, in particular time series of magma discharge rates.

## Acknowledgments

This work was supported by NERC research grant reference NE/C509958/1. OM and AB acknowledge Russian Foundation for Basic Research (08-01-00016) and President of Russian Federation program (NCH-4710.2006.1). RSJS acknowledges a Royal Society Wolfson Merit Award. The Royal Society exchange grants and NERC grants had supported the Bristol/Moscow work over the last 10 years.

## Bibliography

### Primary Literature

1. Balmforth NJ, Burbidge AS, Craster RV (2001) Shallow Lava Theory. In: Balmforth NJ, Provenzale A (eds) *Geomorphological Fluid Mechanics*. Lecture Notes in Physics, vol 582. Springer, Berlin, pp 164–187
2. Balmforth NJ, Burbidge AS, Craster RV, Rust AC, Sassi R (2006) Viscoplastic flow over an inclined surface. *J Non-Newtonian Fluid Mech* 139:103–127
3. Barmin A, Melnik O, Sparks RSJ (2002) Periodic behaviour in lava dome eruptions. *Earth Planet Sc Lett* 199:173–184
4. Bingham EC (1922) *Fluidity and Plasticity*. McGraw–Hill, New York, p 215
5. Blake S (1990) Viscoplastic models of lava domes. In: Fink JH (ed) *Lava flows and domes, Vol 2*. In: Fink JH (ed) *Lava flows and domes; emplacement mechanisms and hazard implications*. Springer, Berlin, pp 88–126
6. Blundy JD, Cashman KV, Humphreys MCS (2006) Magma heating by decompression-driven crystallisation beneath andesite volcanoes. *Nature* 443:76–80. doi:10.1038/nature05100
7. Calder ES, Lockett R, Sparks RSJ, Voight B (2002) Mechanisms of lava dome instability and generation of rockfalls and pyroclastic flows at Soufrière Hills Volcano, Montserrat. In: Druitt TH, Kokelaar BP (eds) *The Eruption of Soufrière Hills Volcano, Montserrat, from 1995 to 1999*, Geological Society, London, Memoirs No 21, pp 173–190
8. Calder ES, Cortés JA, Palma JL, Lockett R (2005) Probabilistic analysis of rockfall frequencies during an andesite lava dome eruption: The Soufrière Hills Volcano, Montserrat. *Geophys Res Lett* 32:L16309. doi:10.1029/2005GL023594
9. Caricchi L, Burlini L, Ulmer P, Gerya T, Vassalli M, Papale P (2007) Non-Newtonian rheology of crystal-bearing magmas and implications for magma ascent dynamics. *Earth Planet Sci Lett* 10.1016/j.epsl.2007.09.032
10. Cashman KV, Blundy JD (2000) Degassing and crystallization of ascending andesite and dacite. In: Francis P, Neuberg J, Sparks RSJ (eds) *Causes and consequences of eruptions of andesite volcanoes*. *Phil Trans Royal Soc London A*, pp 1487–1513
11. Chouet B, Dawson P, Arciniega-Ceballos A (2005) Source mechanism of Vulcanian degassing at Popocatepetl Volcano, Mexico, determined from waveform inversions of very long period signals. *J Geophys Res* 110:B07301. doi:10.1029/2004JB003524
12. Christiansen RL, Peterson DW (1981) Chronology of the 1980 Eruptive Activity. In: Lipman PW, Mullineaux DR (eds) *The 1980 Eruptions of Mount St. Helens*. Washington, US Geological Survey Professional Paper 1250, p 844

13. Clarke AB, Stephens S, Teasdale R, Sparks RSJ, Diller K (2007) Petrological constraints on the decompression history of magma prior to Vulcanian explosions at the Soufrière Hills volcano, Montserrat. *J Volcanol Geotherm Res* 161:261–274. doi:10.1016/j.jvolgeores.2006.11.007
14. Cole P, Calder ES, Sparks RSJ, Clarke AB, Druitt TH, Young SR, Herd R, Harford CL, Norton GE (2002) Deposits from dome-collapse and fountain-collapse pyroclastic flows at Soufrière Hills Volcano, Montserrat. In: Druitt TH, Kokelaar BP (eds) The eruption of the Soufrière Hills Volcano, Montserrat from 1995 to 1999. Geological Society, London, Memoir No 21, pp 231–262
15. Connor CB, Sparks RSJ, Mason RM, Bonadonna C, Young SR (2003) Exploring links between physical and probabilistic models of volcanic eruptions: the Soufrière Hills Volcano, Montserrat. *Geophys Res Lett* 30. doi:10.1029/2003GL017384
16. Costa A (2005) Viscosity of high crystal content melts: dependence on solid fraction. *Geophys Res Lett* 32:L22308. doi:10.1029/2005GL02430
17. Costa A, Macedonio G (2002) Nonlinear phenomena in fluids with temperature-dependent viscosity: an hysteresis model for magma flow in conduits. *Geophys Res Lett* 29(10). doi:10.1029/2001GL014493
18. Costa A, Macedonio G (2003) Viscous heating in fluids with temperature-dependent viscosity: implications for magma flows. *Nonlinear Proc Geophys* 10:545–555
19. Costa A, Macedonio G (2005) Viscous heating in fluids with temperature-dependent viscosity: triggering of secondary flows. *J Fluid Mech* 540:21–38
20. Costa A, Melnik O, Sparks RSJ (2007) Controls of conduit geometry and wallrock elasticity on lava dome eruptions. *Earth Planet Sci Lett* 260:137–151. doi:10.1016/j.epsl.2007.05.024
21. Costa A, Melnik O, Sparks RSJ, Voight B (2007) The control of magma flow in dykes on cyclic lava dome extrusion. *Geophys Res Lett* 34:L02303. doi:10.1029/2006GL027466
22. Couch S, Sparks RSJ, Carroll MR (2001) Mineral disequilibrium in lavas explained by convective self-mixing in open magma chambers. *Nature* 411:1037–1039
23. Denlinger RP, Hoblitt RP (1999) Cyclic eruptive behaviour of silicic volcanoes. *Geology* 27(5):459–462
24. Diller K, Clarke AB, Voight B, Neri A (2006) Mechanisms of conduit plug formation: Implications for vulcanian explosions. *Geophys Res Lett* 33:L20302. doi:10.1029/2006GL027391
25. Dirksen O, Humphreys MCS, Pletchov P, Melnik O, Demyanchuk Y, Sparks RSJ, Mahony S (2006) The 2001–2004 dome-forming eruption of Shiveluch Volcano, Kamchatka: Observation, petrological investigation and numerical modelling. *J Volcanol Geotherm Res* 155:201–226. doi:10.1016/j.jvolgeores.2006.03.029
26. Druitt TH, Young S, Baptie B, Calder E, Clarke AB, Cole P, Harford C, Herd R, Luckett R, Ryan G, Sparks RSJ, Voight B (2002) Episodes of cyclic Vulcanian explosive activity with fountain collapse at Soufrière Hills volcano, Montserrat. In: Druitt TH, Kokelaar BP (eds) The eruption of the Soufrière Hills Volcano, Montserrat from 1995 to 1999. Geological Society, London, Memoir No 21, pp 231–262
27. Eichelberger JC, Carrigan CR, Westrich HR, Price RH (1986) Non-explosive silicic volcanism. *Nature* 323:598–602
28. Fedotov SA, Dvigalo VN, Zharinov NA, Ivanov VV, Seliverstov NI, Khubunaya SA, Demyanchuk YV, Markov LG, Osipenko LG, Smelov NP (2001) The eruption of Shiveluch volcano on May–July 2001. *Volcanol Seis* 6:3–15
29. Fink JH, Griffiths RW (1990) Radial spreading of viscous gravity currents with solidifying crust. *J Fluid Mech* 221:485–509
30. Fink JH, Griffiths RW (1998) Morphology, eruption rates, and rheology of lava domes: Insights from laboratory models. *J Geophys Res* 103:527–545
31. Green, DN, Neuberg J (2006) Waveform classification of volcanic low-frequency earthquake swarms and its implication at Soufrière Hills Volcano, Montserrat. *J Volcanol Geotherm Res* 153:51–63. doi:10.1016/j.jvolgeores.2005.08.003
32. Hale AJ, Bourgouin L, Mühlhaus HB (2007) Using the level set method to model endogenous lava dome growth. *J Geophys Res* 112:B03213. doi:10.1029/2006JB004445
33. Hale AJ, Wadge G (2003) Numerical modeling of the growth dynamics of a simple silicic lava dome. *Geophys Res Lett* 30(19). doi:10.1029/2003GL018182
34. Hammer JE, Rutherford MJ (2002) An experimental study of the kinetics of decompression-induced crystallization in silicic melt. *J Geophys Res* 107:(B1). doi:10.1029/2001JB000281
35. Harris AL, Rose WI, Flynn LP (2002) Temporal trends in Lava Dome extrusion at Santiaguito 1922–2000. *Bull Volcanol* 65:77–89
36. Hess KU, Dingwell DB (1996) Viscosities of hydrous leucogranite melts: A non-Arrhenian model. *Am Mineral* 81:1297–1300
37. Hoblitt RP, Wolfe EW, Scott WE, Couchman MR, Pallister JS, Javier D (1996) The preclimactic eruptions of Mount Pinatubo, June 1991. In: Newhall CG, Punongbayan RS (eds) Fire and Mud: Eruptions and Lahars of Mount Pinatubo, Philippines. Philippine Institute of Volcanology and Seismology, Quezon City, and University of Washington Press, Seattle, pp 457–511
38. Hort M (1998) Abrupt change in magma liquidus temperature because of volatile loss or magma mixing: effects of Nucleation, crystal growth and thermal history of the magma. *J Petrol* 39:1063–1076
39. Humphreys M, Blundy, JD, Sparks RSJ (2006) Magma Evolution and Open-system processes at Shiveluch Volcano: insights from phenocryst zoning. *J Petrol* 47:(12) 2303–2334. doi:10.1093/petrology/eg1045
40. Huppert HE, Shepherd JB, Sigurdsson H, Sparks RSJ (1982) On lava dome growth, with application to the 1979 lava extrusion of the Soufrière, St Vincent. *J Volcanol Geotherm Res* 14: 199–222
41. Huppert HE, Woods AW (2002) The role of volatiles in magma chamber dynamics. *Nature* 420:493–495
42. Ida Y (1996) Cyclic fluid effusion accompanied by pressure change: Implication for volcanic eruptions and tremor. *Geophys Res Lett* 23:1457–1460
43. Iverson RM et al (2006) Dynamics of seismogenic volcanic extrusion at Mount St. Helens in 2004–05. *Nature* 444:439–443
44. Jaquet O, Sparks RSJ, Carniel R (2006) Magma Memory recorded by statistics of volcanic explosions at the Soufrière Hills Volcano, Montserrat. In: Mader HM, Coles SG, Connor CB, Connor LJ (eds) Statistics in Volcanology. Geological Society, London, Special Publication of IAVCEI, vol 1. pp 175–184
45. Jaupart C, Allegre CJ (1991) Gas content, eruption rate and instabilities of eruption regime in silicic volcanoes. *Earth Planet Sci Lett* 102:413–429
46. Kirkpatrick R (1976) Towards a Kinetic Model for the Crystallization of Magma Bodies. *J Geophys Res* 81:2565–2571

47. Landau L, Lifshitz E (1987) *Fluid Mechanics*, 2nd edn. Butterworth–Heinemann, Oxford
48. Lejeune A, Richet P (1995) Rheology of crystal-bearing silicate melts: An experimental study at high viscosity. *J Geophys Res* 100:4215–4229
49. Lensky NG, Sparks RSJ, Navon O, Lyakhovsky V (2007) Cyclic activity at Soufriere Hills volcano, Montserrat: degassing-induced pressurization and stick-slip extrusion. In: Lane SJ, Gilbert JS (eds) *Fluid motions in volcanic conduits: a source of seismic and acoustic signals*. Geological Society, London, Special Publications, vol 307, pp 169–188. doi:10.1144/SP307.100305-8719/08/\$15.00 The Geological Society of London 2008
50. Lister JR, Kerr RC (1991) Fluid mechanical models of crack propagation and their application to magma transport in dykes. *J Geophys Res* 96:10049–10077
51. Llewellyn EW, Manga M (2005) Bubble suspension rheology and implications for conduit flow. *J Geotherm Res* 143: 205–217
52. Loitsyansky LG (1978) *Fluid and gas mechanics*. Nauka, Moscow, pp 847 (in Russian)
53. Maeda I (2000) Nonlinear visco-elastic volcanic model and its application to the recent eruption of Mt. Unzen. *J Volcanol Geotherm Res* 95:35–47
54. Mason RM, Starostin AB, Melnik O, Sparks RSJ (2006) From Vulcanian explosions to sustained explosive eruptions: The role of diffusive mass transfer in conduit flow dynamics. *J Volcanol Geotherm Res* 153:148–165. doi:10.1016/j.volgeores.2005.08.011
55. Marsh BD (2000) Reservoirs of Magma and Magma chambers. In: Sigurdsson H (ed) *Encyclopedia of volcanoes*. Academic Press, New York, pp 191–206
56. Mastin GL, Pollard DD (1988) Surface Deformation and Shallow Dike Intrusion Processes at Inyo Craters, Long Valley, California. *J Geophys Res* 93(B11):13221–13235
57. Matthews SJ, Gardeweg MC, Sparks RSJ (1997) The 1984 to 1996 cyclic activity of Lascar Volcano, northern Chile: Cycles of dome growth, dome subsidence, degassing and explosive eruptions. *Bull Volcanol* 59:72–82
58. Mattioli G, Dixon TH, Farina F, Howell ES, Jansma PE, Smith AL (1998) GPS measurement of surface deformation around Soufriere Hills volcano, Montserrat from October 1995 to July 1996. *Geophys Res Lett* 25(18):3417–3420
59. Melnik O (2000) Dynamics of two-phase conduit flow of high-viscosity gas-saturated magma: large variations of sustained explosive eruption intensity. *Bull Volcanol* 62:153–170
60. Melnik O, Barmin A, Sparks RSJ (2005) Dynamics of magma flow inside volcanic conduits with bubble overpressure buildup and gas loss through permeable magma. *J Volcanol Geotherm Res* 143:53–68
61. Melnik O, Sparks RSJ (1999) Non-linear dynamics of lava dome extrusion. *Nature* 402:37–41
62. Melnik O, Sparks RSJ (2002) Dynamics of magma ascent and lava extrusion at Soufriere Hills Volcano, Montserrat. In: Druitt TH, Kokelaar BP (eds) *The eruption of the Soufriere Hills Volcano, Montserrat from 1995 to 1999*. Geological Society, London, Memoir No 21, pp 223–240
63. Melnik O, Sparks RSJ (2005) Controls on conduit magma flow dynamics during lava dome building eruptions. *J Geophys Res* 110(B02209). doi:10.1029/2004JB003183
64. Mériaux C, Jaupart C (1995) Simple fluid dynamic models of volcanic rift zones. *Earth Planet Sci Lett* 136:223–240
65. Murphy MD, Sparks SJ, Barclay J, Carroll MR, Brewer TS (2000) Remobilization origin for andesite magma by intrusion of mafic magma at the Soufriere Hills Volcano. In: Montserrat WI (ed) *A trigger for renewed eruption*. *J Petrol* 41:21–42
66. Muskhelishvili N (1963) *Some Basic Problems in the Mathematical Theory of Elasticity*. Noordhoff, Leiden, The Netherlands
67. Nakada S, Eichelberger JC (2004) Looking into a volcano: drilling Unzen. *Geotimes* 49:14–17
68. Nakada S, Shimizu H, Ohta K (1999) Overview of the 1990–1995 eruption at Unzen Volcano. *J Volcanol Geoth Res* 89:1–22
69. Navon O, Lyakhovsky V (1998) Vesiculation processes in silicic magmas. In: Gilbert J, Sparks RSJ (eds) *The Physics of explosive volcanic eruption*. Geological Society London, Special Publication, vol 145. pp 27–50
70. Neuberg JW, Tuffen H, Collier L, Green D, Powell T, Dingwell D (2006) The trigger mechanism of low-frequency earthquakes on Montserrat. *J Volcanol Geotherm Res* 153:37–50
71. Newhall CG, Melson WG (1983) Explosive activity associated with the growth of volcanic domes. *J Volcanol Geoth Res* 17:111–131
72. Norton GE, Watts RB, Voight B, Mattioli GS, Herd RA, Young SR, Devine JD, Aspinnall WP, Bonadonna C, Baptie BJ, Edmonds M, Harford CL, Jolly AD, Loughlin SC, Luckett R, Sparks RSJ (2002) Pyroclastic flow and explosive activity of the lava dome of Soufriere Hills volcano, Montserrat, during a period of no magma extrusion (March 1998 to November 1999). In: Druitt TH, Kokelaar BP (eds) *The eruption of the Soufriere Hills Volcano, Montserrat from 1995 to 1999*. Geological Society, London, Memoir No 21, pp 467–482
73. Ohba T, Kitade Y (2005) Subvolcanic hydrothermal systems: Implications from hydrothermal minerals in hydrovolcanic ash. *J Volcanol Geotherm Res* 145:249–262
74. Robertson R, Cole P, Sparks RSJ, Harford C, Lejeune AM, McGuire WJ, Miller AD, Murphy MD, Norton G, Stevens NF, Young SR (1998) The explosive eruption of Soufriere Hills Volcano, Montserrat 17 September, 1996. *Geophys Res Lett* 25:3429–3432
75. Roman DC (2005) Numerical models of volcanotectonic earthquake triggering on non-ideally oriented faults. *Geophys Res Lett* 32, doi:10.1029/2004GL021549
76. Roman DC, Neuberg J, Luckett RR (2006) Assessing the likelihood of volcanic eruption through analysis of volcanotectonic earthquake fault-plane solutions. *Earth Planet Sci Lett* 248:244–252
77. Rubin AM (1995) Propagation of magma-filled cracks. *Annu Rev Planet Sci* 23:287–336
78. Saar MO, Manga M, Katharine VC, Fremouw S (2001) Numerical models of the onset of yield strength in crystal–melt suspensions. *Earth Planet Sci Lett* 187:367–379
79. Sahagian D (2005) Volcanic eruption mechanisms: Insights from intercomparison of models of conduit processes. *J Volcanol Geotherm Res* 143(1–3): 1–15
80. Slezin YB (1984) Dispersion regime dynamics in volcanic eruptions, 2. Flow rate instability conditions and nature of catastrophic explosive eruptions. *Vulkanol Seism* 1:23–35
81. Slezin YB (2003) The mechanism of volcanic eruptions (a steady state approach). *J Volcanol Geotherm Res* 122:7–50
82. Sparks RSJ (1978) The dynamics of bubble formation and growth in magmas – a review and analysis. *J Volcanol Geotherm Res* 3:1–37

83. Sparks RSJ (1997) Causes and consequences of pressurization in lava dome eruptions. *Earth Planet Sci Lett* 150:177–189
84. Sparks RSJ (2003) Forecasting Volcanic Eruptions. *Earth and Planetary Science Letters Frontiers in Earth Science Series* 210:1–15
85. Sparks RSJ, Aspinall WP (2004) Volcanic Activity: Frontiers and Challenges. In: *Forecasting, Prediction, and Risk Assessment. AGU Geophysical Monograph "State of the Planet" 150, IUGG Monograph 19*, pp 359–374
86. Sparks RSJ, Murphy MD, Lejeune AM, Watts RB, Barclay J, Young SR (2000) Control on the emplacement of the andesite lava dome of the Soufriere Hills Volcano by degassing-induced crystallization. *Terra Nova* 12:14–20
87. Sparks RSJ, Young SR (2002) The eruption of Soufrière Hills volcano, Montserrat (1995–1999): overview of scientific results. In: Druitt TH, Kokelaar BP (eds) *The eruption of the Soufrière Hills Volcano, Montserrat from 1995 to 1999. Geological Society, London, Memoir No 21*, pp 45–69
88. Sparks RSJ, Young SR, Barclay J, Calder ES, Cole PD, Darroux B, Davies MA, Druitt TH, Harford CL, Herd R, James M, Lejeune AM, Loughlin S, Norton G, Skerrett G, Stevens NF, Toothill J, Wadge G, Watts R (1998) Magma production and growth of the lava dome of the Soufrière Hills Volcano, Montserrat, West Indies: November 1995 to December 1997. *Geophys Res Lett* 25:3421–3424
89. Swanson DA, Holcomb RT (1990) Regularities in growth of the Mount St. Helens dacite dome 1980–1986. In: Fink JH (ed) *Lava flows and domes; emplacement mechanisms and hazard implications*. Springer, Berlin, pp 3–24
90. Voight B, Hoblitt RP, Clarke AB, Lockhart AB, Miller AD, Lynch L, McMahon J (1998) Remarkable cyclic ground deformation monitored in real-time on Montserrat, and its use in eruption forecasting. *Geophys Res Lett* 25:3405–3408
91. Voight B, Sparks RSJ, Miller AD, Stewart RC, Hoblitt RP, Clarke A, Ewart J, Aspinall W, Baptie B, Druitt TH, Herd R, Jackson P, Lockhart AB, Loughlin SC, Lynch L, McMahon J, Norton GE, Robertson R, Watson IM, Young SR (1999) Magma flow instability and cyclic activity at Soufrière Hills Volcano, Montserrat. *Science* 283:1138–1142
92. Walker GPL (1973) Lengths of lava flows. *Philos Trans Royal Soc A* 274:107–118
93. Watson IM et al (2000) The relationship between degassing and ground deformation at Soufriere Hills Volcano, Montserrat. *J Volcanol Geotherm Res* 98(1–4):117–126
94. Watts RB, Sparks RSJ, Herd RA, Young SR (2002) Growth patterns and emplacement of the andesitic lava dome at Soufrière Hills Volcano, Montserrat. In: Druitt TH, Kokelaar BP (eds) *The eruption of the Soufrière Hills Volcano, Montserrat from 1995 to 1999. Geological Society, London, Memoir No 21*, pp 115–152
95. Whitehead JA, Helfrich KR (1991) Instability of flow with temperature-dependent viscosity: a model of magma dynamics. *J Geophys Res* 96:4145–4155
96. Williams SN, Self S (1983) The October 1902 Plinian eruption of Santa Maria volcano, Guatemala. *J Volcanol Geotherm Res* 16:33–56
97. Woods AW, Koyaguchi T (1994) Transitions between explosive and effusive eruption of silicic magmas. *Nature* 370:641–645
98. Wylie JJ, Voight B, Whitehead JA (1999) Instability of magma flow from volatile-dependent viscosity. *Science* 285:1883–1885
99. Yokoyama I, Yamashita H, Watanabe H, Okada H (1981) Geophysical characteristics of dacite volcanism – 1977–1978 eruption of Usu volcano. *J Volcanol Geotherm Res* 9:335–358

### Books and Reviews

- Dobran F (2001) *Volcanic Processes: Mechanisms In Material Transport*. Kluwer, New York, pp 620
- Gilbert JS, Sparks RSJ (eds) (1998) *The Physics of Explosive Volcanism*. Special Publication of the Geological Society of London, vol 145, pp 186
- Gonnermann H, Manga M (2007) *The fluid mechanics inside a volcano*. *Ann Rev Fluid Mech* 39:321–356
- Mader HM, Coles SG, Connor CB, Connor LJ (2006) *Statistics in Volcanology*. IAVCEI Publications, Geological Society Publishing House, p 296

## Volcanic Eruptions: Stochastic Models of Occurrence Patterns

MARK S. BEBBINGTON

Massey University, Palmerston North, New Zealand

### Article Outline

Glossary

Definition of the Subject

Introduction

Data

Temporal Models

Volcanic Regimes

Spatial Aspects

Yucca Mountain

Interactions with Earthquakes

Future Directions

Bibliography

### Glossary

**Onset** The beginning of an eruption. The term **event** will be used interchangeably.

**Repose** Periods during which an eruption is not in progress.

**Onset time** Time at which an eruption begins. The  $i$ th onset time will be denoted  $t_i$ , for  $i = 1, \dots, n$ .

**Inter-onset time** Time between successive onsets, denoted by  $r_i = t_{i+1} - t_i$ , for  $i = 1, \dots, n - 1$ . In many papers this is termed the **repose time**, which we shall use interchangeably, although the latter is strictly the time between the *end* of one eruption and the *start* of the next.

**Polygenetic vent** Site of multiple events, in contrast to **monogenetic vents**, at which only one eruption occurs. The latter occur predominately in **volcanic fields**.

**Absolute time** is denoted by  $t$  or  $s$ , with the latter usually being the time of the last known event. The **elapsed time** since the last known event is denoted by  $u$ , or, equivalently,  $t - s$ .

**A parameter estimate** of the parameter  $\lambda$ , say, is denoted by  $\hat{\lambda}$ .

### Definition of the Subject

Depending on the definition, there are approximately 1300 Holocene (last 10,000 years) active volcanoes [83], some 55–70 of which are typically active in any given year. Being tectonic phenomena, they are spatially over-represented along littorals and in island arcs where populations congregate, and millions of lives are potentially at risk. Hence

the quantitative forecasting of hazard has long been a key aim of volcanology. This requires models to quantify the likelihood of future outcomes, which may involve the size, location and/or style of eruption, in addition to temporal occurrence. While a wide range of complex deterministic models exist to model various volcanic processes, these provide little in the way of information about future activity. Being the (partially) observed realization of a complex system, volcanological data are inherently stochastic in nature, and need to be modeled using statistical models. To paraphrase Decker [25] concerning potentially active volcanoes, geologists can find out what did happen, geophysicists can determine what is happening, but to find out what might happen requires statistics. Reminding ourselves that “all models are wrong, but some models are useful”, the most useful models for volcanic eruptions are ones that provide new insight into the physical processes, and can in turn be informed by our present understanding of those physical processes.

### Introduction

The hazard is a product of many elements, including the likelihood of an eruption, the distribution of the eruption size, and the style of the eruption. Depending on the nature of the volcanic products, climatological or geographical factors may also play a part.

The style of an eruption strongly determines the relative risks to life and property. Effusive eruptions of molten lava pose little threat to life with, for example, only one life being lost due to a volcanic eruption on Hawaii in the 20th century. However, in the same time period 5% of the island has been covered by new lava flows [25]. Explosive eruptions, in particular pyroclastic flows, travel at high speeds and can cause extensive fatalities such as the 29,000 killed in the Mont Pelee (West Indies) eruption of 1902. Lahars, or mudflows, such as those that killed 25,000 people in the Nevado del Ruiz (Columbia) eruption of 1985, are also a major danger.

Forecasting volcanic eruptions is easier than forecasting earthquakes on several grounds. First, at least in the case of polygenetic volcanoes, there need be no spatial element involved. Second, it is possible using geological techniques to look backward (often a long way, depending on geography and finances) to see what happened in the past. Thirdly, and of most importance for short-term warning, magma intrusions can be detected by seismic networks, although not all seismic signals indicate magma of course, and other monitoring techniques [25].

In this article we will limit our consideration to models of eruption occurrence, omitting techniques for fore-

casting the nature and effect of the eruption. As the track record of a potentially active volcano provides the best method of assessing its future volcanic hazards on a long-term basis [21,25], we will first briefly review the provenance and characteristics of the data available. Various taxonomies for stochastic models are possible. As the majority of models applied to volcanoes are purely temporal, and this also includes the simplest models, we will look at these first. Such models can be further broken down into (nonhomogeneous) Poisson processes, which consider absolute time relative to some fixed origin, and renewal processes, which consider only the time since the previous event. Markov processes, where the state of the process moves between various eruptive and repose states, and models based on eruptive volume, are the simplest ways of incorporating some element of eruption size into the model. Fractals have been used as a descriptive technique for volcanoes, and are reviewed next. The above models, with the exception of the nonhomogeneous Poisson process, all assume that the volcano is in steady state, i. e., that the activity pattern does not change over time. For volcanoes with lengthy records, this appears to be an untenable assumption over longer time scales [99]. This raises the issues of, first, how to detect a departure from stationary stochastic behavior, or change in volcanic regime and, secondly, how to model it, which are discussed in Sect. “**Volcanic Regimes**”, illustrated with results from Mount Etna. While purely temporal models are suitable for polygenetic volcanoes, monogenetic volcanic fields require a spatial element, to model the location, as well as the time, of future events. This is usually estimated using kernel smoothing, but in cases where the location can be treated as a discrete variable, such as summit versus flank eruptions, Markov chain techniques can be used. These are reviewed, along with methods for detecting clustering and linear alignments, in Sect. “**Spatial Aspects**”. A very high profile, particularly in the USA, exercise in volcanic hazard estimation concerns the risk to a proposed high-level radioactive waste repository planned for Yucca Mountain. The history of published hazard estimates is used in Sect. “**Yucca Mountain**” to illustrate the application of both temporal and spatio-temporal models. The final substantive section looks at methods for detecting links between volcanic eruptions and large earthquakes. Determining whether correlation in the occurrence patterns exists is a necessary precursor to formulating models to take advantage of any potential new information from the additional process.

The focus of this article is squarely on the techniques. Hence the applications jump around a bit, as they are those presented in the original paper(s) dealing with the tech-

nique. The examples of Mount Etna and Yucca Mountain were selected for more detailed examination partly on the basis that many, somewhat contradictory, results exist. Different models make different assumptions, and vary in how much information they can extract from data. In addition, the data used often varies from study to study, and the sensitivity of models to data is important, but too often ignored. Subjective judgments on the merits of various techniques is minimized, except for comments on details necessary to their application.

## Data

Information about past eruptions forms the basis of any stochastic model for future behavior. Using such data poses a number of challenges because of its inhomogeneity, particularly in the longer records. This inhomogeneity is due to incomplete observation, which becomes more incomplete the further back one goes, and by variations in dating and measurement precision, again becoming less precise the further into the past one looks. We will now outline the nature and limitations of the available information.

There are several functional distinctions in the types of data used in volcanic hazard models. The most fundamental is at the question of scale. Certain volcanoes, notably Stromboli (which lends its name to the behavior) and Soufriere Hills, erupt in distinct events separated by minutes or hours. Hence a catalog of hundreds of events may only cover a few days of activity. More commonly, eruptions are events with durations ranging from days to years. Eruptions are usually modeled simply as a point process of onsets, but a few studies consider duration, volume, or spatial extent.

A global catalog of Holocene volcanism is maintained by Seibert and Simkin [82]. This gives onset dates, of various precision, and Volcanic Explosivity Index (VEI) for all known eruptions. The VEI, as a logarithmic measuring scale of eruption size analogous to the magnitude of an earthquake, was proposed by Newhall and Self [68]. It is assigned to historical eruptions on the basis of (in decreasing order of reliability): explosion size; volume of ejecta; column height, classification (‘Hawaiian’, ‘Strombolian’, ‘Vulcanian’, ‘(Ultra-)Plinian’); duration; most explosive activity type; tropospheric and stratospheric injection. Many historical (i. e., observed and recorded at the time) observations also have a duration, and some an estimate of eruptive volume. Estimates of volume, where they exist, typically contain uncertainties of 10–50% [101] or more [54] and, provided the eruption styles are similar, the duration of an eruption can be used as a proxy [3,67].

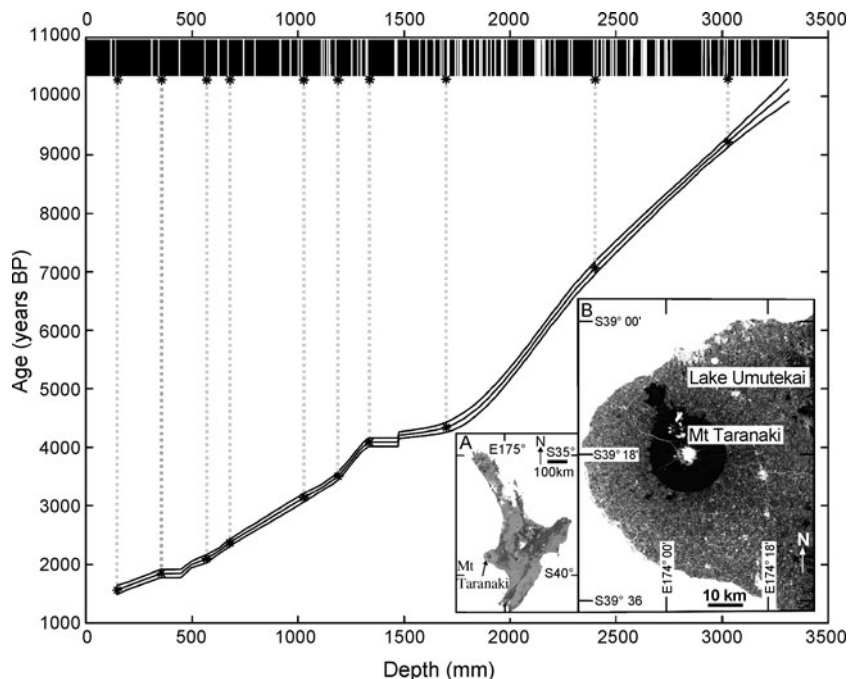


The frequency of observed volcanic eruptions has increased over the last few hundred years, although this is largely an artifact of the global population increase [84], which indicates that many of the historical records are incomplete. A number of published catalogs for individual volcanoes are of greater detail, and probably more complete. Some of the better known are Etna [65,67,74] and Vesuvius [11,77] (Italy), Ruapehu (New Zealand) [5], Colima (Mexico) [27,63] and Kilauea/Mauna Loa (Hawaii) [46].

While it is reasonably easy, although sometimes controversial, to define events in an historical catalog, it is not so easy in pre-historical cases. Ideally events correspond to eruptions, but subsequent geologic processes can obscure, or even obliterate, the evidence of previous eruptions. The identification of events is dependent on looking in the right place. Pre-historical eruptions can be identified from a combination of stratigraphy (which provides an ordering for the eruptions) and dating techniques. Radiocarbon dating is feasible for material within the last 50,000 years, and provides an estimated age and (Gaussian) error given in  $C^{14}$  years. See [28] for an example of such a catalog. Because of the variable rate of  $C^{14}$  deposition, this has to be converted to calendar years [8], in

the process losing the Gaussian distribution (and possibly even becoming disjoint). One way [94] of dealing with the resulting untidiness is to treat the data in Monte Carlo fashion, repeatedly drawing random realizations of the eruption sequence from the distributions (ensuring that the stratigraphic ordering is not violated), and fitting the model to each realization to provide a population of hazard models. Beyond 50,000 years, dating is via a number of other techniques, such as K-Ar radiometric age determinations which, with resolutions of the order of  $10^6 - 10^7$  years, are much less precise. Conway et al. [20] take the approach of simulating event dates from a uniform distribution, bounded by the  $\pm 2\sigma$  error limits on the dates, in order to determine the uncertainty in the estimated recurrence rate.

Because of the expense, and due to the difficulty of finding suitable material in the sample to date, it is common to date only a portion of the samples. Geomorphology is then used to interpolate or bound the unknown dates. For example, in a single core, a spline can be fitted to the depth and known dates, representing the sediment deposition rate, and the unknown dates calculated from their depths [94], as shown in Fig. 1. This gets much more complicated when there is a spatial element involved [14].



Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 1

Radiocarbon age versus depth model (mm) for tephra layers in the sediment core from Lake Umutekai (*inset A*: North Island of New Zealand and *inset B*: Taranaki – the lake is directly above the “k-” in “Lake”). The core stratigraphy is displayed at the top of the plot with white representing tephra deposits. Radiocarbon dated layers are indicated by dotted lines

While deposition cores can provide a complete (above a certain minimum size), and ordered tephra record for a given site, the record extracted from another site may only partially overlap. Various criteria of tephra composition and geochemistry can be used to identify common events. Composite records from many sources will almost certainly be incomplete due to missing events, and may well unknowingly contain duplicates of other events.

**Temporal Models**

Models for the occurrence of volcanic eruptions are based on the idea that the past behavior of a volcano is the best predictor of its future activity. This was dramatically illustrated by Crandell et al. [21] in their somewhat non-quantitative forecast of an eruption of Mount St Helens:

“The repetitive nature of the eruptive activity at Mount St. Helens during the last 4000 years, with dormant intervals typically of a few centuries or less, suggests that the current quiet period will not last a thousand years. Instead, an eruption is likely within the next hundred years, possibly before the end of this century.”

Published in 1975, as a warning that dormant intervals can be a recurring element in the life of a volcano, this foreshadowed the 1980 eruption and subsequent activity.

Let the number of events in a time interval  $(s, t)$  be denoted  $N(s, t)$ , where we use the shorthand  $N(t)$  if  $s = 0$ .

The *intensity* of a process is

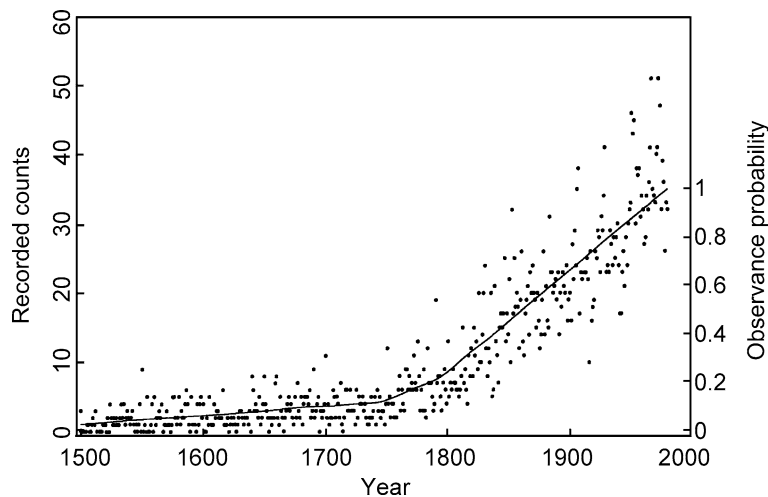
$$\lambda(t) = \lim_{\Delta t \downarrow 0} \frac{\Pr(N(t, t + \Delta t) = 1)}{\Delta t}, \tag{1}$$

i. e., in the short period of time  $(t, t + \Delta t)$ , the probability of an event is approximately  $\lambda(t)\Delta t$ .

Guttorp and Thompson [35] provided a non-parametric method of estimating the intensity, based on the observed onset counts  $Y_k$  in intervals of length  $\Delta$ , indexed by  $k = 1, \dots, K$ . They first estimate the reporting probability  $p(t)$  by smoothing the observed onset counts (see [86] for an alternative approach assuming a non-decreasing reporting probability), and calculate  $Z_k = Y_k / \int_{(k-1)\Delta}^{k\Delta} p(t)dt$ . This is a reconstruction of the underlying process obtained by re-inflating the observed counts by the corresponding observance probability. Assuming that the intensity is locally fairly smooth, an estimate, based on the hanning operation from time-series analysis, is

$$\hat{\lambda}(j\Delta) = \frac{\sum_{i=1}^{K-j-1} Z_i Z_{i+j-1} + 2 \sum_{i=1}^{K-j} Z_i Z_{i+j} + \sum_{i=1}^{K-j+1} Z_i Z_{i+j-1}}{4\Delta(K-j) \sum_{k=1}^K Z_k / K}. \tag{2}$$

Application to the global catalog, assuming a constant rate of underlying events, produced the estimated observance probability in Fig. 2 that increases slowly from 1500 until the second half of the 18th century, and then more sharply at a constant rate to an assumed probability  $p(1980) = 1$ . The resulting estimated intensity for global eruption starts



Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 2  
 Recorded counts (left-hand scale) and estimated observance probability (right-hand scale) for yearly eruption starts in the Simkin catalog from A.D. 1500 on. Reprinted with permission from the *Journal of the American Statistical Association*. Copyright 1991 by the American Statistical Association

did not differ from a Poisson (constant intensity) process. However, in order to see if more structure could be observed in a series of eruptions from a restricted area with homogeneous tectonic structure, the methodology was applied to a catalog of Icelandic eruptions. As expected, the estimated observance probability for the last 200 years was approximately constant, which resulted in a periodicity of about 40 years in the estimated intensity, as shown in Fig. 3.

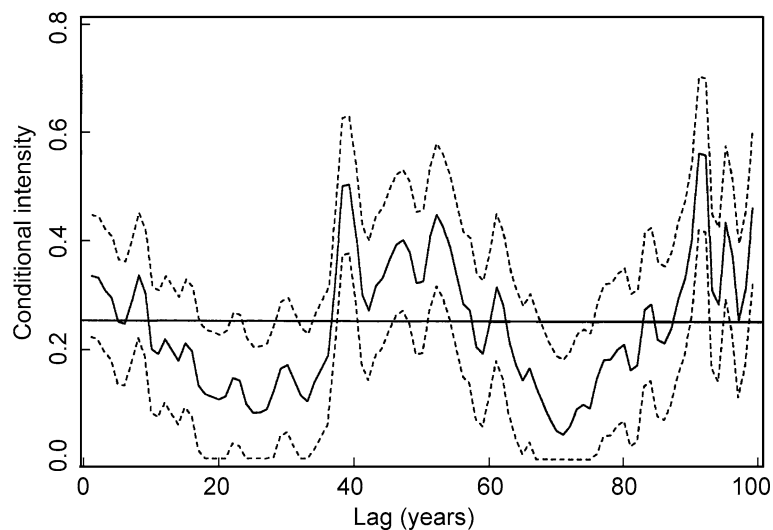
There have been a few other attempts to discern periodic fluctuations in volcanic activity. Mauk and Johnston [62] examined the correlation of the onset dates of 680 eruptions 1900–1971 with the fortnightly solid earth tide. There was a statistically significant tendency for both andesitic and basaltic eruptions to occur at the maxima of the earth tide. In addition, basaltic eruptions had a significant peak at the minima of the earth tide. A more detailed examination of 18 Japanese volcanoes indicated that this may be related to Bouguer anomalies. Similarly, Martin and Rose [54], in Fig. 4, found that eruption onsets of Fuego Volcano, Guatemala were correlated with the fortnightly lunar tidal maxima, with 23 out of 48 eruptions occurring within  $\pm 2$  days of the peak gravity acceleration. Casetti et al. [12] binned eruptions at Mt Etna and examined the resulting statistics, in particular the month of an eruption. Eruptions were shown to be significantly more likely during March–April and November, when there is an increase in the earth-velocity rate. Although

Stothers [89] repeated this analysis for a global catalog of eruptions with  $VEI \geq 3$ , 1500–1980, and detected no significant monthly or seasonal variation, Mason et al. [61] found that seasonal fluctuations account for 18% of the historical average monthly eruption rate. The idea of a deterministic cycle is incorporated in a renewal model by Jupp et al. [45], but the result was not fitted to data.

### Poisson Processes

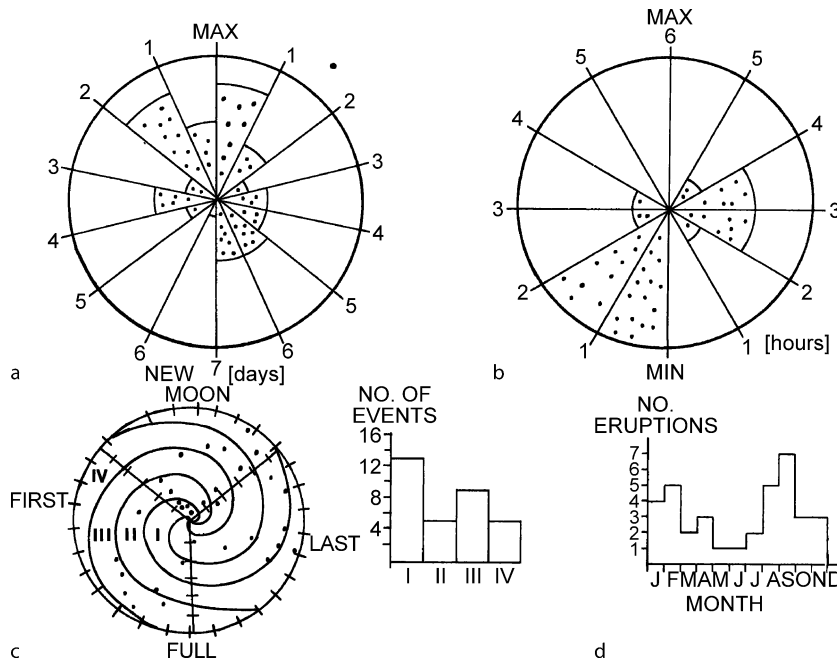
The simplest model is a *simple (homogeneous) Poisson process*, which is characterized by  $\lambda(t) = \lambda_0$ , a constant independent of  $t$ . In this case  $N(t)$  has a Poisson distribution with mean  $\mu = \lambda_0 t$ , and the inter-event times have an exponential distribution with mean  $1/\lambda_0$ . This is sometimes referred to as *Poissonian behavior*. The parameter is estimated as  $\hat{\lambda}_0 = N(T)/T$ , for an observed history spanning the time interval  $(0, T)$ .

In order to weaken the requirement that the mean and variance of  $N(t)$  be equal, Ho [36] suggested a Bayesian approach, in which the parameter  $\lambda_0$  has a gamma prior distribution. This leads to  $N(t)$  having a negative binomial distribution. Solow [87] augmented this to an empirical Bayes formulation by suggesting that an informative prior could be constructed from the eruption records of a group of similar volcanoes. Bebbington and Lai [6] introduced a generalized negative binomial distribution incorporating



Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 3

Estimated conditional intensity function for Icelandic eruption starts using data from the last 200 years. The *dashed lines* are approximate 95% pointwise confidence bands under the assumption of a homogeneous Poisson process. Reprinted with permission from the *Journal of the American Statistical Association*. Copyright 1991 by the American Statistical Association



Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 4

**a** Lambert equal area polar histogram showing the percentage of eruptions of Fuego since 1800 in relationship to the fortnightly tidal maximum of the vertical gravity acceleration. **b** Lambert equal-area polar histogram showing the percentage of all the known times of the beginning of eruptions since 1957 compared to the 12.4-hour (semi-diurnal) minimum of the vertical gravity acceleration. **c** A spiral diagram indicating the position of an eruption in relation to the synodic month (days plotted around the circumference) and the reduced anomalistic age (days plotted as the radius, with perigee at the center), in which it occurred. There are 27.5 days in the anomalistic cycle, and 29.6 days in the synodic cycle. Right of the spiral diagram, is a histogram of the four spirals of the above diagram, which are drawn around  $\pm 3.5$  days of the maxima and minima of the 2 cycles. **d** Monthly histogram of Fuego’s eruptions which have occurred since 1800. Reprinted from [54] © 1981 Elsevier B.V., with permission from Elsevier

serial dependence and applied it to eruptions from Mount Sangay (Ecuador).

The Poisson process is stationary in time, in that the distribution of the number of events in an interval depends only on the length of the interval, not its location. In other words, events occur at the same average rate at all times, i.e., there is no trend in the occurrence rate with time. Marzocchi [55] analyzed possible trends by calculating the (biased) autocorrelation function

$$a(s) = \frac{\sum_{i=1}^{n-1-|s|} (r_i - \bar{r})(r_{i+|s|} - \bar{r})}{\sum_{i=1}^{n-1} (r_i - \bar{r})^2}, \tag{3}$$

for  $s = 0, \pm 1, \dots, \pm m$ , where  $r_i, i = 1, \dots, n - 1$  is the time between the  $i$ th and  $(i + 1)$ th onsets,  $\bar{r} = (1/(n - 1)) \sum_{i=1}^{n-1} r_i$  and  $m < n - 1$ . Under the null hypothesis that the process is purely random, the distribution of  $a(s)$  is asymptotically normal with mean zero and variance  $1/(n - 1)$  for  $s \neq 0$ .

More generally, a process can be a *nonhomogeneous Poisson process*, where  $N(t)$  has a Poisson distribution with

mean  $\mu(t) = \int_0^t \lambda(s) ds$ . Ho [37] used an example of a nonhomogeneous Poisson process known as the *Weibull Process* [2], with

$$\lambda(t) = \frac{\beta}{\theta} \left( \frac{t}{\theta} \right)^{\beta-1}, \tag{4}$$

where  $t$  is absolute time with respect to some fixed origin, not the elapsed time since the previous eruption. This includes the homogeneous Poisson process as a special ( $\beta = 1$ ) case, while if  $\beta \neq 1$  the process is non-stationary. The intensity (4) is monotonic, and hence can model either an increase or decrease in volcanic activity, but not both. Note that the parameter estimates,

$$\hat{\beta} = N(T) / \sum_{i=1}^{N(T)} \ln(T/t_i) \quad \text{and} \quad \hat{\theta} = T/N(T)^{1/\hat{\beta}}, \tag{5}$$

for the intensity (4) are sensitive to the position of the time origin [4]. Furthermore, estimates of  $\hat{\beta} > 2$ , which are quite feasible, indicate a constantly accelerating, or

convex, intensity. Neither of these properties is particularly desirable from a physical viewpoint, and hence the Weibull process is not suitable to model entire volcanic histories. Salvi et al. [73] used a statistical technique to identify that the most recent part of the Mount Etna sequence was non-stationary, and fitted the Weibull process to it. The Weibull process can be tested for goodness of fit to a series of data [4] either via a standard  $\chi^2$  test, or by using the fact that  $x_1, \dots, x_{n-1}$ , where  $x_i = \ln(t_n/t_{n-i})$ , should be a random sample from an exponential distribution of unknown mean.

Jaquet et al. [44] proposed a temporal occurrence model using the Cox process. This is a doubly stochastic process with

$$\Pr(N(s, s + t) = k) = \mathbb{E} \left( \frac{Z(s, s + t)^k}{k!} e^{-Z(s, s + t)} \right), \quad (6)$$

where

$$Z(s, s + t) = \int_s^{s+t} \omega(t) dt, \quad (7)$$

and  $\omega(t)$  is the conditional random frequency of events. The process of determining  $\omega(t)$  is not spelt out in detail, involving geostatistical methods, simulation, variograms and a conditioning technique. The advantage of (6) over the Poisson process is that it introduces the possibility of correlation in time.

### Renewal Processes

The homogeneous Poisson process without trend is a special case of a *renewal process*. In a renewal process, the intervals between events are independent and identically distributed, and so

$$\lambda(t) = \frac{f(t - s; \theta)}{1 - F(t - s; \theta)}, \quad t > s, \quad (8)$$

where the most recent event occurred at time  $s < t$ , and  $f = F'$  is a density with parameter vector  $\theta$ . This was first suggested for eruption onsets by Wickman [103], who coined the term *age-specific eruption rate* for (8). The character of the renewal process is that only the elapsed time since the last eruption controls the time to the next eruption. Previous eruptions exert an influence only through their contribution to the parameter estimates  $\hat{\theta}$ . A number of tests to check whether a series of eruptions is consistent with a renewal process can be found in the paper by Reymont [72]: One first checks for the existence of a trend using the standard normal statistic

$$U = \frac{\sum_{i=1}^n t_i/n - T/2}{T\sqrt{n/12}}, \quad (9)$$

where onsets are observed at times  $0 < t_1 < \dots < t_n < T$ . If no trend is observed, one can then examine the serial correlation coefficients

$$\rho_j = \frac{\text{Cov}(\mathbf{r}, \mathbf{r}^{(j)})}{\text{Var}(\mathbf{r})}, \quad (10)$$

where  $\mathbf{r} = (r_1 \dots r_{n-1-j})$ ,  $\mathbf{r}^{(j)} = (r_{1+j} \dots r_{n-1})$ , and  $r_i$  is the time between the  $i$ th and  $(i + 1)$ th onsets. For a renewal process, there should be no correlation, and  $\text{Var}(\rho_j) \sim 1/(n - j)$ . Hence a test of  $\rho_j = 0, j = 1, 2, \dots$  can be performed using the standard normal statistic  $\rho_j \sqrt{(n - j)}$ . One can also examine the periodogram [72]. The special case of a Poisson process can be tested in many ways, the easiest of which is to test the inter-onset intervals for exponentiality using, for example, the Kolmogorov–Smirnov statistic of the maximum distance between the empirical and hypothesized distribution functions. Other tests are mentioned in [4,5].

Given a density  $f(u; \theta)$  and observed inter-onset times  $r_i, i = 1, \dots, n - 1$ , the parameters  $\theta$  can be estimated by maximum likelihood. That is, the values are chosen, either algebraically or numerically, to maximize the likelihood

$$L(r_1, \dots, r_{n-1}, t - s; \theta) = [1 - F(t - s; \theta)] \prod_{i=1}^{n-1} f(r_i; \theta), \quad (11)$$

where the first term accounts for the current unfinished interval, and is often omitted. Occasionally other methods are used. The method of moments involves setting the parameter values to equalize the observed and theoretical moments. As many moments are used as one has parameters. For example, two parameters can be fitted using the mean and standard deviation. An alternative is based on minimizing the sum of squared differences between an observed curve and that produced by the model. This is usually done using linear regression, but direct numerical optimization is possible.

If  $f$  is the exponential density  $f(u) = \nu e^{-\nu u}$ , then

$$\lambda(t) = \frac{\nu e^{-\nu(t-s)}}{1 - (1 - e^{-\nu(t-s)})} = \nu, \quad (12)$$

and we recover the homogeneous Poisson process. Klein [46] tested the eruptive patterns of Hawaiian volcanoes against a Poisson process (random model) using this formulation, and showed that reposes following large eruptions differed significantly from the whole. The Eq. (12) is sometimes referred to as the memoryless property, as it says that the time elapsed since the last eruption provides no information about the time of the next eruption.

Thorlaksson [93] suggested that an over-dispersed (standard deviation  $\sigma$  greater than the mean  $\mu$ ) sequence of inter-onset intervals should be modeled by the Pareto density  $f(u) = a(1 + bu)^{-a/b-1}$ ,  $u > 0$ , with intensity

$$\lambda(t) = \frac{a}{1 + b(t-s)}, \quad t > s, \tag{13}$$

where  $\hat{a} = 2\sigma^2/(\mu^3 + \mu\sigma^2)$  and  $\hat{b} = (\sigma^2 - \mu^2)/(\mu^3 + \mu\sigma^2)$  are estimated by the method of moments. This models a decrease in eruption probability with time, and was applied to Colima by Medina Martinez [63]. For under-dispersed inter-onset intervals, Thorlaksson [93] suggested the density

$$f(u) = \begin{cases} (a + bu) \exp(-au - (b/2)u^2) & \sqrt{4/\pi - 1} < \sigma/\mu < 1 \\ b(u - a) \exp(-(b/2)(u - a)^2) & \sigma/\mu < \sqrt{4/\pi - 1} \end{cases}, \quad u > 0, \tag{14}$$

leading to the intensity

$$\lambda(t) = \begin{cases} a + b(t-s) & \sqrt{4/\pi - 1} < \sigma/\mu < 1 \\ b(t-s-a) H(t-s-a) & \sigma/\mu < \sqrt{4/\pi - 1} \end{cases}, \quad t > s, \tag{15}$$

where

$$H(x) = \begin{cases} 1 & x > 0 \\ 0 & x \leq 0 \end{cases} \tag{16}$$

is the Heaviside function. These imply an increase in eruption probability with time, while the case  $\sigma/\mu < \sqrt{4/\pi - 1}$  adds a loading time  $a$ , and the parameters can be estimated as  $\hat{a} = \mu - \sigma/\sqrt{4/\pi - 1}$  and  $\hat{b} = (2 - \pi/2)/\sigma^2$ . This latter case was applied to the volcanoes Hekla and Katla, which had ratios  $\sigma/\mu$  of 0.44 and 0.43 respectively. Of the other 28 volcanoes examined, none had a ratio less than 0.76, and 23 had a ratio of greater than unity.

Settle and McGetchin [78] fitted a Gaussian density to the inter-onset times of eruptions at Stromboli over four days. Note that the formulation (8) is always strictly positive, even for distributions such as the Gaussian whose support includes the negative axis. However, fitting and simulating such processes presents certain practical difficulties.

Bebbington and Lai [4] fitted renewal models to a number of volcanoes, using the Weibull density with scale parameter  $\beta$  and shape parameter  $\alpha$ ,

$$f(u) = \alpha\beta(\beta u)^{\alpha-1} \exp(-(\beta u)^\alpha), \quad u > 0. \tag{17}$$

The intensity is then

$$\lambda(t) = \alpha\beta(\beta(t-s))^{\alpha-1}, \quad t > s. \tag{18}$$

We see that while  $\alpha = 1$  is the exponential distribution,  $\alpha < 1$  corresponds to an ‘over-dispersed’, or clustering, distribution (cf. [34]), and  $\alpha > 1$  to a more periodic distribution with a mode at  $u = \beta^{-1}(1 - 1/\alpha)^{1/\alpha}$ . The density and intensity are shown in Fig. 5.

De la Cruz-Reyna and Carrasco-Nunez [28] suggested using the gamma density, which also allows for clustering and periodicity, but the computational details are more complex. Also, the Weibull is the consequence of a material failure model [98], and in any case, little qualitative difference is usually observed between Weibull and gamma distributions for small samples, such as eruptive records. Note that the intensity (18) is monotonic, and can model either increasing probability of eruption as the repose time increases, such as at Hekla [103], or decreasing probability with increasing repose time, such as at Colima [63]. Bebbington and Lai [4,5] provided a number of tests for assessing the applicability of the Weibull distribution, based on the fact that  $r_i^\beta$ , where  $r_i = t_{i+1} - t_i$ , should be exponentially distributed.

The lognormal density,

$$f(u) = \frac{1}{u\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln u - \mu)^2}{2\sigma^2}\right], \quad u > 0, \tag{19}$$

with intensity

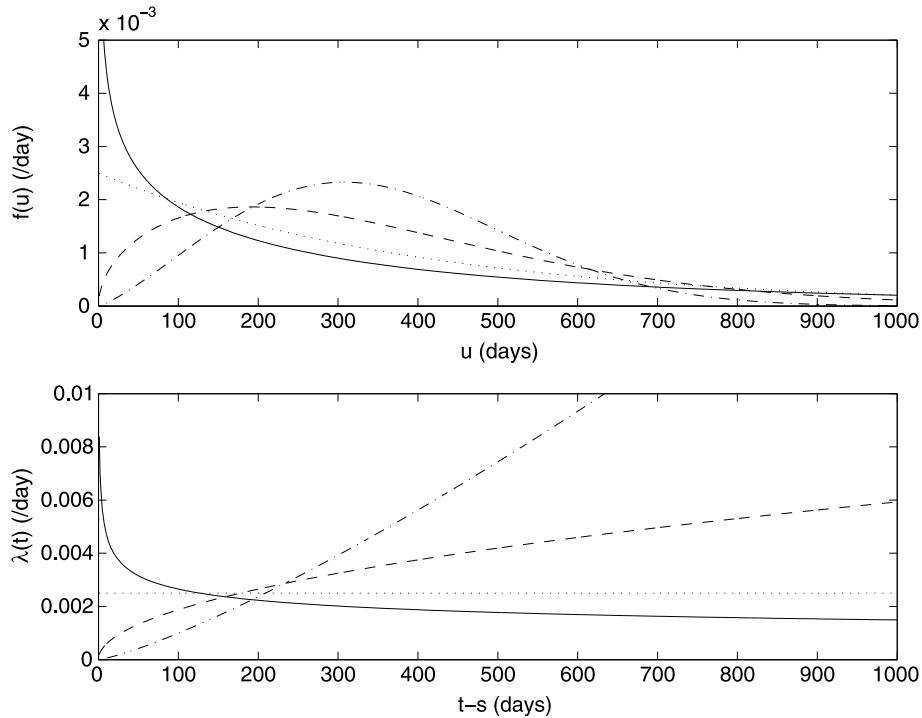
$$\lambda(t) = \frac{\frac{\sqrt{2}}{(t-s)\sigma\sqrt{\pi}} \exp\left[-\frac{(\ln(t-s) - \mu)^2}{2\sigma^2}\right]}{\operatorname{erfc}\left(\frac{\ln(t-s) - \mu}{\sqrt{2}\sigma}\right)}, \quad t > s, \tag{20}$$

was also considered by Bebbington and Lai [4], and used by Marzocchi and Zaccarelli [59] and Eliasson et al. [30]. As the density (19) is unimodal at  $u = \exp(\mu - \sigma^2)$ , this works better for non-clustering events, such as the very large eruptions considered in [59]. Eliasson et al. [30] noted significant departure from the distribution in one of their models. The density and intensity are shown in Fig. 6. Note that (20) is ‘upside-down bathtub shaped’, i. e., that as  $t$  increases it rises to a maximum and then declines.

The log-logistic density,

$$f(u) = \frac{\eta\gamma u^{\gamma-1}}{(1 + \eta u^\gamma)^2}, \quad u > 0, \tag{21}$$

with a mode at  $u = \max\{0, (\gamma - 1)/((\gamma + 1)\eta)^{1/\gamma}\}$  was used by Connor et al. [18] to model the intervals between Vulcanian explosions of Soufriere Hills volcano, Montserrat.



Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 5  
 Density  $f$  and conditional intensity  $\lambda$  for the Weibull distribution with  $\beta = 0.0025$  and  $\alpha = 0.75$  (solid line),  $\alpha = 1$  (dotted line),  $\alpha = 1.5$  (dashed line) and  $\alpha = 2.25$  (dot-dash line)

The density and the intensity

$$\lambda(t) = \frac{\eta\gamma(t-s)^{\gamma-1}}{1 + \eta(t-s)^\gamma}, \quad t > s, \tag{22}$$

are shown in Fig. 7. The intensity (22) has a maximum at  $t = s + \max\{0, ((\gamma - 1)/\eta)^{1/\gamma}\}$ , prior to which it is increasing, subsequently decreasing. Thus we can see that the logarithistic distribution possesses some of the character of the Weibull and of the lognormal distributions.

Pyle [71] used the power-law density

$$f(u | u > c) = \frac{u^{-1-1/b}}{bc^{-1/b}}, \quad u > c > 0 \tag{23}$$

to model the tail of the inter-onset distribution. Note that the density is not defined for  $c = 0$ . The intensity is

$$\lambda(t) = \frac{1}{b(t-s)}, \quad t > s, \tag{24}$$

from which we see that the parameter  $b$  controls the rate of decay of the intensity, or equivalently, the ‘thickness of the tail’.

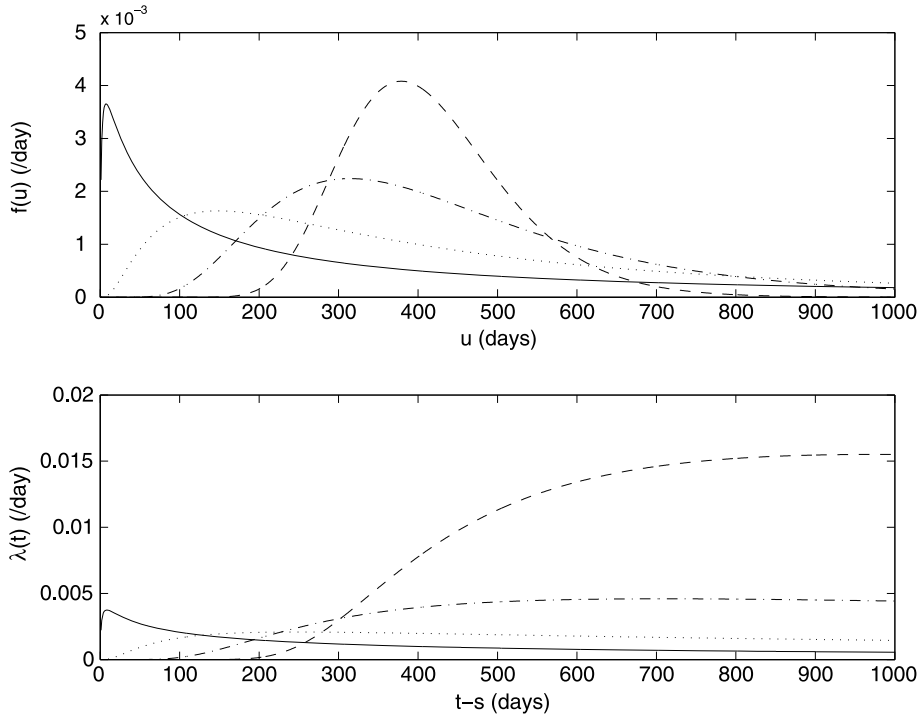
Turner et al. [94] fitted a renewal density consisting of a mixture of Weibull densities, as exemplified in

Fig. 8, to model a multimodal repose distribution. Cronin et al. [22] proposed a variant on this, where onsets occurred in ‘episodes’, each of variable numbers of eruptions, with different repose distributions corresponding to inter- and intra-episode repose.

The drawback of renewal models is that they commonly fail to explain variations in eruption rate, corresponding to changes in activity level, although Wickman [103] did suggest using a small number of different levels. This idea was taken up in [3], where a renewal process with parameters controlled by a hidden Markov model was fitted to flank eruptions from Mt Etna. Using the Weibull distribution (17) for the inter-onset times, the preferred model had long sequences of approximately Poisson behavior, interspersed with shorter sequences of more frequent and regular eruptions.

### Markov Processes

A *Markov process* (continuous time Markov chain)  $Z(t), t \geq 0$  has transition intensity matrix  $Q = (q_{ij})$ ,



Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 6  
 Density  $f$  and conditional intensity  $\lambda$  for the lognormal distribution with  $\mu = 6$  and  $\sigma = 2$  (solid line),  $\sigma = 1$  (dotted line),  $\sigma = 0.5$  (dashed line) and  $\sigma = 0.25$  (dot-dash line)

where

$$q_{ij} = \lim_{t \downarrow 0} \frac{\Pr(Z(s+t) = j \mid Z(s) = i)}{t} \geq 0, \quad (25)$$

for all  $i \neq j$ , and  $-q_i = q_{ii} = -\sum_{j \neq i} q_{ij}$ . Thus the sojourn time in state  $i$  is exponential with mean  $1/q_i$ , and when it ends, the process moves to state  $j$  with probability  $p_{ij} = q_{ij} / \sum_{k \neq i} q_{ik}$ .

Wickman [104] proposed the use of Markov processes to model eruption-repose patterns. This entailed defining a number of eruptive and repose states, with prescribed permitted transitions between them. The initial models [104] allowed for one or two magma chambers, with eruptions and ‘solidification’ of the vent occurring in Poisson processes of differing rate. In the two-state case this solidification can only be removed by an eruption from the second, larger and deeper, magma chamber. If the two magma chambers erupt at rates  $\nu$  and  $\mu$  ( $\nu > \mu$ ), respectively, and the vent solidifies at rate  $\eta$ , then the eruption rate can be derived as

$$\lambda(t) = \frac{\mu\eta + \nu(\mu + \nu + \eta) \exp[-(\nu + \eta)(t - s)]}{\eta + \nu \exp[-(\eta + \nu)(t - s)]}, \quad t > s, \quad (26)$$

where the previous eruption occurred at time  $s$ . The pa-

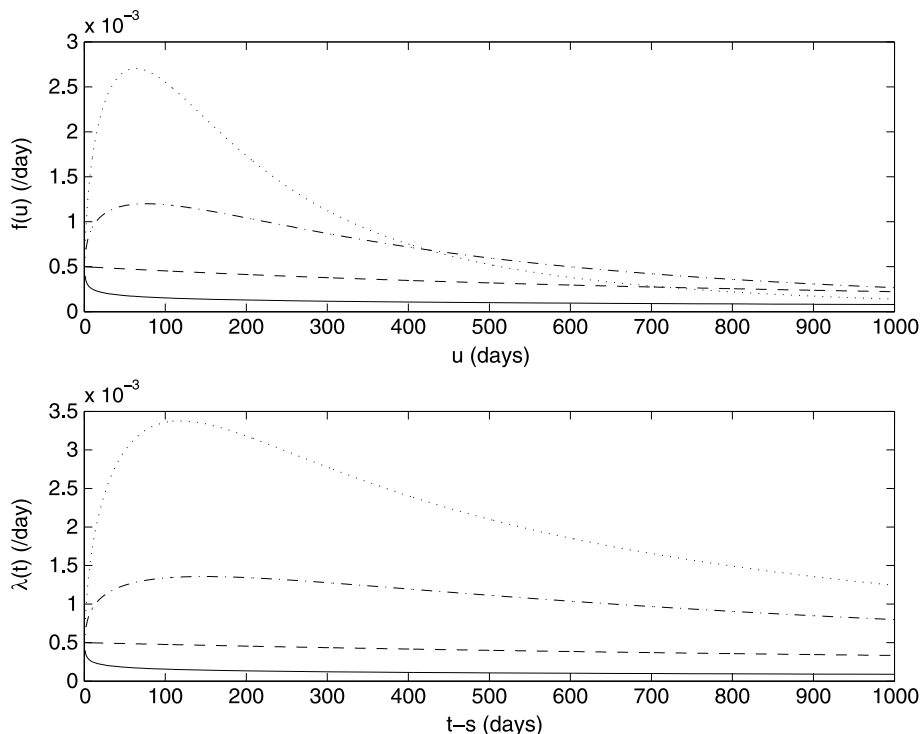
rameters can be estimated by numerically fitting this equation to the empirical renewal intensity.

Wickman [105] extended this to six possibilities modeled on various well-known volcano types, and simulated them to produce reasonable appearing synthetic data. The process of fitting these models to observed data has required manual identification of the various states in the data [11,54], but the use of hidden Markov model techniques to automate the process has been suggested [3]. In a Markov process, the time between onsets is the passage time through the eruptive and repose states, which will have a compound exponential distribution, i.e., be the sum of a number of exponential random variables of differing means.

### Time and Size Predictable Models

Bacon [1] noted that the intervals between rhyolitic and basaltic eruptions in the Coso Range (California) appeared to be proportional to the volume of the preceding eruption. This *time-predictable* behavior (cf. [81]) is most simply explained by assuming that the rate of magma input is constant, and an eruption occurs when a certain magma level is reached. However, Bacon [1] suggested the covariate of accumulated extensional strain as the controlling





Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 7  
 Density  $f$  and conditional intensity  $\lambda$  for the log-logistic distribution with  $\eta = 0.0005$  and  $\gamma = 0.8$  (solid line),  $\gamma = 1$  (dashed line),  $\gamma = 1.2$  (dot-dash line) and  $\gamma = 1.4$  (dashed line)

variable, and considered eruptions a passive response of the magmatic system to the tectonic stress field.

Similar behavior was observed for Fuego Volcano, Guatemala [54], together with a weaker proportionality between volume and the length of repose before an eruption, although the authors did note that “the volume estimates probably have an error of less than  $\pm 1$  order of magnitude”. Martin and Rose [54] also examined a possible link between the length of repose before an eruption and its geochemical composition. Santacroce [75] found a similar phenomenon at the Somma-Vesuvius complex, and suggested that eruptions there occur in cycles, of approximately 3000 years, each initiated by a large-scale Plinian eruption, and terminated by a long repose of several centuries. Klein [46] found evidence of time-predictable behavior at both Kilauea and Mauna Loa (Hawaii).

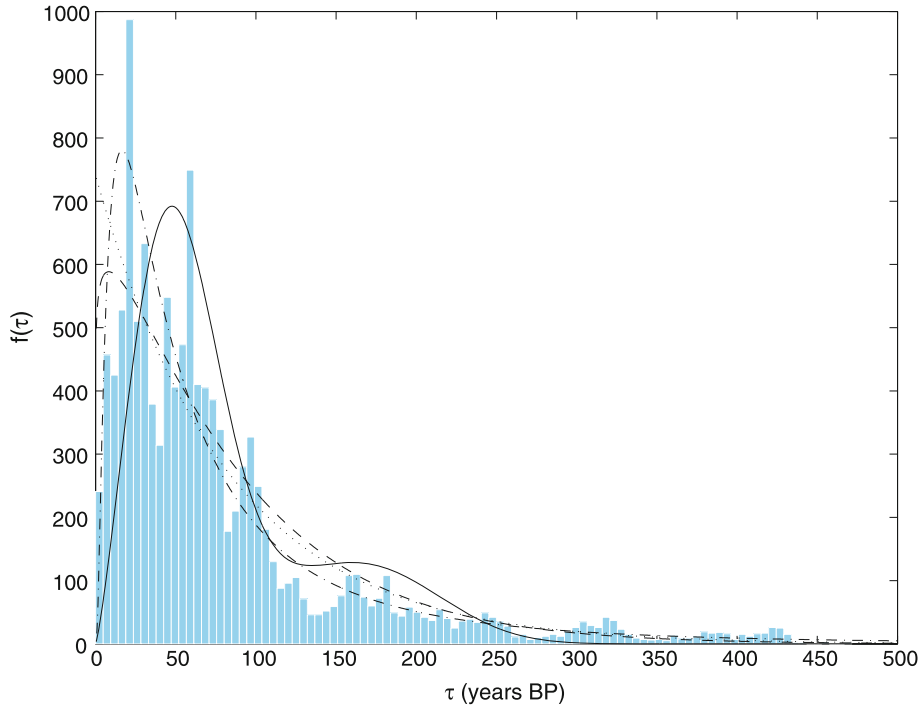
Wadge and Guest [101] suggested that the time-predictable model has a limited predictive ability, based on drawing an empirical envelope on the recent eruptive volume curve. This has a slope determined as the mean eruptive volume rate, and limits defined by the extreme points of the curve in the considered history. This enables a forecast of both the longest the current repose is likely to

last, and the maximum volume of the next eruption, provided that steady-state is maintained. Figure 9 shows an example using flank eruptions from Mt Etna 1970–2006. Wadge [99] estimated, however, that volcanoes may spend only 1/4–1/2 of the time in steady-state, being dormant much of the remainder. Changes in eruptive rate have also been found in a number of historical records [9,26,65].

De la Cruz-Reyna [26] proposed a general *load-and-discharge* model, where the ‘energy’ of the volcano increases at a constant rate  $\sigma$  between eruptions. The  $i$ th eruption occurs when this exceeds the threshold  $H_i$ , during which the stored energy drops to the threshold  $L_i$ , releasing an energy  $\xi_i = H_i - L_i$ . The interval between eruptions is thus

$$r_i = t_{i+1} - t_i = \frac{H_{i+1} - H_i + \xi_i}{\sigma} . \tag{27}$$

Hence successive repose intervals are not independent, as they include a common threshold term, and should be negatively correlated. In general, the thresholds can be random, but this formulation includes the time-predictable model. In this, the thresholds  $H_i$  are constant, and



Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 8  
 Histogram of 122,000 sampled inter-event times for Mt Taranaki based on 1000 Monte Carlo runs. Curves show the fitted densities for this data set: Dotted line = exponential distribution, dot-dash = lognormal, dashed = Weibull, solid = mixture of Weibulls

thus the interval

$$r_i = \xi_i / \sigma \tag{28}$$

is proportional to the energy of the preceding eruption, and successive intervals are independent. Alternatively, if the thresholds  $L_i$  are held constant ( $= L$ , say),  $H_i = \xi_i - L$  and from (27)

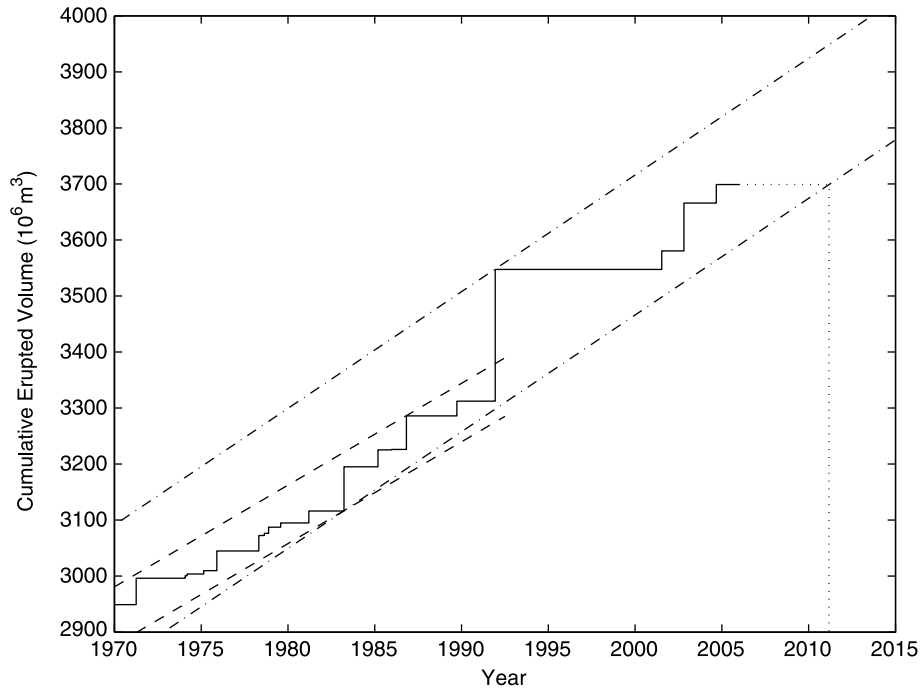
$$r_i = (H_{i+1} - L) / \sigma . \tag{29}$$

Thus successive intervals are again uncorrelated, and the energy drop of the eruption is proportional to the length of the interval preceding it. This is termed the *size-predictable* model.

In a global analysis of eruptions, De la Cruz-Reyna [26] noted the existence of a regression relationship  $\log \lambda_0 = 3.494 - 0.789\text{VEI}$  with correlation  $-0.999$ , where  $\lambda_0$  is the Poisson intensity for eruptions of a given VEI. This led De la Cruz-Reyna [27] (see also [28]) to model the onsets of different VEI events as independent Poisson processes, in a special case of the size-predictable model. This approach was also used in the quantification of volcanic risk due to Vesuvius by Scandone et al. [76], based on the sequence of

equilibrium states (repose, persistent activity, intermediate eruption, final eruption) identified by Carta et al. [11].

Burt et al. [9], in an examination of the basaltic volcano Nyamuragia (Zaire), formulated the ‘pressure-cooker’ and ‘water-butt’ models. The former, where an eruption occurs when the magma volume reaches a threshold determined by the rock strength and reservoir shape, is the time-predictable model, while the latter, where eruptions always drain the magma reservoir to the same level, is the size-predictable model. Burt et al. [9] suggested physical explanations for stochastic perturbations observed in these models, and proposed that a volcano could be tested for time-predictability by performing a regression analysis of  $(v_i^{(1)})$  on  $(r_i)$ , where  $r_i = t_{i+1} - t_i$ , and  $t_i, v_i^{(1)}$  are the onset time and eruptive volume of the  $i$ th eruption. Similarly the size-predictable model entails a regression analysis of  $(v_i^{(2)})$  on  $(r_i)$ , where  $v_i^{(2)} = v_{i+1}^{(1)}$ . Sandri et al. [74] generalized the test for time-predictability to a regression analysis of  $(\log v_i^{(1)})$  on  $(\log r_i)$ , so that an estimated slope of  $b$  significantly greater than zero implies a time-predictable relation  $v_i^{(1)} \propto r_i^b$ . Marzocchi and Zaccarelli [59] performed a similar regression analysis of  $(\log v_i^{(2)})$  on  $(\log r_i)$  for the size predictable model, resulting in a relation  $v_i^{(2)} \propto r_i^b$ . Be-



Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 9  
 Volume-time curve for Mt Etna flank eruptions (solid line). Dashed limits based on 1970–1990 data were exceeded, or equivalently the volcano departed from steady-state, by the 1991 eruption. Dot-dash limits based on 1970–2006 data have a steeper slope and wider limits to accommodate the 1991 eruption. The dotted lines show that the method predicts the next eruption to occur by the end of February 2011, and the eruptive volume to be less than  $142.1 + 20.8t$  ( $10^6 \text{ m}^3$ ), where  $t$  is the number of years to the date of the eruption from January 2007

cause both the repose and volume distributions are highly skewed, the logarithmic transformation is advisable to reduce the high leverage of the tail points.

Marzocchi et al. [57] estimated the VEI of the next eruption of Mt Vesuvius by considering a) the global catalog, b) a catalog of ‘analog’ volcanoes, and c) the catalog of Vesuvius itself. In each case, the catalogs are trimmed to those eruptions following a repose interval of at least the presently observed 60 years. The estimated probability of an eruption of  $\text{VEI} \geq 5$  ranged from 1% to 20%, and could not be considered negligible. This has implications for the Emergency Plan of Mount Vesuvius, which assumes a maximal  $\text{VEI} = 4$ .

Marzocchi and Zaccarelli [59] proposed a renewal model based on the time-predictable model. In their *open conduit system* model, the probability density of the time  $u$  to the next eruption is conditional on the previous eruptive volume  $v$ . Using a log-normal distribution, this is

$$f(u | v) = \frac{1}{u\sigma\sqrt{2\pi}} \exp \left[ -\frac{(\log u - (a + bv))^2}{2\sigma^2} \right], \quad (30)$$

where  $a, b$  are the intercept and slope regression coefficients from the regression of  $(\log v_i^{(1)})$  on  $(\log r_i)$ .

### Chaos and Fractals

A sufficiently complicated deterministic model is indistinguishable by finite observation from a random process. However, it is possible to construct systems with few degrees of freedom, but very complex dynamics, which are known as *chaotic systems*, or *strange attractors*. This possibility can be tested by examining the correlation properties of the observation sequence. The underlying principle of fractals is one of *self-similarity*, i.e., that the statistical properties of the process are invariant at differing time (or space) scales.

Sornette et al. [88] computed the dimension  $D$  of the attractor as a function of the dimension  $d$  of the embedding phase space. Given onset times  $t_i$ , let  $X_i^{(k)} = r_{i+k-1}$  for  $k = 1, \dots, d - 1$ , where  $r_i = t_{i+1} - t_i$ , for  $i = 1, \dots, n - 1$ . Then the correlation function is determined as

$$C_d(w) = \frac{1}{n^2} \#\{i, j: \|\mathbf{X}_i - \mathbf{X}_j\| < w\}, \quad (31)$$

where  $\mathbf{X}_i$  is the vector  $(X_i^{(1)} \dots X_i^{(d)})$ . The correlation dimension is then defined by the power law relation

$$C(w) \sim w^{D(d)}. \quad (32)$$

If this saturates at some value  $D_{\max}$  as  $d$  increases, this is taken as evidence that the underlying dynamics is deterministic and has dimension of the order of  $D_{\max}$ . Applications to the volcanoes Piton de la Fournaise (La Reunion Island) and Mauna Loa/Kilauea (Hawaii) produced estimates of  $D_{\max}$  of 1.7 and 4.6, respectively. The latter value was interpreted as evidence of two independent dynamical systems for the two Hawaiian volcanoes, contradicting the conclusion of Klein [46]. Sornette et al. [88] also constructed a series of ‘return maps’ ( $r_{i+1}$  as a function of  $r_i$ ) by truncating the data in various ways.

Dubois and Cheminee [29] used the Cantor dust model for volcanic eruption onsets. In this method, the eruptive record is divided into segments of length  $T$ , and the fraction of these intervals containing at least one onset is denoted  $X$ . If fractal clustering with fractal dimension  $D$  is present, then  $X \sim T^{1-D}$ . This implies a renewal model with intensity  $\lambda(t) = D/(t - s)$ ,  $t > s$ , where the last eruption occurred at time  $s$  (cf. Eq. (24)). An obvious method for estimating  $D$  is a regression analysis of  $(\log X_i)$  against  $(\log T_i)$  for various interval lengths  $T_i$ . Applied to the Piton de la Fournaise (La Reunion Island), two slopes were observed, one with  $D \approx 0.5$  for  $T < 10$  months, and one with  $D \approx 0.7$  for  $T > 10$  months. Kilauea had two regimes with  $D = 0.58$  for  $T < 24$  months and  $D = 0.81$  for  $T > 24$  months, while Mauna Loa had  $D \approx 0.35$  for  $T < 6$  years and  $D \approx 0.85$  for  $T > 6$  years. The two regimes all indicated clustering at short repose lengths, and more regular activity for eruptions separated by longer repose. This was interpreted as evidence of multiple magma chambers, with different sizes and recharge rates, at different depths. Applied to a combined catalog of flank and summit eruptions of Mt Etna 1950–1987, the result was a single regime with  $D = 0.88$ . This high value of  $D$  was interpreted as being characteristic of the regular eruptive activity of the volcano.

Marzocchi [55] calculated the statistic

$$a_{\max} = \max_{1 \leq s \leq 5} a(s), \tag{33}$$

where  $a(s)$  is defined in (3), although a process might first be de-trended by subtracting an auto-regressive model. A value  $a_{\max} > a_c^*$  indicates a chaotic system, while a value  $a_{\max} < a_s^*$  indicates a stochastic system. Intermediate values provide insufficient evidence for either. The critical values  $a_c^*$  and  $a_s^*$  are determined by simulation. Both Vesuvius and Etna evidenced significantly stochastic behavior.

A multifractal analysis using the Grassberger-Procaccia algorithm was performed by Godano and Ci-

vetta [33]. The correlation interval is calculated as

$$C(t, q) = \frac{1}{n-1} \sum_{i=1}^{n-1} \left[ \frac{1}{n-2} \sum_{j \neq i} H(t - |r_i - r_j|) \right]^q, \tag{34}$$

where  $r_1, \dots, r_{n-1}$  are the inter-onset times, and  $H(\cdot)$  is the Heaviside function. The *generalized dimensions* are then evaluated as

$$D_{q+1} = \phi(q)/q, \tag{35}$$

where  $\phi(q)$  is the estimated slope from a regression analysis of  $\log C(t, q)$  versus  $\log t$ . The fractal, information, and correlation dimensions are  $D_0, D_1$  and  $D_2$ , respectively. More generally,  $D_{q+1}$  is constant as  $q$  varies for a homogeneous fractal distribution, but varies with  $q$  for a multifractal distribution. The latter implies that local clustering properties are different from global clustering properties. Applied to eruptions of Vesuvius, the estimated dimensions were  $D_0 = 1.05 \pm 0.07$ ,  $D_2 = 0.75 \pm 0.04 \approx D_\infty = 0.73 \pm 0.02$  and  $D_{-\infty} = 1.09 \pm 0.13$ . This indicates that at a global scale, the catalog exhibits Poissonian behavior ( $D = 1$ ), and the most intense clustering is not a particularly strong effect. The catalog exhibits only weak multifractal behavior, and this only for  $q < 1$ . Thus the cyclic behavior outlined by Carta et al. [11] cannot be identified statistically, as the intervals between events in a cycle are not distinguishable from those between cycles.

Nishi et al. [69] examined the long term memory of the time series of 3796 eruptions of Sakurajima 1981–1999 using the Hurst exponent. Let  $r_i = t_{i+1} - t_i$ ,  $i = 1, \dots, n-1$ , be the inter-onset times, and set

$$X(k, t, m) = \sum_{i=1}^k \left( r_{i+t-1} - \frac{1}{m} \sum_{j=1}^m r_{j+t-1} \right). \tag{36}$$

Then the *self-adjust range* is

$$R(t, m) = \max_{1 \leq k \leq t} X(k, t, m) - \min_{1 \leq k \leq t} X(k, t, m), \tag{37}$$

which is re-scaled by the standard deviation

$$S(t, m) = \sqrt{\frac{1}{t} \sum_{i=1}^t \left( r_{i+m-1} - \frac{1}{m} \sum_{j=1}^m r_{j+t-1} \right)^2}, \tag{38}$$

and averaging over the length of the time series, the *Hurst exponent*  $H$  is then defined by

$$\frac{\sum_{m=1}^{n-1} R(t, m)}{\sum_{m=1}^{n-1} S(t, m)} = \left( \frac{t}{2} \right)^H. \tag{39}$$

The value of  $H$  can be estimated by a regression analysis of (39), after taking logarithms. The Poisson process produces a value of  $H = 0.5$ . A value of  $H$  greater than this implies persistent behavior, with each value depending on the previous one. Anti-persistence behavior is characterized by  $H < 0.5$ . The Hurst exponents for Sakurajima were  $H = 0.72$  for the whole series, and  $H = 0.74$  for a ‘high frequency’ period 1983–1985. A Monte Carlo test consisting of randomly reordering the observed inter-onset times and recalculating  $H$  confirms that these values are significantly greater than 0.5. Telesca and Lapenna [92] performed the same analysis for eruptions of  $VEI \geq 0$  of 14 volcanoes worldwide. The requirements were a minimum of 40 events during 1800–2000. Seven of the volcanoes, with Hurst exponents ranging from 0.79 to 1.3, were characterized by persistent behavior, while the remainder, with Hurst exponents ranging from 0.60 to 0.73, were not significantly different from a Poisson process.

Telesca et al. [91] examined 35 sequences of  $VEI \geq 0$  eruptions (minimum 25 eruptions 1800–2000) using the Fano factor method. The Fano factor is defined as the variance of the number of events in a specified interval width  $\Delta$  divided by the mean,

$$FF(\Delta) = \frac{(\Delta/T) \sum_{i=1}^{T/\Delta} N_i^2 - \left( (\Delta/T) \sum_{i=1}^{T/\Delta} N_i \right)^2}{(\Delta/T) \sum_{i=1}^{T/\Delta} N_i}, \quad (40)$$

where  $N_i = N((i - 1)\Delta, i\Delta)$ . Telesca et al. [91] represented the FF of a fractal point process as a monotonic power law

$$FF(\Delta) = 1 + \left( \frac{\Delta}{\Delta_0} \right)^\alpha, \quad (41)$$

for  $\Delta > \Delta_0$ , where  $\Delta_0$  is the fractal onset time, marking the lower limit for significant scaling behavior, with negligible clustering below this. Again the fractal exponent  $\alpha$  is estimated by a regression analysis of (41), after taking logarithms. For Poisson processes, the FF is approximately one for all interval lengths, and so  $\alpha \approx 0$ . Of the 35 sequences examined, 30 had fractal exponents ranging from 0.2 to 0.9 (mean of 0.5). The remaining five exhibited no value  $\Delta_0$  above which (41) held.

Gusev et al. [34] examined clustering of eruptions on Kamchatka during the last 10,000 years using the Weibull density (17) for the inter-onset times, the correlation dimension (32), and the Hurst exponent (39). No self-similar behavior was found for intervals greater than 800 years, but for shorter delays there was a significant degree of clustering. Gusev et al. [34] also generalized the correla-

tion dimension by replacing (31) by a weighted version, where the contribution of a pair of events is proportional to the product of their eruptive volumes. This identified self-similar behavior over the entire range of inter-onset times (100–10,000 years).

Fractals have basically been used in volcanology as a descriptive technique. They can provide, at the level of a single volcano, some indication of the sort of stochastic model that might fit the observed behavior. Applied to multiple volcanoes, they can provide a similarity index, useful for sorting volcanoes into homogeneous groups.

### Volcanic Regimes

Apart from the nonhomogeneous model (4), the parametric models examined above have all been stationary, in the sense that the statistical properties of the intensity function do not vary over time. However, Wadge [99] postulated that, while a volcano could be in steady state on a time scale of years to decades, in the longer term the activity can wax and wane. Thus, particularly for long-term hazard forecasts, it is desirable to formulate methods of detecting (and modeling) such non-stationary activity.

Wickman [103] suggested that the activity of a volcano could change over time, perhaps in discrete steps, which we will term *regimes*. Mulargia et al. [65] addressed the importance of objectively identifying regimes of a volcano, noting however that a) the number of regimes is unknown, b) they may follow different distributions, and c) sample sizes are generally small. Different patterns of eruption and magma output are supposedly features of most volcanoes, and fundamental to the understanding and modeling of eruptions. These regimes may represent changes in the eruption mechanism, the mechanism for transport of magma to the surface, or the eruptive style (see, for example, the interpretation by Wadge et al. [102]). The estimated properties of such regimes can exclude certain models or mechanisms of volcanic eruption [46].

Various methods have been used to identify changes in regime, one based on cumulative eruptive volume [99,102] already having been illustrated in Fig. 9. Another uses a running mean of time between onsets. Suppose that the individual reposees are denoted by  $r_i = t_{i+1} - t_i$ ,  $i = 1, \dots, n - 1$ , and that  $\sum_{i=1}^{n-1} r_i = T$ . If the reposees are divided into disjoint groups of  $m$  consecutive reposees, then  $(2(n - 1)/T) \sum_{i=(k-1)m+1}^{km} r_i$ , for  $k = 1, \dots, \lfloor (n - 1)/m \rfloor$ , where  $\lfloor \cdot \rfloor$  is the greatest integer function, are independent and have a chi-square distribution with  $2m$  degrees of freedom [46].

Mulargia et al. [65] presented a method based on the theory of change-point problems. Suppose that we have

a time series of data  $X_1, \dots, X_n$ , which might be, for example, inter-onset intervals, eruptive volumes, or effusion rates. The algorithm begins by setting a significance level  $\alpha$ , and examining the time series to determine the most significant change point. If a change point is found, the time series is divided into two parts at this point, and each examined to determine the most significant change point. The algorithm proceeds recursively until no further significant change points are found. The most significant change point is identified by a search process: For  $m = 3, \dots, n-3$ , divide the time series into two segments  $X^{(1)} = X_1, \dots, X_m$ , and  $X^{(2)} = X_{m+1}, \dots, X_n$ , and calculate the Kolmogorov–Smirnov (two-tail) statistic as

$$J = \left( \frac{m(n-m)}{n} \right)^{1/2} \max_x |F_1(x; m) - F_2(x; n-m)|, \quad (42)$$

where  $F_j(x; k) = \#\{i: X_i^{(j)} \leq x\}/k$  is the empirical distribution function of the first segment, et cetera. For large  $m, n-m (> 30)$ , the approximate critical value  $J_\alpha$  of the statistic can be determined from

$$\Pr(J < J_\alpha) = \sum_{k=-\infty}^{\infty} (-1)^k \exp(-2k^2 J_\alpha^2) = 1 - \alpha. \quad (43)$$

Exact critical values for small  $m, n-m$  are also available.

Later approaches used the cumulative count of eruptions in a statistical control chart [39], applied by Burt et al. [9] to Nyamuragia, and rank order statistics for the size of event [71]. Bebbington [3] noted that a hidden Markov model for the activity, with the unobserved state representing the regime, provides for regime identification via the Viterbi algorithm, which finds the most likely path through the hidden states.

## Mount Etna

Mount Etna possesses one of the most complete records of volcanism known. A variety of eruption styles have been displayed, but since 1500 the activity has been predominantly effusive eruptions from lateral vents on the volcano's flanks, or from the central craters. The record of summit eruptions is incomplete prior to 1970 [101], but data for flank eruptions is considered complete since 1600 or so [65]. The volumes of all such eruptions have been estimated [74], and these have been shown to be well correlated with the observed duration [67]. As the summit eruptions appear to be of different style [102], and to have different mechanisms to the flank eruptions [66] the longer record of the latter can be analyzed in isolation. In addition,

only flank eruptions pose a major threat to inhabited areas [73].

Using the cumulative erupted volume curve, Wadge and Guest [101] found that the volume output was in steady state 1971–1981. Mulargia et al. [67] found that the events 1600 to 1980 were satisfactorily fit by a stationary Poisson process, but that the distribution of eruptive durations changed at the end of the 17th century, confirming the observation of Wadge et al. [102]. However, Mulargia et al. [65] identified a change point in the inter-onset times at 1865, both parts being consistent with a Poisson process, and, when augmented by volume data, change points at 1670 and 1750. Using the effusion rate (= volume/duration), Mulargia et al. [65] found a change point at 1950 in the 1600–1980 data, with the later sequence being of significantly lower rate. A repeat of the analysis by Gasperini et al. [31] for the data 1978–1987 identified a change point in the inter-onset times at November 1978, although no change point occurred in the volume series. Ho [39] fitted a statistical control chart based on the Weibull process, and found departures from the trend (4) in 1702 and 1759. Marzocchi [55] found some evidence of a trend in inter-onset times in the 1600–1994 data, while Sandri et al. [74] found no change points in the inter-onset times or eruptive volumes in the 1971–2002 data. Salvi et al. [73] determined that the activity 1536–2001 was not Poissonian, and although there was no trend 1536–1980, there was a trend in the data 1536–2001. They therefore concluded that the last 20 years of data is from a nonhomogeneous Poisson process, and fitted a Weibull process model with increasing trend.

The regime change points identified in the Mount Etna flank eruption data are summarized in Table 1. The lack of consistency is due to differences in both models and, via the fitting of them, in the coverage of the data. These models are 'first order' in the data being analyzed. Using hidden Markov models, Bebbington [3] was able to examine the 'second order' quantity of the correlation between durations and subsequent repose. The conclusion was that changes in volcanic regime may be more frequent and/or fleeting, and thus undetectable by change-point methods, than has been thought. Notably, the identified change points corresponded closely to those identified manually by Behncke and Neri [7], showing that these are statistically identifiable. The complex relation between the duration of an eruption and the subsequent repose appears to be at least partially explained by an open/closed conduit system model with transitions between them. On the other hand, the correlation between repose and subsequent durations may be sensitive to the longer of the two cycles identified by Behncke and Neri [7]. This would indi-

Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Table 1  
 Regime change points for Etna flank eruptions

Data type	Source	Date Range	Change Points
Onset counts	Mulargia et al. [67]	1600–1980	None
Inter-onset times	Mulargia et al. [65]	1600–1978	1865
	Gasperini et al. [31]	1978–1987	1987
	Ho [39]	1600–1978	1702, 1759
	Marzocchi [55]	1600–1994	(Trend)
	Sandri et al. [74]	1971–2002	None
	Salvi et al. [73]	1536–2001	1980
Volume -repose	Wadge et al. [102]	1535–1974	1610, 1669, 1759/1763
	Wadge and Guest [101]	1971–1981	None
	Mulargia et al. [65]	1600–1978	1670, 1750, 1950
	Gasperini et al. [31]	1978–1987	None
	Sandri et al. [74]	1971–2002	None

cate that flank eruptions are being fed from more than one magma chamber, of widely different sizes and recharge rates.

**Spatial Aspects**

In the case of polygenetic cones, the major aspect of hazard modeling is temporal. Spatial aspects are limited to the direction of flank eruptions (cf. [73]), and the likely paths of lava flows and lahars. In monogenetic volcanic fields, where cones correspond to single eruptions, there is a true spatio-temporal element to modeling occurrence data.

**Spatio-temporal Intensities**

The models used to describe spatio-temporal eruption occurrence have been largely nonparametric, or kernel-type. Connor and Hill [17] gave three methods:

1. Spatio-temporal nearest-neighbor estimate: Suppose we have  $n$  volcanoes, that the formation of the  $i$ th volcano occurred at time  $t_i$ , and let  $u_i$  be the area of a circle with radius equal to the distance between the point  $x$  of interest and the  $i$ th volcano. Further, let  $j = 1, \dots, m$  index the  $m$ th nearest neighbors to the point  $x$  using the distance metric  $u_i(t - t_i)$ . Then the estimated intensity at the point  $x$  is

$$\lambda(x, t; m) = \frac{m}{\sum_{j=1}^m u_j(t - t_j)}, \tag{44}$$

where  $m$  is the number of nearest neighbors used in the estimation, which was set equal to the number of volcanoes, although a smaller choice is possible. The temporal hazard for an area  $A$  can then be obtained by integrating the intensity (44) over the area. Note that in

practice, to avoid singularities,  $u_i > c$ , a constant, and the intensity is summed over a grid. The choice of  $c$  may be an interesting question.

Condit and Connor [14] proposed that the optimal  $m$  can be determined by matching the average recurrence rate  $\int_{x \in A} \lambda(x, t; m) dx$  to the observed average recurrence rate  $\bar{\lambda}(t) = N(t-s, t)/As$ , in a window of length  $s$ , for the area  $A$ . Further, Condit and Connor [14] also suggested that overestimation during rapidly waning stages of activity can be addressed by defining a threshold time beyond which prior eruptions are not included in the calculation.

2. Kernel estimate: If we have  $n$  volcanoes, then using the Epanechnikov kernel

$$\kappa_i = \max \left\{ 0, \frac{2}{\pi} \left[ 1 - \left( \frac{d_i}{h} \right)^2 \right] \right\}, \tag{45}$$

where  $d_i$  is the distance from  $x$  to the  $i$ th volcano, we have the spatial intensity

$$\lambda(x) = \frac{1}{e_h} \sum_{i=1}^n \frac{\kappa_i}{h^2}, \tag{46}$$

where  $e_h$  is an edge correction. Conway et al. [20] use the Gaussian kernel instead. The parameter  $h$  is a smoothing constant. A small value of  $h$  concentrates the probability close to existing volcanoes, while a large value distributes it more uniformly. Estimating the best value of  $h$  is a difficult task – see Vere-Jones [96] for an earth sciences based discussion.

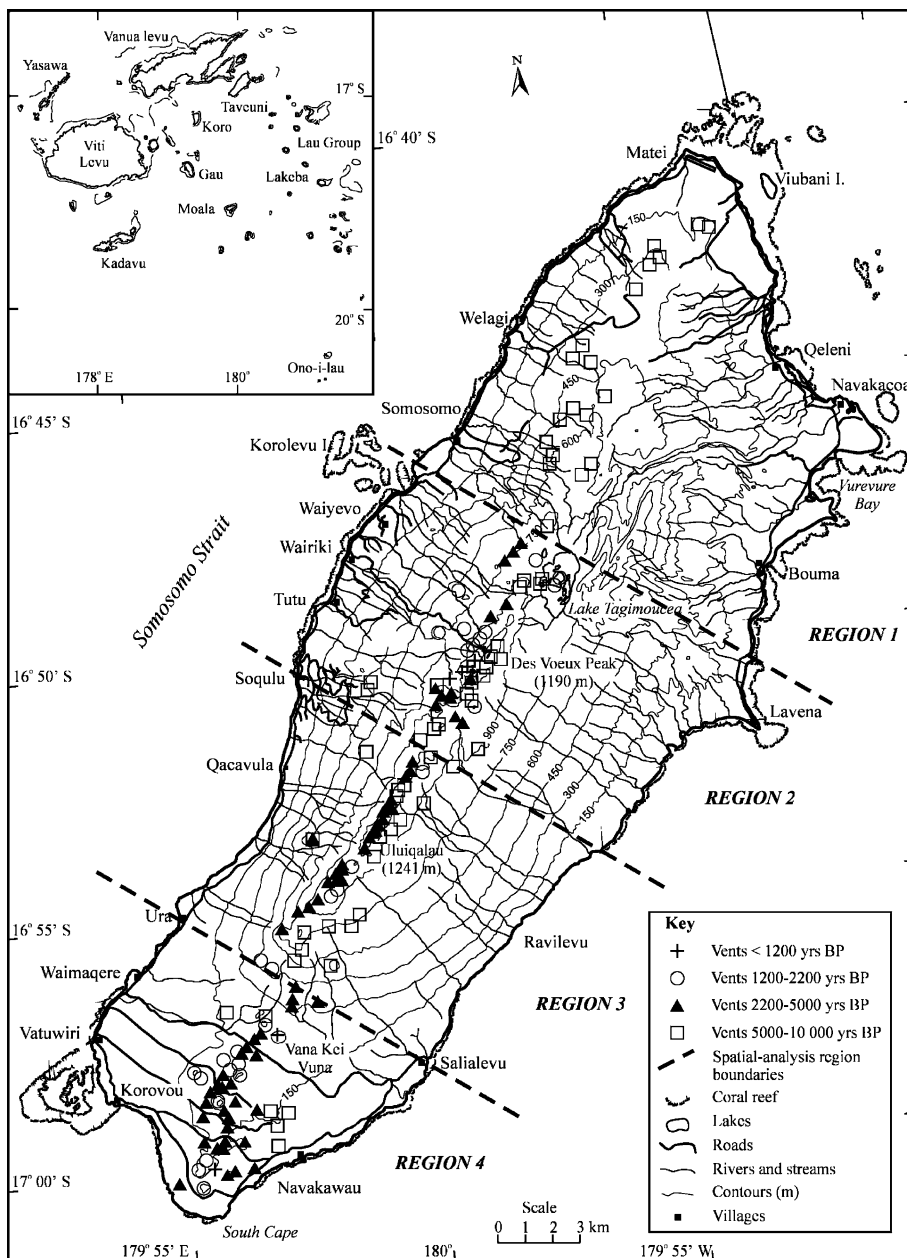
3. Nearest-neighbor kernel estimate: Here we replace the smoothing constant  $h$  in the Kernel estimate with  $d_{(m)}$ , the distance to the  $m$ th nearest neighbor using the  $d_i$  metric.

The two (spatial) kernel estimates above can then be made into spatio-temporal estimates by multiplying them by a spatially independent temporal rate  $\hat{\lambda}_0 = N/T$ , resulting in

$$\lambda(x, t) = \lambda(x)\hat{\lambda}_0 . \tag{47}$$

Ho and Smith [42] generalized (47) by replacing  $\hat{\lambda}_0$  by the Weibull process intensity (4).

Martin et al. [53] extended the spatio-temporal idea with a Bayesian formulation to incorporate information from other geophysical data such as P-wave velocity perturbations or geothermal gradients. The spatio-temporal



Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 10  
 Location map of Taveuni (inset: Fiji group), with main villages, positions of Holocene mapped vents, and regions used for the spatial analysis. From [22], Figure 1, © Springer-Verlag 2001. With kind permission of Springer Science and Business Media



intensity (46) is used as an a priori intensity  $P(x)$ , and a likelihood function  $L(\theta | x)$  is generated by conditioning the geophysical data  $\theta$  on the locations of volcanic events. The a posterior intensity can then be obtained from Bayes' theorem as

$$P(x | \theta) = \frac{P(x)L(\theta | x)}{\int_{y \in A} P(y)L(\theta | y)dA}, \tag{48}$$

where  $A$  is the volcanic field.

**Markov Chains**

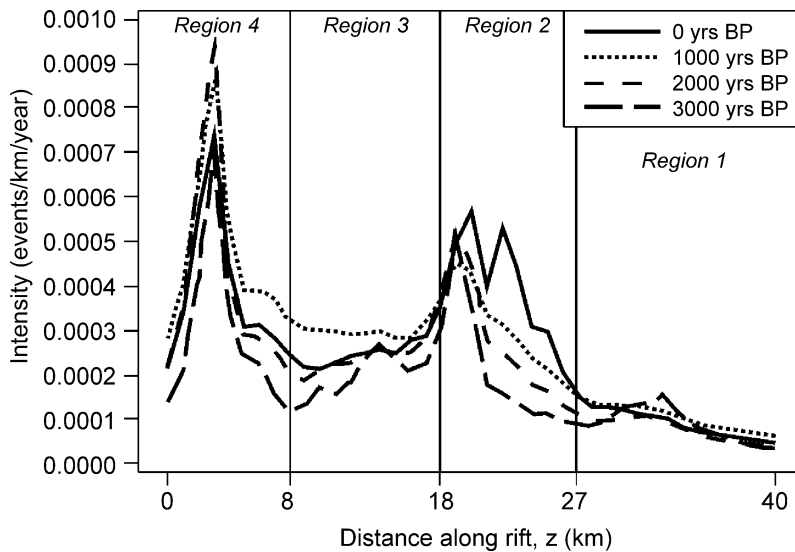
In the case where there is a geologically useful spatial classification, and sufficient data, a simple Markov chain approach can be used. If the  $i$ th eruption occurs in spatial region  $X_i$ , then the transition matrix is  $P = (p_{jk})$ , where  $p_{jk} = \text{Pr}(X_{i+1} = k | X_i = j)$ . The maximum likelihood estimates are  $\hat{p}_{jk} = N_{jk} / \sum_l N_{jl}$ , where  $N_{jk}$  is the number of transitions from  $j$  to  $k$  observed in the data. The Markov chain can be tested against a null hypothesis of independence using the statistic

$$\chi^2 = 2 \sum_j \sum_k N_{jk} \log \left[ \frac{N_{jk} / \sum_l N_{jl}}{\sum_l N_{lk} / \sum_l \sum_m N_{lm}} \right], \tag{49}$$

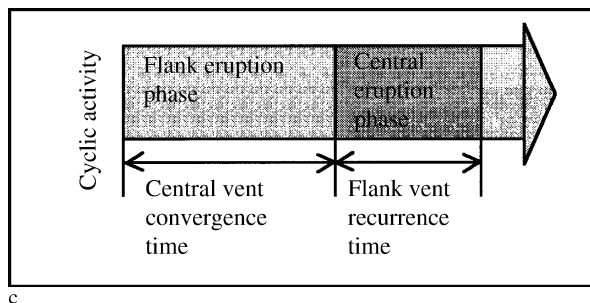
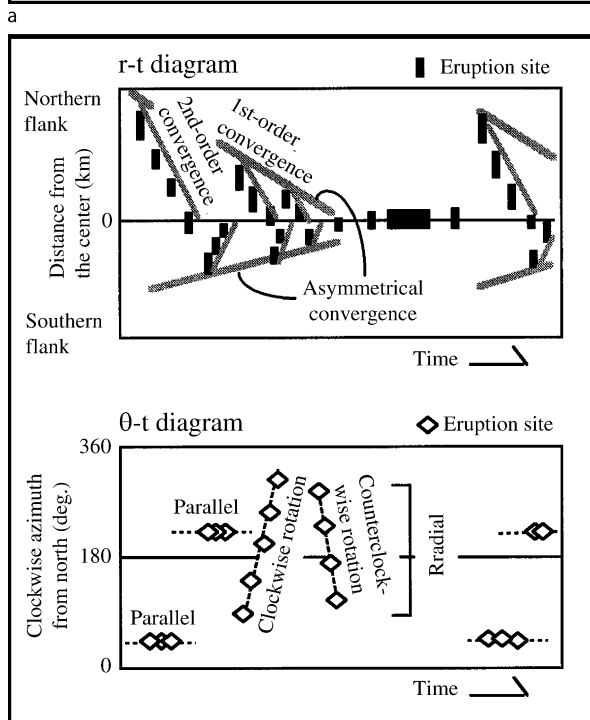
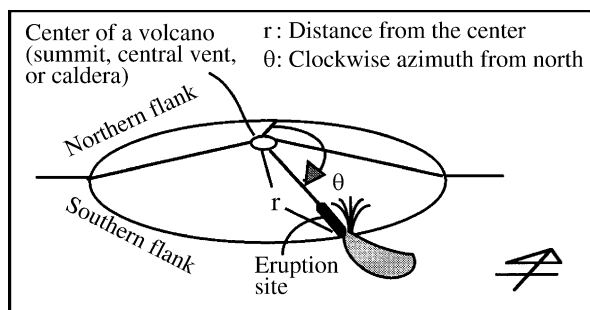
which asymptotically has a  $\chi^2$  distribution with  $K^2$  degrees of freedom, where  $K$  is the number of states. Cronin et al. [22] used this approach, made feasible by the essentially linear rift formation shown in Fig. 10, to estimate the hazard at the heavily populated south-west end of Taveuni,

Fiji. This was combined (cf. Eq. (47)) with a hierarchical renewal model for the temporal aspect to provide an alternative to the spatio-temporal hazard estimate from (44) shown in Fig. 11. Eliasson et al. [30] likewise divided the Katla caldera, Iceland, into three sectors to examine the spatial progression of volcanogenic flood events.

In the same vein, a few studies have considered possible correlations between multiple vents on the same volcano. Klein [46] showed using a runs test that eruptions and intrusions at Kilauea tend to cluster in time, and that intrusions occur in place of eruptions during long repose. By dividing Kilauea repose into summit-summit, summit-flank, flank-summit and flank-flank pairs, and finding that their distributions did not differ statistically, Klein [46] showed that the location of the previous event provided no information about the time of the next eruption. At Mauna Loa, however, repose following flank eruptions were significantly longer than those following summit eruptions, consistent with the fact that the former are generally more voluminous than the latter. Examination of the sequences of summit and flank eruptions disclosed that Kilauea summit eruptions tend to occur in runs, due solely to the long summit sequence of 1924–1954, while Mauna Loa displays no tendency for clustering or alternation of summit and flank eruptions. However, Lockwood [49], in an analysis of 170 well-dated prehistoric lava flows from Mauna Loa, identifies cycling between summit overflows and flank eruptions with a periodicity of about 2000 years.



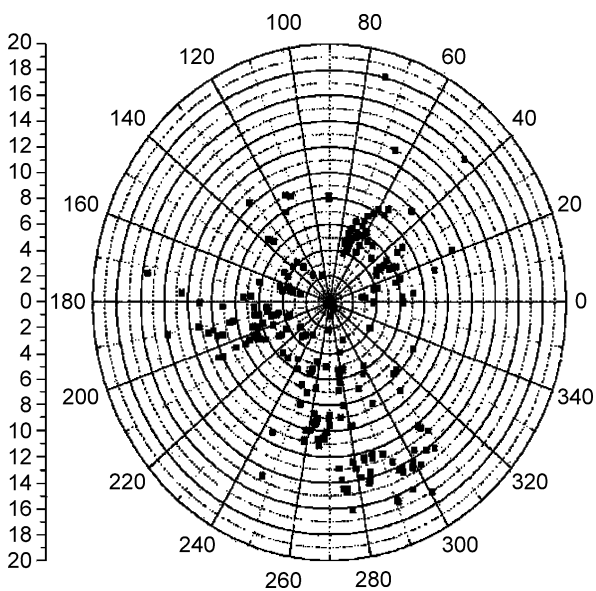
Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 11  
 Plot of spatio-temporal intensities (events/km year<sup>-1</sup>) over the length of the Taveuni rift zone and at different times between 3000 B.P. and the present. From [22], Figure 8, © Springer-Verlag 2001. With kind permission of Springer Science and Business Media



Takada [90] proposed using time-series plots, as shown in Fig. 12, to visually express the temporal relationship between flank and summit eruptions, in particular plots of flank-vent distance and direction from the central vent. Central eruption phases, defined as periods where at least 70% of eruptions are central vent eruptions without accompanying flank eruptions, alternate with flank erup-

◀ Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 12

**a** Parameters representing the position of an eruption site. The flank of a volcano is divided into two regions, e.g., the northern flank and the southern one. **b** Two types of time-series diagram for eruption sites, representing the distance of eruption sites from the center ( $r$ - $t$  diagram) and the clockwise azimuth from the north ( $\theta$ - $t$  diagram), respectively. **c** Schematic cyclic activity consisting of flank eruption and central eruption phases From [90], Figure 1, © Springer-Verlag 1997. With kind permission of Springer Science and Business Media



Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 13

**a** Azimuthal distribution of cones of lateral eruption on Etna. The center is located on "La Voragine" central crater. The scale on the left is in km from the center. Reprinted from [73], © 2006 Elsevier B.V., with permission from Elsevier

tion phases. Salvi et al. [73] considered the azimuth distribution of flank eruptions from Mount Etna, as in Fig. 13, determining that there is no significant difference in the pre- and post-1536 distributions.

Klein [46] also examined the relation between activity at the nearby volcanoes of Kilauea and Mauna Loa. Separating the record into three parts at the longest repose of each volcano disclosed that the activities were inversely related. The rank correlation of the number of events in five year periods 1915–1980 was  $-0.646$ , significant at a level of 0.01, which indicates that they share the same magma source. Bebbington and Lai [5] repeated this analysis for the New Zealand volcanoes Ruapehu and Ngauruhoe, finding that the combined eruption sequence was indistinguishable from a superposition of indepen-

dent Poisson processes. This independence of the eruption sequences argues for separate magma sources.

**Alignments and Clusters**

There has also been considerable interest in detecting alignments or clusters of volcanic vents, in the hope that these can be related to geological features.

Lutz [50] proposed the ‘two-point azimuth method’, where the azimuth between every two vents is measured, generating  $n(n-1)/2$  measurements for  $n$  vents. These are then binned in  $10^\circ$  intervals, although the bin size could vary, and the result tested for departure from randomness. As non-circular fields and/or a non-homogeneous density of points can produce a preferred orientation, a correction must be made in these cases. One way of doing this is to use Monte Carlo simulation of random points to produce a reference distribution from which departures can be detected. Lutz and Gutmann [51] noted that using kernel smoothing methods for the vent location provides an improved distribution from which to perform the Monte Carlo simulation.

Wadge and Cross [100] suggested the use of the Hough transform. In this technique, a point  $(x, y)$  is converted into the normal parameterization  $(\rho, \theta)$ , where  $\rho = x \cos \theta + y \sin \theta$ . Thus each point generates a curve in the  $(\rho, \theta)$  plane as  $\theta$  varies, and multiple (more than two curves) intersections of these define the alignment of co-linear points. In practice the scales for  $\rho$  and  $\theta$  are quantized into discrete bins, which allows for a small degree of departure from strict co-linearity. Significance levels can be obtained via Monte Carlo techniques.

Connor [15] suggested the use of uniform kernel density fusion cluster analysis, using the distance metric

$$d(i, j) = \begin{cases} \frac{1}{2} \left( \frac{1}{m_i(w)} + \frac{1}{m_j(w)} \right), & \|x_i - x_j\| < w \\ \infty, & \text{otherwise,} \end{cases} \quad (50)$$

where  $\|\cdot\|$  denotes map distance, and  $m_i(w)$  is the number of cones within a circle of radius  $w$  centered on the  $i$ th vent. Cones are linked into clusters starting with the smallest values of  $d$ . Clusters are only combined if the ‘density fusion’  $d$  between the clusters is greater than the maximum  $d$  between any two events within either cluster. This makes it possible to recognize overlapping clusters. The search radius  $w$  is determined experimentally.

Connor [15] and Connor et al. [16] used the two-point azimuth method, Hough transform and cluster analysis to map clusters of multiple cones with common alignments in the TransMexican volcanic belt and Springerville volcanic field (Arizona).

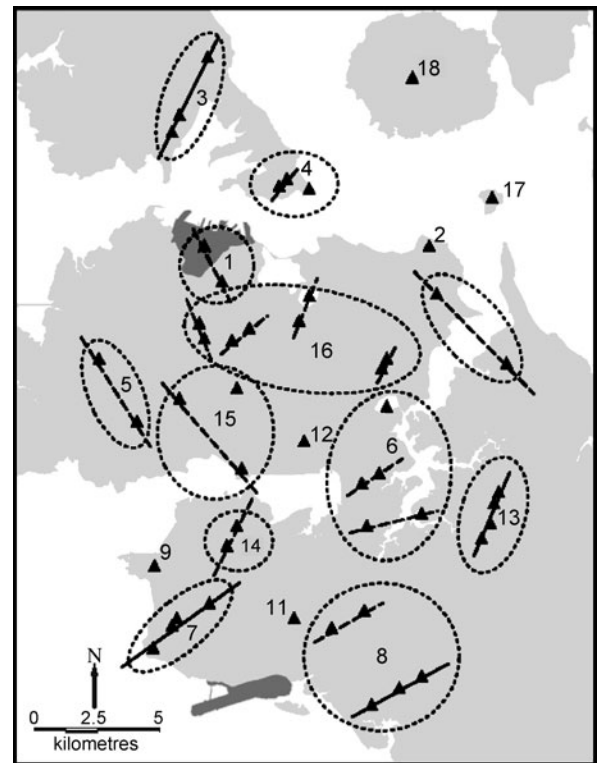
Magill et al. [52] measured vent clustering in the Auckland volcanic field using the statistic

$$L(d) = \sqrt{\frac{K(d)}{\pi}} - d, \quad (51)$$

where

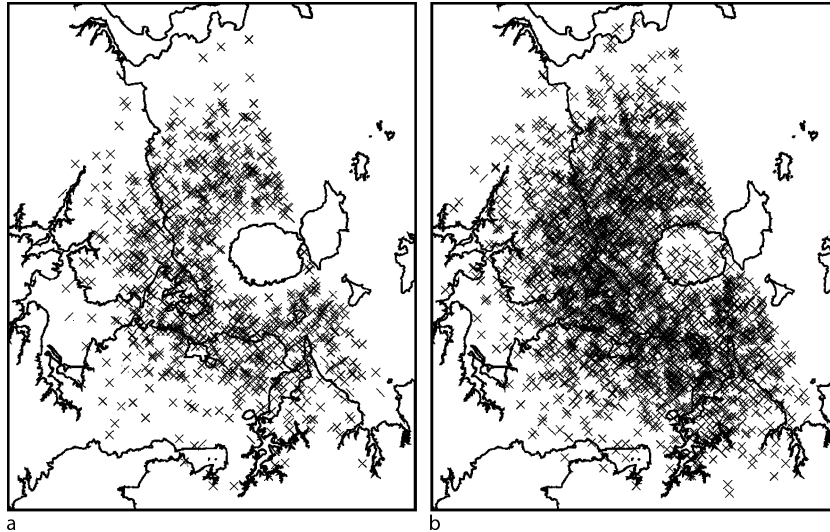
$$K(d) = \frac{A}{n} \sum_{j=1}^n \sum_{i \neq j} \frac{H(d - d_{ij})}{w_{ij}} \quad (52)$$

is Ripley’s K-function,  $A$  is the area of the field,  $d_{ij}$  is the distance between the  $i$ th and  $j$ th vents,  $H(\cdot)$  is the Heaviside function, and  $w_{ij}$  is an edge correction. Peaks of positive  $L(d)$  indicate clustering, and troughs of negative values regularity, at that distance. The Auckland field exhibits clustering at between 900 and 1600 m, and spatial regularity at about 4600 m. This clustering was used to collapse



Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 14

Groups and alignments of vents within the Auckland volcanic field, identified by position and age, and their relative order of eruption. Solid lines indicate alignments for which there are three or more vents. Dashed lines are alignments inferred from two vents. From [52], Figure 5, © 2005 International Association for Mathematical Geology. With kind permission of Springer Science and Business Media



Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 15  
 a event centroid locations, and b eruption locations for 1000 simulated events. From [52], Figure 8, © 2005 International Association for Mathematical Geology. With kind permission of Springer Science and Business Media

49 cones into 18 events. Magill et al. [52] then examined the distance separating consecutively erupted vents, and the distance between clustered eruptions, concluding that eruptions do not occur randomly, and occur preferentially closer to the previous eruption. Vent alignments in NE–SW and NW–SE directions were observed within clusters, as shown in Fig. 14. Using these observations, Magill et al. [52] constructed a Monte Carlo simulation model for the location of the next vent. Based on the last vent location (Rangitoto), the distance to the center of the next cluster was simulated from a distribution fitted to the observed distances. The number of vents was described by a negative binomial distribution, and vent locations were then simulated from a fitted distribution for the distance from the center of the cluster, modified by the observed alignment distribution, with the results shown in Fig. 15. Some 57% of the simulated eruption locations occur either on land, or within 1 km of the coast, rising to 69% if one considers the probability of at least one eruption from the next cluster falling within these limits.

### Yucca Mountain

After Mount Etna, the next most studied volcano is probably the volcanic field at Yucca Mountain, Nevada (cf. Fig. 16), due to the high-level radioactive waste repository planned there. The available data [13] consists of a large number of basaltic volcanoes, of ages 80,000 years to 10,500,000 years. These of course have large dating er-

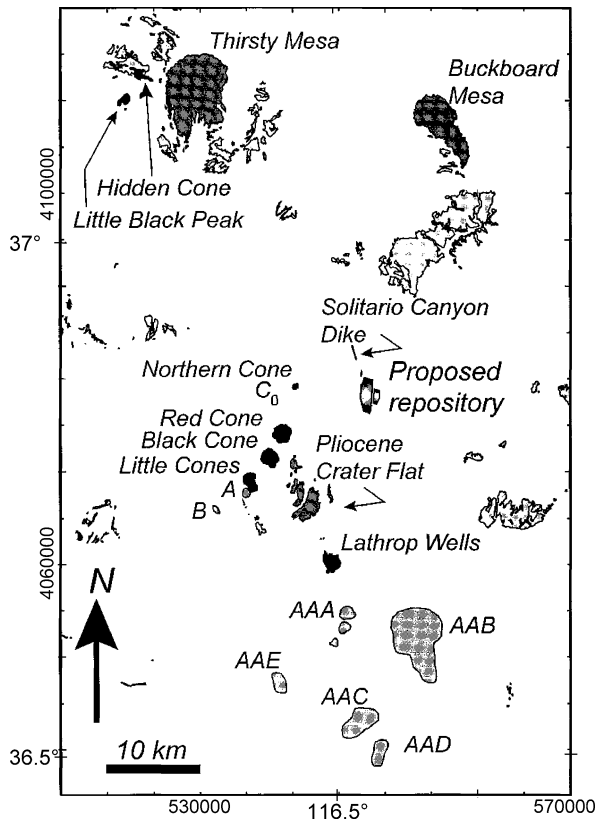
rors associated with them, but most analysis do not seem to have included this. The object is to estimate the probability of disruption of the repository by a volcanic intrusion. This is dependent on the probability of an intrusion, the distribution of the intrusion's dimensions, and on the dimensions of the repository itself.

The first analysis [23] used a Poisson process for the probability of an event, coupled with a Bernoulli probability of an event disrupting the repository:

$$\begin{aligned} \Pr(\text{no disruption before time } t) &= \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} (1-p)^n \\ &= e^{-\lambda t p}. \end{aligned} \quad (53)$$

The probability of an event being disruptive was estimated as  $p = a/A$ , where  $a$  is the area of the repository, and  $A$  the area, including the repository, used to define  $\lambda$ . The latter was estimated as  $N/T$ , where  $N$  is the number of scoria cones, and  $T$  the corresponding time period. Ho et al. [43] elaborated on means of estimating  $\lambda$  through event counts, repose intervals and magma volumes, and Ho [38] suggested a Bayesian framework for estimating  $p$ .

The next approach [38] attempted to incorporate a trend in the rate of volcanism using the Weibull process (4). As the data consisted of a number of clusters each containing several cones, but only the former had been dated, the event dates input to the Weibull process



Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 16  
 Distribution of basaltic rocks in the Yucca Mountain Region: *light gray shading*, Miocene; *dark gray shading*, Pliocene; *black shading*, Quaternary. Magnetic anomalies are shown in *light gray shading*, but only AAB is dated (Pliocene). Location is given in latitude/longitude, and in Universal Transverse Mercator (UTM) coordinate pairs (in meters). From [19], Figure 1b, © 2000 by the American Geophysical Union

consisted of multiples of each of a few dates. Depending on the details of the implementation, there was either a strong ( $\beta > 2$ ) increasing trend, or no significant trend at all. Notably, Ho [38] used a homogeneous Poisson process, with rate equal to the fitted Weibull process intensity at the present, for future prediction, although a continuing Weibull process was one possibility considered in the sensitivity analysis [40]. Ho and Smith [41] combined the Weibull process and Bayesian estimation of  $p$  in a formulation allowing the incorporation of ‘expert knowledge’.

Sheridan [80] included a spatial element by fitting a bivariate Gaussian distribution with five parameters (the  $(x, y)$  coordinates of the center, the length of the major and minor axes, and their orientation) to the volcanic field

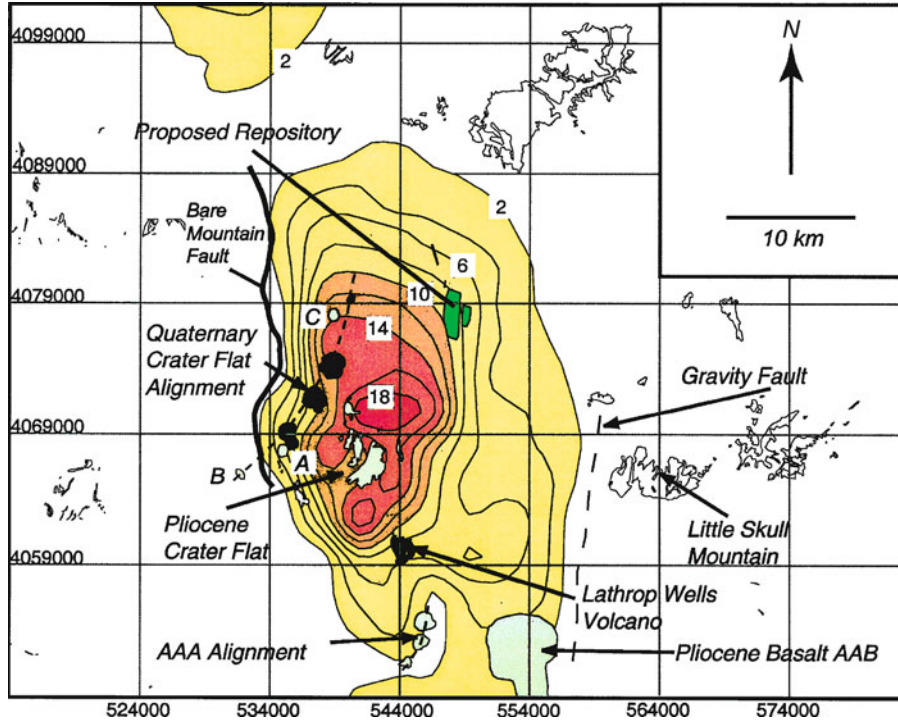
as a whole. Connor and Hill [17] used the spatio-temporal estimate (44), and integrated the resulting intensity over the area of the proposed repository, using a Gamma distribution for the area of a new volcano. The effects on the intensity of future events beyond the vicinity of the repository were ignored. Similar results were obtained using kernel-type estimators. Ho and Smith [42] elaborated on the latter by making the independent temporal component a Weibull process, while Connor et al. [19], while retaining the temporal Poisson process component, imposed a variety of geological constraints from crustal density, vent and fault alignments, and vent positions, on the spatial Gaussian kernel intensity in Fig. 17. A sensitivity analysis restricting the vent locations to various date-subsets was also conducted.

Crowe et al. [24], in their ‘final report’ of hazard studies at Yucca Mountain, summarized previous work, in particular that of Geomatrix Consultants [32]. They also fitted uniform, triangular (two varieties) and normal distributions to event count data, and simulated the results to obtain hazard estimates.

Smith et al. [85] suggested that the volcanism in the Yucca Mountain area could be episodic volcanism, and linked to an area of hot mantle. This would make possible another future peak of volcanism, with consequent underestimation of the recurrence rate. It was suggested that recurrence rates up to four times those previously used might be feasible. On the other hand, Coleman et al. [13] considered the lack of any detected dikes in or above the potential repository block, treating intrusions as an isolated Poisson process with zero observations in 13,000,000 years (although there was one ‘near-miss’). The conclusion was that high probabilities ( $10^{-6}$  per year) of disruption are ‘unrealistic’. Table 2 summarizes the published hazard estimates.

### Interactions with Earthquakes

A common question is whether an eruption is related to a preceding large regional earthquake (or vice-versa). Stress changes from earthquakes resolved at dike locations tend to be small, in fact smaller than those associated with solid earth tides. The exception are those earthquakes within a few fault-lengths. However, small increments of stress may of course advance the occurrence of an already imminent eruption. On a longer scale, as both eruptions and earthquakes are tectonic in nature, they must be related. Any such correlation may provide additional information about the likelihood of an imminent eruption which may be caused, or at least ‘detected’, by seismic activity. Alternatively, one might be interested in the likeli-



Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 17

The spatial recurrence rate (volcanic events/km<sup>2</sup>) contoured for the Yucca Mountain Region, based on the distribution of Quaternary volcanism and its relationship to the Bare Mountain Fault. The contour interval is  $2 \times 10^{-4}$  volcanic events/km<sup>2</sup>. Location is given in Universal Transverse Mercator (UTM) coordinate pairs (in meters). From [19], Plate 2, © 2000 by the American Geophysical Union

Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Table 2  
Summary of hazard assessment for Yucca Mountain

Source	Estimation Method		Probability of disruption in 10 <sup>5</sup> years
	Temporal	Spatial	
Crowe et al. [23]	Poisson process	$p = a/A$	$10^{-5}$ to $10^{-3}$
Ho et al. [43]	Poisson process	$(p)$	$0.5p$
Sheridan [80]	$(\lambda)$	bivariate Gaussian	$6\lambda \times 10^{-3}$ to $1.7\lambda \times 10^{-2}$
Ho [38]	Weibull (past)/ Poisson (future) process	Bayesian	$10^{-3}$ to $7 \times 10^{-3}$ (10 <sup>4</sup> years)
Ho [40]	Weibull process	Bayesian	$2 \times 10^{-5}$ to $7 \times 10^{-3}$
Connor and Hill [17]	Spatio-temporal		$10^{-4}$ to $5 \times 10^{-4}$ (10 <sup>4</sup> years)
Connor and Hill [17]	Poisson process	Kernel	$10^{-4}$ to $5 \times 10^{-4}$ (10 <sup>4</sup> years)
Geomatrix [32]	Expert Panel: Various		$5 \times 10^{-5}$ to $5 \times 10^{-3}$
Ho and Smith [41]	Weibull process	Bayesian	$10^{-3}$ to $3 \times 10^2$
Crowe et al. [24]	Summary: Various		$7 \times 10^{-5}$ to $4 \times 10^{-3}$
Ho and Smith [42]	Weibull process	Kernel	$10^{-4}$ to $6 \times 10^{-4}$
Connor et al. [19]	Weighted spatio-temporal		$10^{-3}$ to $10^{-2}$
Coleman et al. [13]	Poisson process		$10^{-3}$ to $10^{-2}$

hood of potentially dangerous earthquakes during a period of volcanic activity.

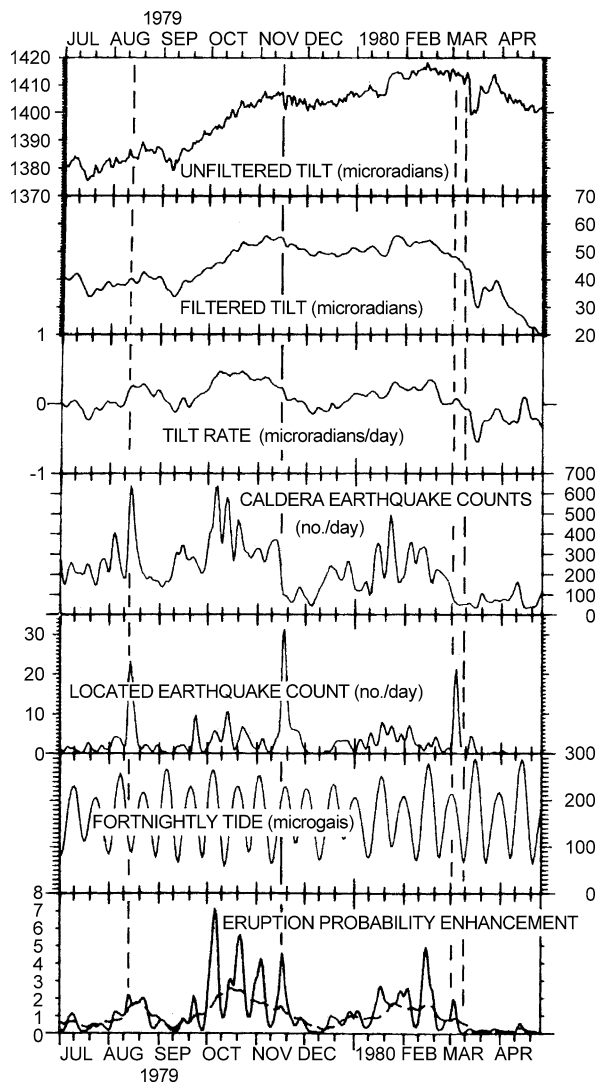
Sharp et al. [79] considered the relationship between eruptions of Mount Etna and earthquakes from various

structural zones of southern Italy. Both the earthquakes and eruption onsets were treated as Poisson processes, an assumption that had to be tested, with intensities  $\lambda_a$  and  $\lambda_b$ , respectively. A variable  $\xi(t)$  is formed by summing the

number of A-series events in the intervals  $(B_i, B_i + t)$ , where  $(B_i)$  are the observed times of the B-series events. If the two series are independent, then  $\xi(t)$  should have a Poisson distribution with mean  $\hat{\lambda} = \hat{\lambda}_a \hat{\lambda}_b T$ , where  $T$  is the length of the observation period, and  $\hat{\lambda}_i = N_i(T)/T$ , for  $i = a, b$ . For example, if B-series events consistently lead A-series events by some interval less than  $t$ ,  $\hat{\xi}$  will be significantly larger than  $\hat{\lambda}$ . The significance of an observed  $\hat{\xi}(t)$ , or confidence limits, can be obtained from the null distribution, which is Poisson with mean  $\hat{\lambda}$ . Repeating the procedure for varying  $t$  identifies significant correlations. Sharp et al. [79] found that the earthquakes were not Poissonian, but could be represented as a non-stationary Poisson process. A significant relationship was found between local earthquakes and subsequent flank eruptions within a few hundred days. Also, summit eruptions were found to occur before flank eruptions not preceded by earthquakes, but the summit eruptions themselves were not correlated with earthquakes. This suggests two mechanisms for flank eruptions of Etna: Those preceded by summit eruptions may be caused by magma pressure in the reservoir, while those preceded by earthquakes may be due to fracturing of the flank by earthquakes caused by tensile forces associated with the east-west extension of eastern Sicily.

Klein [47] examined various precursors of volcanic eruptions at Kilauea by examining the ratio of these precursors preceding an eruption to their long-term average value. The candidates included summit tilt, daily counts of small caldera earthquakes, repose since the last eruption, earth tides, and rainfall, as shown in Fig. 18. The number of caldera earthquakes was highly significant as a short-term (10 days or less) predictor, but tilt rate was a better predictor beyond 20 days. The fortnightly earth tidal modulation was also significant (cf. [54]), as it appears to trigger a volcano that is nearly ready to erupt anyway. Similarly, Mulargia et al. [66] used a statistical non-parametric pattern-recognition algorithm to examine possible precursors of eruptions at Mount Etna 1974–1990, including the number of earthquakes in five surrounding tectonic regions, and their maximum magnitude. An eighty day interval straddling the eruption onset was considered, and the pattern of 12 factors in each interval noted. The only statistically significant result was that flank eruptions were linked to seismicity with more than six earthquakes in the surrounding regions during the 80 day period. All 11 of the flank eruptions considered satisfied this condition, but only 11 out of 24 (in a total of 62) windows with this level of seismicity corresponded to earthquakes. Summit eruptions possessed no discernable pattern.

Mulargia [64] used the intensity cross product for two series of events at times  $0 < s_1 < \dots < s_{M(T)} < T$  and



Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 18

Kilauea volcano: Several parameters and eruption probability enhancement at an expanded time scale. The figure spans 10 months during 1979–1980 and includes the November 16, 1979, eruption (solid vertical line), and intrusions on August 12, 1979, and March 2 and 10, 1980 (dashed vertical lines). Intrusions, like eruptions, are generally accompanied by summit deflation, an earthquake swarm and harmonic tremor, but magma fails to reach the surface. The eruption probability enhancement is plotted as a 1-day outlook (solid line) and 30-day outlook (dashed line). The 1-day probability depends on the fortnightly tide and oscillates with it. The 30-day probability is a function of smoothed tilt and earthquake parameters and does not depend on the tide. Note that eruption probability enhancement was higher than average (greater than 1) for about 2 months prior to the November 1979 eruption and March 1980 intrusions. From [47], Figure 4 © 1984 by the American Geophysical Union

$0 < t_1 < \dots < t_{N(T)} < T$ . For a lag  $u$ , this is estimated by

$$\hat{p}_{MN}(u) = \frac{n(u, h)}{2hT}, \tag{54}$$

where  $n(u, h) = \#\{i: t_j + u - h \leq s_i \leq t_j + u + h, \text{ for some } j\}$  is the number of events from the process  $M$  that occur within a window of length  $2h$  of the lag of a point in the process  $N$ . Under a null hypothesis of two independent Poisson processes, the significance level can be calculated as

$$P = \sum_{k=n(u, h)}^{N(T)} \binom{k}{N(T)} \left(\frac{2hM(T)}{T}\right)^k. \tag{55}$$

This was applied to the flank eruptive activity and local seismicity at Mt Etna 1974–1990. The data consisted of 11 eruptions and 12 seismic sequences [31]. Using a window of length  $2h = 10$  days, Mulargia [64] found a highly significant association peak at a lag  $u$  between  $-7$  and  $3$  days. This indicates that local earthquakes and flank eruptions are largely concomitant phenomena.

Marzocchi et al. [58] examined the correlation between eruptions of Vesuvius and earthquakes of  $M > 5.4$  in the southern Apennines, Calabrian Arc and Sicily, 1631–1944. The eruptions and earthquakes were binned into intervals of 1, 2, 3 and 4 years, and the Spearman rank correlation

$$\rho_{XY}(k) = \frac{\sum_{i=1}^m R(X_{i+k})R(Y_i) - m(m+1)^2/4}{\left(\sum_{i=1}^m R(X_{i+k})^2 - m(m+1)^2/4\right)^{1/2} \cdot \left(\sum_{i=1}^m R(Y_i)^2 - m(m+1)^2/4\right)^{1/2}} \tag{56}$$

calculated, where  $X_i, Y_i$  are the total seismic moment and number of eruptions in the  $i$ th time bin,  $R(\cdot)$  is the rank, from smallest to largest, of the quantities, and  $k$  is the lag. Varying this lag produces the Spearman correlogram. Marzocchi et al. [58] observed changes in eruptive activity following increases in seismicity in the southern Apennines after 6–13 years, and with delays of 27–30 years and 36–39 years for earthquakes in Sicily and the Calabrian arc, respectively. Nostro et al. [70] investigated this coupling (cf. Fig. 19) in terms of the Coulomb stress change on a magma body beneath Vesuvius caused by a slip on Apennine normal faults (promoting eruptions by earthquakes) and on fault planes caused by voiding a buried magma body (promoting earthquakes by eruptions), as shown in Fig. 20. Using these results to select earthquakes and eruptions improved the significance of the Marzocchi et al. [58] analysis.

Cardaci et al. [10] used a cross-correlation technique to investigate the relation between seismic and volcanic data

at Mount Etna 1975–1987. The seismic and eruptive data were averaged over seven day intervals, with the cross-correlation being

$$\rho_{XY}(k) = \frac{\sigma_{XY}(k)}{\sqrt{\sigma_X^2 \sigma_Y^2}} \tag{57}$$

for  $k = 0, \pm 1, \dots, \pm K$ , where

$$\sigma_{XY}(k) = \begin{cases} (1/n) \sum_{i=1}^{n-k} (X_i - \mu_X)(Y_{i+k} - \mu_Y) & k = 0, 1, \dots, K \\ (1/n) \sum_{i=1-k}^n (X_i - \mu_X)(Y_{i+k} - \mu_Y) & k = -1, \dots, -K \end{cases} \tag{58}$$

is the cross covariance function and  $\mu_X, \mu_Y$  and  $\sigma_X^2, \sigma_Y^2$  are the means and variances, respectively, of the two time series  $X_1, \dots, X_n$  and  $Y_1, \dots, Y_n$ . The significance of the cross-correlation can be assessed from the fact that  $\rho_{XY}(k) \sqrt{(m-2)/(1-\rho_{XY}^2(k))}$ , is asymptotically distributed as a  $t$ -distribution with  $m-2$  degrees of freedom, where  $m$  is the number of overlapped positions between the two series. Cardaci et al. [10] arrived at a similar conclusion to Mulargia et al. [66], identifying low frequency events, whose rate of occurrence increases from 17 to 108 days prior to flank eruptions, as the best precursor. Summit events had a less clear relation with volcanic tremor rather than seismic activity.

Linde and Sacks [48] examined the historical record of eruptions ( $VEI \geq 2$ ) and large earthquakes, to see whether there are more of the former within two days of the latter. For earthquakes with ( $M \geq 8$ ), there are a significant excess of eruptions within 750 km of the epicenter, while for earthquakes with  $7.0 \leq M \leq 7.9$ , the effect is only observed within a radius of 200 km. In all, 20 such earthquake-volcano pairs are observed 1587–1974. In addition, Linde and Sacks [48] hypothesized that eruptions in unison of volcanoes separated by hundreds of kilometers could be due to the second eruption being triggered by earthquakes associated with the first. Marzocchi [56] elaborated on this analysis by defining a perturbation function

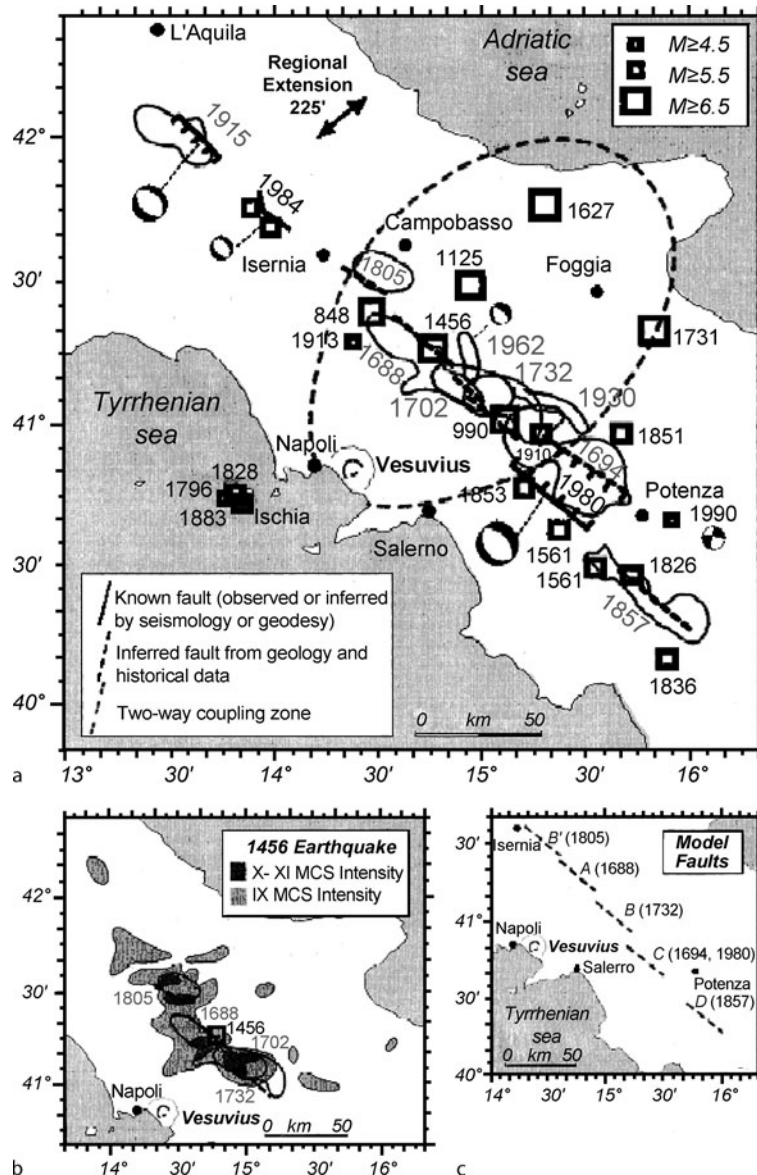
$$\phi_i^{(k)}(\Delta) = \sum_j M_j w(d_{jk}) \mathbf{I}_{((i-1)\Delta, i\Delta)}(t_k - s_j), \tag{59}$$

where

$$\mathbf{I}_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases} \tag{60}$$

is the indicator function,  $s_j$  and  $M_j$  are the time and seismic moment of the  $j$ th earthquake,  $t_k$  is the onset time of

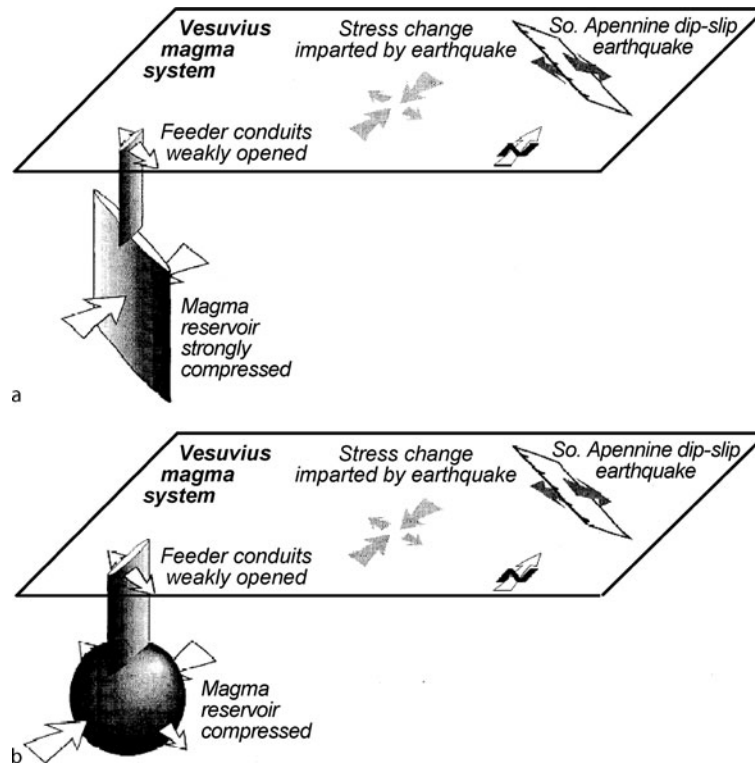




Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 19  
**a** Sites of historical earthquakes in south central Italy, associated faults and Vesuvius volcano. Epicentres or, if unknown, isoseismals of modified Mercalli intensity, are shown. The coupling zone is the region in which normal-faulting earthquakes may promote Vesuvius eruptions and vice versa. **b** Isoseismals for the 1456 earthquake with 1688–1805 isoseismal zones. **c** Simplified model faults A–D used for calculations, with the dates of earthquakes associated with each fault segment. From [70], Figure 1 © 1998 by the American Geophysical Union

the  $k$ th eruption, and  $i$  is the lag. The distance between the epicenter of the  $j$ th earthquake and the  $k$ th eruption is  $d_{jk}$ , which is weighted by a (sharply) decreasing function  $w(\cdot)$ . The statistical significance of the calculated values is assessed by a randomization test. The idea is that the perturbation on the volcano by earthquakes is proportional to the seismic moment and decays with distance. Of the

eight eruptions with  $VEI \geq 5$  examined, using  $\Delta = 3$  and 5 years, five have a significant peak at 0–5 years in the perturbation function, calculated from earthquakes with  $M \geq 7$  (792 events, 1900–1999). Using an earthquake catalog with  $M \geq 6$  (3186 events, 1950–1997), two of four events have a significant peak at 30–35 years before the eruption. Marzocchi et al. [60] expanded this analysis to



Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 20

Schematic illustration of the response of a hypothetical Vesuvius magmatic system to a southern Apennine normal-faulting earthquake for (a) a buried dike in which at least one fissure or feeder conduit strikes NE and (b) a buried spherical magma chamber with a NE striking fissure. From [70], Figure 3 © 1998 by the American Geophysical Union

the 62 20th Century eruptions with  $VEI \geq 4$ , identifying peaks at 0–5 and 20–35 years. The perturbations for  $VEI \geq 4$  eruptions were shown to be significantly larger than those for  $VEI = 1$  and  $VEI = 2$  eruptions. This suggests that the likelihood of large explosive eruptions is affected by post-seismic stress variations induced by large earthquakes up to decades earlier and hundreds of km distant. No spatio-temporal clustering of the earthquakes and eruptions was detected.

### Future Directions

There are two places in the above review where the effect of a second process on the likelihood of eruptions has been identified, but an appropriate model incorporating this is lacking. These are deterministic (fortnightly and/or annual) cycles in the intensity, and correlation between eruptions and large earthquakes. A candidate to model the first is a cyclic Poisson process [97], and the second could be examined via a bivariate Point process, or perhaps a non-homogeneous hidden Markov model.

More generally, much of the relevant literature on stochastic processes has not been applied to problems in volcanology. In particular, (marked) point processes, alternating renewal processes, and semi-Markov processes would seem to be natural candidates for modeling volcanic activity. A number of techniques from time series analysis (see [106] for a brief review) may prove useful for the investigation of volcanic data at various time scales.

In the application of stochastic models, and the understanding to be gleaned from them, volcanology lags well behind seismology. This is despite the inherent advantages of a phenomenon in which spatial degrees of freedom can often be ignored. An obvious future direction is the borrowing of more techniques from seismology. Unlike seismology, a certain amount of effort (and money) can often produce additional data through coring and dating. This is a great advantage in statistical terms, especially as the data can be 'targeted' to some degree as required by the model. This additional data makes it possible in volcanology to detect, and forecast, non-stationary behavior, and to relate this to physical changes in the volcano [95].

Although this article has concentrated on models for occurrence patterns, actual volcanic hazard is dependent on the nature and behavior of volcanic products such as lava, ash fall, lahars and pyroclastic flows. Incorporating existing sophisticated numerical models of such behavior with stochastic models of occurrence, location/direction and magnitude will allow more accurate estimation of the hazard to life and property.

## Bibliography

### Primary Literature

- Bacon CR (1982) Time-predictable bimodal volcanism in the Coso range, California. *Geology* 10:65–69
- Bain LJ (1978) *Statistical Analysis of Reliability and Life Testing Models*. Marcel Dekker, New York
- Bebbington MS (2007) Identifying volcanic regimes using hidden Markov models. *Geophys J Int*, 171:921–942
- Bebbington MS, Lai CD (1996) On nonhomogeneous models for volcanic eruptions. *Math Geol* 28:585–600
- Bebbington MS, Lai CD (1996) Statistical analysis of New Zealand volcanic occurrence data. *J Volcanol Geotherm Res* 74:101–110
- Bebbington MS, Lai CD (1998) A generalised negative binomial and applications. *Commun Statist Theor Methods* 27:2515–2533
- Behncke B, Neri M (2003) Cycles and trends in the recent eruptive behaviour of Mount Etna (Italy). *Canadian J Earth Sci* 40:1405–1411
- Bronk Ramsey C (1995) Radiocarbon calibration and analysis of stratigraphy: The OxCal program. *Radiocarbon* 37:425–430
- Burt ML, Wadge G, Scott WA (1994) Simple stochastic modelling of the eruption history of a basaltic volcano: Nyamuragira, Zaire. *Bull Volcanol* 56:87–97
- Cardaci C, Falsaperla S, Gasperini P, Lombardo G, Marzocchi W, Mulargia F (1993) Cross-correlation analysis of seismic and volcanic data at Mt Etna volcano, Italy. *Bull Volcanol* 55: 596–603
- Carta S, Figari R, Sartoris G, Sassi R, Scandone R (1981) A statistical model for Vesuvius and its volcanological implications. *Bull Volcanol* 44:129–151
- Casetti G, Frazzetta G, Romano R (1981) A statistical analysis in time of the eruptive events on Mount Etna (Italy) from 1323 to 1980. *J Volcanol Geotherm Res* 44:283–294
- Coleman NM, Abramson LR, Marsh BD (2004) Testing claims about volcanic disruption of a potential geologic repository at Yucca Mountain, Nevada. *Geophys Res Lett* 31:L24601
- Condit CD, Connor CB (1996) Recurrence rates of volcanism in basaltic volcanic fields: an example from the Springerville volcanic field, Arizona. *Geol Soc Am Bull* 108:1225–1241
- Connor CB (1990) Cinder cone clustering in the TransMexican volcanic belt: Implications for structural and petrologic models. *J Geophys Res* 95:19395–19405
- Connor CB, Condit CD, Crumpler LS, Aubele JC (1992) Evidence of regional structural controls on vent distribution: Springerville volcanic field, Arizona. *J Geophys Res* 97: 349–359
- Connor CB, Hill BE (1995) Three nonhomogeneous Poisson models for the probability of basaltic volcanism: application to the Yucca Mountain region, Nevada. *J Geophys Res* 100:10107–10125
- Connor CB, Sparks RSJ, Mason RM, Bonadonna C, Young SR (2003) Exploring links between physical and probabilistic models of volcanic eruptions: The Soufriere Hills volcano, Montserrat. *Geophys Res Lett* 30:1701
- Connor CB, Stamatakos JA, Ferrill DA, Hill BE, Ofoegbu GI, Conway FM, Sagar B, Trapp J (2000) Geologic factors controlling patterns of small-volume basaltic volcanism: Application to a volcanic hazards assessment at Yucca Mountain, Nevada. *J Geophys Res* 105:417–432
- Conway FM, Connor CB, Hill BE, Condit CD, Mullaney K, Hall CM (1998) Recurrence rates of basaltic volcanism in SP cluster, San Francisco volcanic field, Arizona. *Geology* 26:655–658
- Crandell DR, Mullineaux DR, Rubin M (1975) Mount St. Helens volcano: Recent and future behaviour. *Science* 187:438–441
- Cronin S, Bebbington M, Lai CD (2001) A probabilistic assessment of eruption recurrence on Taveuni volcano, Fiji. *Bull Volcanol* 63:274–288
- Crowe BM, Johnson ME, Beckman RJ (1982) Calculation of the probability of volcanic disruption of a high-level radioactive waste repository within southern Nevada, USA. *Radioact Waste Manag* 3:167–190
- Crowe BM, Wallmann P, Bowker LM (1998) Probabilistic modeling of volcanism data: Final volcanic hazard studies for the Yucca Mountain site. In: Pery FV et al. (eds) *Volcanism Studies: Final Report for the Yucca Mountain project*. Los Alamos National Laboratory Report LA-13478, Los Alamos National Laboratory, Los Alamos, 415 pp
- Decker RW (1986) Forecasting volcanic eruptions. *Ann Rev Earth Planet Sci* 14:267–291
- De la Cruz-Reyna S (1991) Poisson-distributed patterns of explosive eruptive activity. *Bull Volcanol* 54:57–67
- De la Cruz-Reyna S (1993) Random patterns of occurrence of explosive eruptions at Colima volcano, Mexico. *J Volcanol Geotherm Res* 55:51–68
- De la Cruz-Reyna S, Carrasco-Nunez G (2002) Probabilistic hazard analysis of Citlaltepeltl (Pico de Orizaba) volcano, eastern Mexican volcanic belt. *J Volcanol Geotherm Res* 113: 307–318
- Dubois J, Cheminee JL (1991) Fractal analysis of eruptive activity of some basaltic volcanoes. *J Volcanol Geotherm Res* 45:197–208
- Eliasson J, Larsen G, Gudmundsson MT, Sigmundsson F (2006) Probabilistic model for eruptions and associated flood events in the Katla caldera, Iceland. *Comput Geosci* 10: 179–200
- Gasperini P, Gresta S, Mulargia F (1990) Statistical analysis of seismic and eruptive activities at Mt. Etna during 1978–1987. *J Volcanol Geotherm Res* 40:317–325
- Geomatrix Consultants (1996) Probabilistic volcanic hazard analysis for Yucca Mountain, Nevada. Report BA0000000-1717-220-00082. Geomatrix Consultants, San Francisco
- Godano C, Civetta L (1996) Multifractal analysis of Vesuvius volcano eruptions. *Geophys Res Lett* 23:1167–1170
- Gusev AA, Ponomareva VV, Braitseva OA, Melekestsev IV, Sulerzhitsky LD (2003) Great explosive eruptions on Kamchatka during the last 10,000 years: Self-similar irregularity of the output of volcanic products. *J Geophys Res* 108(B2):2126

35. Guttorp P, Thompson ML (1991) Estimating second-order parameters of volcanicity from historical data. *J Amer Statist Assoc* 86:578–583
36. Ho CH (1990) Bayesian analysis of volcanic eruptions. *J Volcanol Geotherm Res* 43:91–98
37. Ho CH (1991) Nonhomogeneous Poisson model for volcanic eruptions. *Math Geol* 23:167–173
38. Ho CH (1992) Risk assessment for the Yucca Mountain high-level nuclear waste repository site: Estimation of volcanic disruption. *Math Geol* 24:347–364
39. Ho CH (1992) Statistical control chart for regime identification in volcanic time-series. *Math Geol* 24:775–787
40. Ho CH (1995) Sensitivity in volcanic hazard assessment for the Yucca Mountain high-level nuclear waste repository site: The model and the data. *Math Geol* 27:239–258
41. Ho CH, Smith EI (1997) Volcanic hazard assessment incorporating expert knowledge: application to the Yucca Mountain region, Nevada, USA. *Math Geol* 29:615–627
42. Ho CH, Smith EI (1998) A spatial-temporal/3-D model for volcanic hazard assessment: application to the Yucca Mountain region, Nevada. *Math Geol* 30:497–510
43. Ho CH, Smith EI, Feuerbach DL, Naumann TR (1991) Eruptive probability calculation for the Yucca Mountain site, USA: Statistical estimation of recurrence rates. *Bull Volcanol* 54:50–56
44. Jaquet O, Low S, Martinelli B, Dietrich V, Gilby D (2000) Estimation of volcanic hazards based on Cox stochastic processes. *Phys Chem Earth (A)* 25:571–579
45. Jupp TE, Pyle DM, Mason BG, Dade WB (2004) A statistical model for the timing of earthquakes and volcanic eruptions influenced by periodic processes. *J Geophys Res* 109:B02206
46. Klein FW (1982) Patterns of historical eruptions at Hawaiian volcanoes. *J Volcanol Geotherm Res* 12:1–35
47. Klein FW (1984) Eruption forecasting at Kilauea Volcano, Hawaii. *J Geophys Res* 89:3059–3073
48. Linde AT, Sacks IS (1998) Triggering of volcanic eruptions. *Nature* 395:888–890
49. Lockwood JP (1995) Mauna Loa eruptive history – the preliminary radiocarbon record, Hawai'i. In: Rhodes JM, Lockwood JP (eds) *Mauna Loa Revealed: Structure, Composition, History, and Hazards*. American Geophysical Union Monograph, vol 92. American Geophysical Union, Washington DC, pp 81–94
50. Lutz TM (1986) An analysis of the orientation of large-scale crustal structures: A statistical approach based on areal distributions of pointlike features. *J Geophys Res* 91:421–434
51. Lutz TM, Gutmann JT (1995) An improved method for determining and characterizing alignments of pointlike features and its implications for the Pinacate volcanic field, Sonora, Mexico. *J Geophys Res* 100:17659–17670
52. Magill CR, McAnaney KJ, Smith IEM (2005) Probabilistic assessment of vent locations for the next Auckland volcanic field event. *Math Geol* 37:227–242
53. Martin AJ, Umeda K, Connor CB, Weller JN, Zhao D, Takahashi M (2004) Modeling long-term volcanic hazards through Bayesian inference: An example from the Tohoku volcanic arc, Japan. *J Geophys Res* 109:B10208
54. Martin DP, Rose WI (1981) Behavioral patterns of Fuego volcano, Guatemala. *J Volcanol Geotherm Res* 10:67–81
55. Marzocchi W (1996) Chaos and stochasticity in volcanic eruptions the case of Mount Etna and Vesuvius. *J Volcanol Geotherm Res* 70:205–212
56. Marzocchi W (2002) Remote seismic influence on large explosive eruptions. *J Geophys Res* 107:2018. doi:10.1029/2001JB000307
57. Marzocchi W, Sandri L, Gasparini P, Newhall C, Boschi E (2004) Quantifying probabilities of volcanic events: The example of volcanic hazard at Mount Vesuvius. *J Geophys Res* 109:B11201
58. Marzocchi W, Scandone R, Mulargia F (1993) The tectonic setting of Mount Vesuvius and the correlation between its eruptions and the earthquakes of the southern Apennines. *J Volcanol Geotherm Res* 58:27–41
59. Marzocchi W, Zaccarelli L (2006) A quantitative model for the time-size distribution of eruptions. *J Geophys Res* 111: B04204
60. Marzocchi W, Zaccarelli L, Boschi E (2004) Phenomenological evidence in favor of a remote seismic coupling for large volcanic eruptions. *Geophys Res Lett* 31:L04601
61. Mason BG, Pyle DM, Dade WB, Jupp T (2004) Seasonality of volcanic eruptions. *J Geophys Res* 109:B04206
62. Mauk FJ, Johnston MJS (1973) On the triggering of volcanic eruptions by earth tides. *J Geophys Res* 78:3356–3362
63. Medina Martinez F (1983) Analysis of the eruptive history of the Volcan de Colima, Mexico (1560–1980). *Geof Int* 22: 157–178
64. Mulargia F (1992) Time association between series of geophysical events. *Phys Earth Plan Int* 71:147–153
65. Mulargia F, Gasparini P, Tinti S (1987) Identifying regimes in eruptive activity: An application to Etna volcano. *J Volcanol Geotherm Res* 34:89–106
66. Mulargia F, Marzocchi W, Gasparini P (1992) Statistical identification of physical patterns which accompany eruptive activity on Mount Etna, Sicily. *J Volcanol Geotherm Res* 53: 289–296
67. Mulargia F, Tinti S, Boschi E (1985) A statistical analysis of flank eruptions on Etna volcano. *J Volcanol Geotherm Res* 23: 263–272
68. Newhall CG, Self S (1982) The volcanic explosivity index (VEI): An estimate of explosive magnitude for historical volcanism. *J Geophys Res* 87:1231–1238
69. Nishi Y, Inoue M, Tnaka T, Murai M (2001) Analysis of time sequences of explosive volcanic eruptions of Sakurajima. *J Phys Soc Japan* 70:1422–1428
70. Nostro C, Stein RS, Cocco M, Belardinelli ME, Marzocchi W (1998) Two-way coupling between Vesuvius eruptions and southern Apennine earthquakes, Italy, by elastic stress transfer. *J Geophys Res* 103:24487–24504
71. Pyle DM (1998) Forecasting sizes and repose times of future extreme volcanic events. *Geology* 26:367–370
72. Reymont RA (1969) Statistical analysis of some volcanologic data regarded as series of point events. *Pure Appl Geophys* 74:57–77
73. Salvi F, Scandone R, Palma C (2006) Statistical analysis of the historical activity of Mount Etna, aimed at the evaluation of volcanic hazard. *J Volcanol Geotherm Res* 154:159–168
74. Sandri L, Marzocchi W, Gasparini P (2005) Some insights on the occurrence of recent volcanic eruptions of Mount Etna volcano (Sicily, Italy). *Geophys J Int* 163:1203–1218
75. Santacroce R (1983) A general model for the behaviour of the Somma-Vesuvius volcanic complex. *J Volcanol Geotherm Res* 17:237–248

76. Scandone R, Arganese G, Galdi F (1993) The evaluation of volcanic risk in the Vesuvian area. *J Volcanol Geotherm Res* 58:263–271
77. Scandone R, Giacomelli L, Gasparini P (1993) Mount Vesuvius: 2000 years of volcanological observations. *J Volcanol Geotherm Res* 58:5–25
78. Settle M, McGetchin TR (1980) Statistical analysis of persistent explosive activity at Stromboli, 1971: Implications for eruption prediction. *J Volcanol Geotherm Res* 8:45–58
79. Sharp ADL, Lombardo G, David PM (1981) Correlation between eruptions of Mount Etna, Sicily, and regional earthquakes as seen in historical records from AD 1582. *Geophys J R astr Soc* 65:507–523
80. Sheridan MF (1992) A Monte Carlo technique to estimate the probability of volcanic dikes. In: High-Level Radioactive Waste Management: Proceedings of the Third Annual International Conference, Las Vegas, April 12–16, 1992. American Nuclear Society, La Grange Park, pp 2033–2038
81. Shimazaki K, Nakata T (1980) Time-predictable recurrence model for large earthquakes. *Geophys Res Lett* 7:279–282
82. Siebert L, Simkin T (2002) *Volcanoes of the World: an Illustrated Catalog of Holocene Volcanoes and their Eruptions*, Smithsonian Institution, Global Volcanism Program Digital Information Series, GVP-3. <http://www.volcano.si.edu/world/>. Accessed 20 Jun 2008
83. Simkin T (1993) Terrestrial volcanism in space and time. *Ann Rev Earth Planet Sci* 21:427–452
84. Simkin T (1994) Distant effects of volcanism – how big and how often? *Science* 264:913–914
85. Smith EI, Keenan DL, Plank T (2002) Episodic volcanism and hot mantle: Implications for volcanic hazard studies at the proposed nuclear waste repository at Yucca Mountain, Nevada. *GSA Today* April 2002:4–9
86. Solow AR (1993) Estimating record inclusion probability. *The Amer Statist* 47:206–208
87. Solow AR (2001) An empirical Bayes analysis of volcanic eruptions. *Math Geol* 33:95–102
88. Sornette A, Dubois J, Cheminee JL, Sornette D (1991) Are sequences of volcanic eruptions deterministically chaotic? *J Geophys Res* 96:11931–11945
89. Stothers RB (1989) Seasonal variations of volcanic eruption frequencies. *Geophys Res Lett* 16:453–455
90. Takada A (1997) Cyclic flank-vent and central-vent eruption patterns. *Bull Volcanol* 58:539–556
91. Telesca L, Cuomo V, Lapenna V, Macchiato M (2002) Time-clustering analysis of volcanic occurrence sequences. *Phys Earth Planet Int* 131:47–62
92. Telesca L, Lapenna V (2005) Identifying features in time-occurrence sequences of volcanic eruptions. *Environmetrics* 16:181–190
93. Thorlaksson JE (1967) A probability model of volcanoes and the probability of eruptions of Hekla and Katla. *Bull Volcanol* 31:97–106
94. Turner M, Cronin S, Bebbington M, Platz T (2008) Developing a probabilistic eruption forecast for dormant volcanoes; a case study from Mt Taranaki, New Zealand. *Bull Volcanol* 70: 507–515
95. Turner M, Cronin S, Smith I, Bebbington M, Stewart RB (2008) Using titanomagnetite textures to elucidate volcanic eruption histories. *Geology* 36:31–34
96. Vere-Jones D (1992) Statistical methods for the description and display of earthquake catalogs. In: Walden AT, Guttorp P (eds) *Statistics in the Environmental and Earth Sciences*, Edward Arnold, London, pp 220–246
97. Vere-Jones D, Ozaki T (1982) Some examples of statistical estimation applied to earthquake data. *Ann Inst Statist Math* 34:189–207
98. Voight B (1988) A method for prediction of volcanic eruptions. *Nature* 332:125–130
99. Wadge G (1982) Steady state volcanism: Evidence from eruption histories of polygenetic volcanoes. *J Geophys Res* 87:4035–4049
100. Wadge G, Cross A (1988) Quantitative methods for detecting aligned points: An application to the volcanic vents of the Michoacan–Guanajuato volcanic field, Mexico. *Geology* 16: 815–818
101. Wadge G, Guest JE (1981) Steady-state magma discharge at Etna 1971–1981. *Nature* 294:548–550
102. Wadge G, Walker WPL, Guest JE (1975) The output of Etna volcano. *Nature* 255:385–387
103. Wickman FE (1966) Repose-period patterns of volcanoes. I. Volcanic eruptions regarded as random phenomena. *Arch Mineral Geol* 4:291–367
104. Wickman FE (1966) Repose-period patterns of volcanoes. V. General discussion and a tentative stochastic model. *Arch Mineral Geol* 4:351–367
105. Wickman FE (1976) Markov models of repose-period patterns of volcanoes. In: Merriam DF (ed) *Random Processes in Geology*. Springer, New York, pp 135–161
106. Young PC (2006) New approaches to volcanic time-series analysis. In Mader HM, Coles SG, Connor CB, Connor LJ (eds) *Statistics in Volcanology*. Geological Society of London, London, pp 143–160

### Books and Reviews

- Guttorp P (1995) *Stochastic Modeling of Scientific Data*, Chapman and Hall, London
- Hill DP, Pollitz FP, Newhall C (2002) Earthquake-volcano interactions. *Phys Today* 55:41–47
- Lindsay JK (2004) *Statistical Analysis of Stochastic Processes in Time*. Cambridge University Press, Cambridge
- Karr AF (1991) *Point Processes and Their Statistical Inference*, 2nd edn. Marcel Dekker, New York
- Mader HM, Coles SG, Connor CB, Connor LJ (eds) (2006) *Statistics in Volcanology*. Geological Society of London, London

## Volcanic Hazards and Early Warning

ROBERT I. TILLING  
Volcano Hazards Team, US Geological Survey,  
Menlo Park, USA

### Article Outline

Glossary  
Definition of the Subject  
Introduction  
Scope of Problem:  
Challenge for Emergency–Management Authorities  
Variability in Possible Outcomes of Volcano Unrest  
Some Recent Examples of Actual Outcomes  
of Volcano Unrest  
Challenges in Achieving Refined Predictive Capability  
Future Directions  
Acknowledgments  
Bibliography

### Glossary

**Volcano hazards** Potentially damaging volcano-related processes and products that occur during or following eruptions [see [31,37] for overviews]. In quantitative hazards assessments, the probability of a given area being affected by potentially destructive phenomena within a given period of time.

**Volcano risk** Probability of harmful consequences – individual or societal – or expected losses (deaths, injuries, property, livelihoods, economic activity disrupted or environment damaged) resulting from interactions between volcano hazards, human development, and vulnerable conditions. Though definitions vary in detail, risk is conventionally expressed by the general relation:  $risk = hazards \times vulnerability$  [13].

**Vulnerability** The conditions determined by physical, social, economic, and environmental factors or processes, which increase the susceptibility of a community to the impact of hazards [13].

**Volcano status** The current activity or mode of behavior of a volcano. Status is commonly described as follows: *active* (having one or more recorded historical eruptions); *dormant* (currently inactive but having potential for renewed eruptive activity); and *extinct* (dormant for long time and not expected to erupt again). These terms, while commonly used in the scientific literature, are imperfect and are undergoing serious reexamination within the global volcanological community.

**Factual statement** Following the recommended definition of Swanson et al. [35], is the description of the current status and conditions of a volcano but does not anticipate future events.

**Forecast** As defined by Swanson et al. [35], is a comparatively imprecise statement of the time, place, and nature of expected activity; forecasts of eruptions and earthquakes sometimes are probabilistic (e. g., [43,47]). A forecast usually covers a longer time period than that for a prediction.

**Prediction** As defined by Swanson et al. [35], is a comparatively precise statement of the time, place, and, ideally, the nature and size of impending activity. Forecasts and predictions can be either *long term* (typically years to decades or longer) or *short term* (typically hours to months).

**Volcano unrest** Visual and (or) measurable physical and (or) chemical changes, surface or subsurface, in the status of the volcano, relative to its “normal” historical behavior; duration of unrest can vary from hours to decades. The initiation or escalation of volcano unrest, regardless of duration, may or may not culminate in eruption.

**Magma intrusion** The subsurface movement or injection of *magma* (molten rock containing associated crystals and gases) from one part of a volcanic system into another. Typically, an intrusion involves transport of magma from a central zone of storage (i. e., magma “reservoir”) into peripheral, structurally weaker zones (e. g., faults or rifts). Some intrusions culminate in surface eruptive outbreaks, others do not.

**Volcano monitoring** The systematic collection, analysis, and interpretation of visual observations and instrumental measurements of changes at volcanoes before, during, and after the onset of volcano unrest and (or) eruptive activity.

**“Baseline” monitoring data** Volcano-monitoring data acquired for a volcanic system documenting its range of variation during its “normal” behavior prior to volcano unrest and (or) eruptions. The longer the time span covered by “baseline” monitoring, the more diagnostic the dataset for identifying any significant departures from normal behavior in anticipating the possible outcomes of escalating volcano unrest.

**Volcanic crisis** An unstable situation or time of heightened public concern when the level of volcano unrest exceeds its “baseline” level, thereby increasing the prospects of possible eruption at some future, but indeterminate time. In general, during a crisis, emergency managers face a relatively narrow “decision win-

dow” in which to take timely mitigative actions to ensure public safety.

**“Decision window”** The period of time – typically weeks to months, but can be longer – during a developing volcanic crisis between the onset or escalation of volcano unrest. During this time, emergency managers are under high-stress political and public pressure to make decisions regarding mitigative actions, including possible evacuation of populations at risk.

**Warning** An official message issued by government authorities, usually given to a specific community or communities when a direct response to a volcanic threat is required. To be useful, warnings must be credible and effectively communicated – in clear, easily understandable language – with sufficient lead-time, ideally, well before the volcano unrest escalates into a volcanic crisis.

### Definition of the Subject

The hazards and risks posed by volcanic eruptions are increasing inexorably with time. This trend is the direct result of continuing exponential growth in global population and progressive encroachment of human settlement and economic development into hazardous volcanic areas. One obvious strategy in reducing volcano risk is the total abandonment of hazardous volcanic regions for human habitation. Clearly, this is utterly unrealistic; many hazardous volcanoes are located in densely populated areas, for most of which land-use patterns have been fixed by history, culture, and tradition for centuries or millennia. Moreover, people also are exposed to potential volcano hazards by simply being passengers aboard commercial airliners flying over volcanic regions and possibly encountering a drifting volcanic ash cloud from a powerful explosive eruption [5]. Thus, the only viable option in reducing volcano risk is the timely issuance of early warning of possible impending eruptions, allowing emergency-management officials to take mitigative actions. It is then imperative for volcanologists to effectively communicate the best-possible hazards information to emergency-management authorities, and to convince them to implement timely mitigative countermeasures, including evacuation, if necessary, of people at risk. However, volcanologists face daunting challenges in providing reliable, precise early warnings that government officials demand and need to make informed decisions to ensure public safety.

### Introduction

About 1500 volcanoes have erupted one or more times during the Holocene (i. e., past 10,000 years); since A.D.

1600, volcanic disasters have killed about than 300,000 people and resulted in property damage and economic loss exceeding hundreds of millions of dollars [39,40]. During recent centuries, on average 50 to 70 volcanoes worldwide are in eruption each year – about half representing activity continuing from the previous year and the other half new eruptive outbreaks [32]. Most of these erupting volcanoes are located in developing countries, which often lack sufficient scientific and economic resources to conduct adequate volcano-monitoring studies. Moreover, in any given year, many more volcanoes exhibit observable and (or) measurable anomalous behavior (*volcano unrest*) that does not lead to eruption.

For scientists and emergency-management authorities alike, the critical issues in providing early warning of possible eruption, or of possible escalation of an ongoing but weak eruption, involve basic questions such as (1) will the initiation of unrest at a currently dormant volcano culminate in eruption?; (2) if so, what would be its time of onset, eruptive mode (explosive, non-explosive, etc.) and duration?; (3) at an already restless or weakly erupting volcano, what are the prospects of escalation to major activity?; and (4) what would be the nature and severity of the associated hazards and which sectors of the volcano are the most vulnerable?

Systematic geoscience studies of a volcano’s geologic history and past behavior can provide answers to some of the above-listed and related questions relevant to *long-term* forecasts of a volcano’s present and possible future activity. In general, the fundamental premise applied in making long-term forecasts is that the volcano’s past and present behavior provides the best clues to its future behavior. However, data from volcano monitoring – showing increased activity relative to pre-crisis or “background” levels – constitute the only scientific basis for *short-term* forecasts or predictions of a possible future eruption or a significant change during an ongoing eruption. To respond effectively to a developing volcanic crisis, timely early warnings are absolutely essential, and these warnings must be reliable and precise. Improvement in the reliability and precision of warnings can be achieved only by a greatly improved capability for eruption prediction, which in turn depends on the quantity and quality of volcano-monitoring data and the diagnostic interpretation of such information [36,38,40].

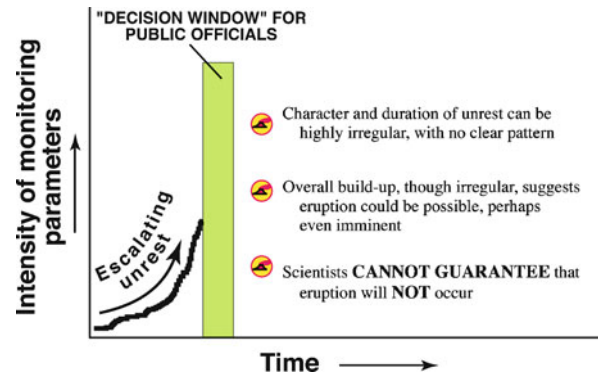
At the outset, it must be emphasized that, in this article, the term *prediction* is used in the strict sense (see Glossary) proposed by Swanson et al. [35]; thus, for a forecast or prediction to be genuine, it must be included in an official warning statement made publicly in advance of the predicted event. While retrospective forecasts and

predictions serve valid scientific purposes in testing or verifying theories or models, they are of little or no use to emergency-management authorities. It is beyond the scope of this brief paper to consider in detail volcano-monitoring studies and their applications to eruption prediction. Suffice it to say that instrumental volcano monitoring is a multi-faceted, multidisciplinary endeavor that involves the ground- and (or) satellite-based measurement of seismicity, ground deformation, gas and fluid chemistry, temperature variations, secular changes in micro-gravity and geomagnetism, etc. For more information about volcano-monitoring approaches, techniques, and instrumentation, interested readers are referred to numerous recent comprehensive reviews and references cited therein (e. g., [6,10,21,22,23,36] ► **Volcanoes, Non-linear Processes in**). Instead, this paper highlights the range in possible outcomes of volcano unrest and reviews some recent examples of the actual outcomes documented for several well-monitored volcanoes. Some implications from the observations discussed in this paper are considered in the final two parts of this article (Sect. “**Challenges in Achieving Refined Predictive Capability**” and “**Future Directions**”).

### Scope of Problem: Challenge for Emergency-Management Authorities

Because earthquake and eruption dynamics intrinsically involve non-linear processes (the focus of this volume), the attainment of a reliable capability to predict eruptions (especially explosive events) still eludes volcanologists. This dilemma persists despite notable advances in volcanology and the availability in recent decades of increasingly sophisticated models, stochastic and deterministic. Such models, which commonly involve pattern-recognition methodologies, are developed from theoretical or statistical analysis of volcano-monitoring data, eruption-occurrence patterns, or other time-space attributes for a volcanic system (e. g., [14,15,18,25,29,33,45,46] ► **Volcanic Eruptions: Stochastic Models of Occurrence Patterns**). The non-attainment of predictive capability to date reflects the reality that most volcanic systems are highly complex and that the relevant datasets for the vast majority of volcanoes are too inadequate and (or) incomplete to enable reliable and precise predictions.

Given the fact that the emerging science of eruption prediction is still nascent, scientists and emergency-management officials are at a substantial disadvantage when a long-dormant or weakly active volcano begins to exhibit unrest above its historical (i. e., “normal”) levels. A principal concern is that the heightened unrest, should it per-



Volcanic Hazards and Early Warning, Figure 1

A common scenario confronting emergency-management authorities when unrest initiates or escalates at a volcano, as documented by increasing intensity of monitoring parameters (seismicity, ground deformation, gas emission, thermal, microgravity, etc.) during a volcanic crisis. The “decision window”, though shown schematically, represents a very real and usually short timeframe during which emergency managers must decide and implement – by legal requirement, political and (or) citizen pressure, or moral conscience – strategies and actions to mitigate the potential risk from volcano hazards. (Figures 1 and 2 modified from unpublished conceptual sketches of C. Dan Miller, US Geological Survey)

sist or escalate, could culminate in renewed or greatly increased, possibly hazardous activity. Because of public-safety and socio-economic issues and consequences, emergency managers in particular must have a definitive answer to the question: What is the most likely outcome of the escalating unrest? The answer(s) to this critical question then shape(s) possible decisions they must make – within a narrow “decision window” (typically weeks to months), generally under confusing, stressful conditions – in managing the growing volcanic crisis.

Figure 1 depicts a common scenario during a volcanic crisis and lists some major factors that complicate the determination of the most likely outcome of the unrest. Unfortunately, despite advances in volcanology and volcano-monitoring techniques, scientists are unable to give emergency-management authorities the specific information they require. The greatest frustration for the authorities is that scientists cannot guarantee that an eruption will *not* occur. In the section below, I review some common possible outcomes of volcano unrest, and, in Sect. “**Some Recent Examples of Actual Outcomes of Volcano Unrest**”, some actual outcomes of recent episodes of volcano unrest.

### Variability in Possible Outcomes of Volcano Unrest

Experience worldwide for historical eruptions demonstrates a great range in possible outcomes of volcano un-

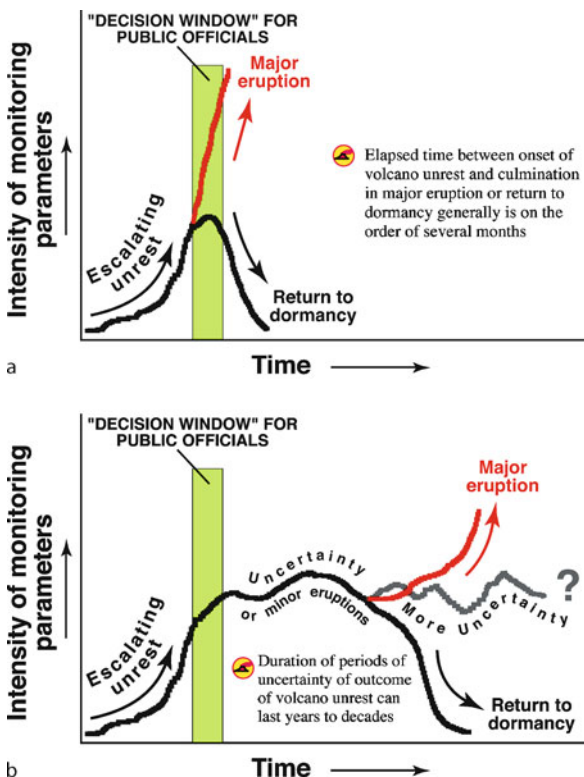


rest. The three types of outcomes discussed below – for which adequate eruptive-history and volcano-monitoring data exist – probably represent the commonest possibilities. Given the non-linear nature of eruptive phenomena, however, other scenarios can be envisaged.

### Culmination in Major Eruption or Return to Dormancy Following Short Duration of Unrest

Perhaps the simplest scenario – and certainly the most easily manageable for public officials – is when a volcano reawakens from a long dormancy (i. e., greater than a century) and begins to exhibit significant unrest. Such unrest then persists at a high level for a relatively short duration (several weeks or months) before culminating in a major eruption (Fig. 2a). With adequate volcano-monitoring studies, preferably in real time or near-real time, conducted during the precursory build-up, scientists may be able to make a short-term forecast or pre-

diction of the impending eruption. The emergency-management authorities then could have the needed guidance during the “decision window” in issuing early warnings and implementing mitigative countermeasures. Alternatively, diagnostic analysis of the volcano-monitoring data, combined with detailed knowledge of eruptive history, could lead the scientists to conclude that an eruption is not likely. In this situation, emergency managers also receive the required scientific input they need to make informed decisions to manage the volcanic crisis, including the declaration of the end of the crisis and the cancellation of any warnings in effect. In either case, scientists need sufficient information to try to determine whether an imminent eruption or a return of the volcano to dormancy is the most likely outcome. However, it must be emphasized that scientists in assessing possible outcomes can reach incorrect conclusions, even using abundant scientific data and constraints from quantitative models.



Volcanic Hazards and Early Warning, Figure 2  
**a** Two common possible outcomes of volcano unrest that are arguably the most tractable for public officials managing a volcanic crisis (see text). **b** Some other possible outcomes of volcano unrest that pose much greater challenges for scientists and public officials alike during a volcanic crisis (see text)

### Culmination in Major Eruption or Return to Dormancy Following a Long Lull in Unrest

A scenario fraught with greater uncertainty is one that, after an initial short period (weeks to months) of sustained and heightened volcano unrest, the volcano neither erupts nor quickly returns to quiescence. Instead, the volcanic system exhibits a decreased level of activity characterized by fluctuating but low-level unrest and (or) weak, non-hazardous eruptions that may persist for years or decades, before possibly exhibiting another episode of heightened unrest. Under such circumstances, scientists and public officials confront anew the crisis conditions and challenges inherent in attempting to anticipate the outcome of the renewed unrest.

This scenario is similar to that previously discussed above (Sect. “Culmination in Major Eruption or Return to Dormancy Following Short Duration of Unrest”) but differs by the complication posed by another possible outcome in addition to those of an imminent eruption or a return to dormancy: namely, the possibility that the volcano could revert to a mode of continuing uncertainty involving only low-level activity or cessation of unrest for an indeterminate period of time (Fig. 2b). The prolongation of the periods of uncertainty has some unfortunate consequences, including (1) the potential for scientists and public officials to lose credibility by incorrectly anticipating the outcomes during the pulses of heightened unrest and by issuing warnings (“false alarms”) of events that fail to materialize (see Sect. “The Dilemma of “False Alarms””); and (2) increased difficulty in maintain-

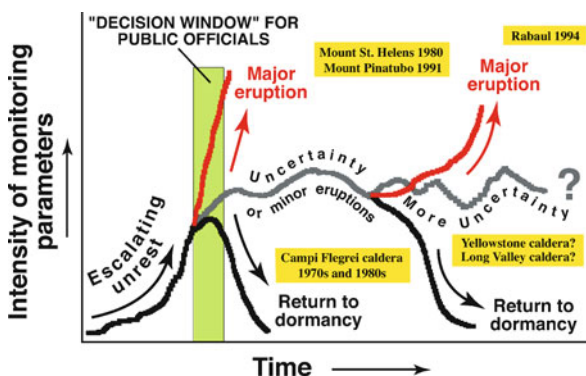
ing public interest in, and awareness of, potential hazards posed by some eventual volcanic crisis that actually culminates in eruption.

### Ongoing Unrest over Long Periods with no Clear Indication of Possible Future Activity

Some long-dormant volcanic systems undergo unrest that can persist for many years to decades, fluctuating in intensity but not showing any long-term definitive trend suggestive of possible renewed eruptive activity (Fig. 2b). It is extremely difficult, perhaps impossible, to anticipate possible outcomes of such long-duration, ongoing unrest. The reasons for this problem are simply that the eruptive histories and dynamics of long-dormant systems are poorly understood and (or) their eruption frequencies are longer than the time span of the available pre-eruption *baseline* monitoring data. Some examples of such modes of volcano unrest include: Three Sisters volcanic center, Oregon, USA [48]; Long Valley caldera, California, USA [2,12]; Yellowstone caldera, Wyoming–Montana, USA [48]; and Campi Flegrei Caldera, Italy [3,42].

### Some Recent Examples of Actual Outcomes of Volcano Unrest

The previous discussion of possible hypothetical outcomes of volcano unrest (Sect. “Variability in Possible Outcomes of Volcano Unrest”) has its basis in several recent examples of actual volcano unrest, as documented by detailed investigations. These examples are reviewed herein (Fig. 3).



Volcanic Hazards and Early Warning, Figure 3  
Some actual outcomes of recent episodes of volcano unrest at selected volcanoes. Even though these well-studied volcanoes are closely monitored in real time or near-real time by instrumental arrays, it was or is not possible to determine the outcome of the unrest at most of them (see text for discussion)

### Mount St. Helens (USA) and Mount Pinatubo (Philippines)

In 1975, on the basis of many years of detailed stratigraphic and dating studies, together with an analysis of repose intervals, scientists of the US Geological Survey (USGS) made a long-term forecast of possible renewed eruptive activity at Mount St. Helens Volcano, Washington State, which had been dormant since 1857. This forecast, albeit qualitative, was remarkably prescient, stating: “...an eruption is likely within the next hundred years, possibly before the end of this century” (see p. 441 in [7]). Five years after the forecast, phreatic eruptions began at Mount St. Helens in late March 1980 following a week of precursory seismicity. Sustained seismicity, accompanied by large-scale ground deformation of the volcano’s north flank, ultimately produced a cataclysmic magmatic eruption on 18 May 1980 (Fig. 4). This catastrophic event caused the worst volcanic disaster in the history of the United States (see [16] and chapters therein for detailed summary).

For Mount St. Helens, the volcano unrest, which was well monitored visually and instrumentally, lasted for only about three months before culminating in a major eruption (Fig. 3). Scientists monitoring Mount St. Helens were not able to make a precise prediction of the 18 May events (flank collapse, laterally directed blast, and vertical mag-



Volcanic Hazards and Early Warning, Figure 4  
19 km-high eruption column of the climactic magmatic eruption of Mount St. Helens Volcano, Washington State, USA, on 18 May 1980. This eruption occurred following about two months of volcano unrest that began after a 123-year quiet period. (Photograph by Austin Post taken about noon on 18 May 1980)

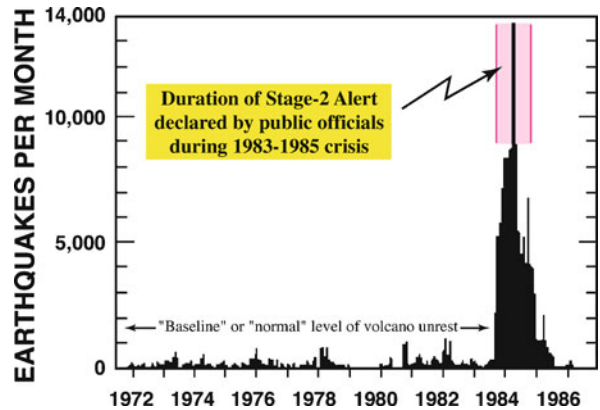
matic eruption). From analysis of the volcano-monitoring data and hazards assessments, however, they were able to give early warning and to convince emergency-management authorities to order evacuation and take other risk-reduction actions well within the “decision window” prior to the paroxysmal eruption.

Another illustrative example of a major eruptive activity following a brief duration of volcano unrest was the 1991 eruption of Mount Pinatubo, Luzon, Philippines (see [26], and chapters therein). Prior to 1991, Pinatubo had been quiescent for about 600 years with no recorded historical eruptions; it had been little studied and was not monitored. In March 1991, Pinatubo began to stir from its long sleep, as indicated by increased fumarolic activity and precursory earthquakes, some large enough to be felt by local inhabitants. Phreatic explosions began in early April, and the first magmatic activity (dome intrusion) was observed on 7 June. On the basis of the analysis and interpretation of data from volcano monitoring (initiated in late April), combined with observation of increased frequency of high (> 15 km) eruption columns, scientists of the Philippine Institute of Volcanology and Seismology (PHIVOLCS) and the USGS Volcano Disaster Assistance Program (VDAP) issued warnings and made a forecast on 12 June that a major eruption was imminent. The PHIVOLCS-USGS scientific team recommended to the Commandant of the US Clark Air Base and the Philippine emergency-management authorities to order major evacuations of the air base and other population centers around the volcano. In all, more than 300,000 people were evacuated from hazardous zones. Three days later on 15 June, Pinatubo’s climactic eruption occurred. Even though this eruption was the largest in the world since the 1912 eruption of Novarupta (Alaska, USA), the timely early warnings and evacuations saved many thousands of lives and minimized economic loss.

### Rabaul Caldera (Papua New Guinea)

The outcomes for the volcano unrest at Mount St. Helens (1980) and Pinatubo (1991) represent examples of simple (linear?) and relatively short progression (several months) from onset of precursory activity to major eruption. However, most other recent episodes of volcano unrest involve complex or uncertain outcomes; below I discuss one well-documented example of a complex outcome.

Rabaul caldera, Papua New Guinea, after being quiescent for several decades began to exhibit volcano unrest in the mid-1980s, as manifested by dramatically increased seismicity (Fig. 5) and accompanying caldera uplift of more than 1 m [20]. This intense and rapidly escalating



Volcanic Hazards and Early Warning, Figure 5  
Occurrences of earthquakes at Rabaul Caldera, Papua New Guinea, during the period 1972–1986 (modified from McKee et al. [20]). The dramatically increased seismicity during 1983–1985, together with significant ground deformation, constituted a serious volcanic crisis and prompted public officials to declare a “Stage-2 Alert” (see text). The time span bracketed by this alert (pink shading) is somewhat comparable to the “decision window” schematically shown in Figs. 1–3. However, the 1983–1985 unrest did not immediately culminate in eruption, which did occur ten years later

unrest prompted the scientists of the Rabaul Volcanological Observatory (RVO) to warn of possible imminent eruption. In response, emergency-management authorities declared in October 1983 a “Stage-2 Alert” (Fig. 5), stating that the volcano could erupt within weeks. Accordingly, the civil authorities implemented a number of mitigative measures – including widening of roads to serve as evacuation routes, preparation of contingency plans, conduct of evacuation drills, etc. – to prepare for the expected soon-approaching eruption. However, soon after the issuance of the alert, the volcano unrest, after peaking in early 1985, began to diminish sharply and returned to normal levels within 6 months (Fig. 5). No eruption occurred, and the “Stage-2 Alert” was lifted in November 1984.

The abrupt cessation of volcano unrest and the non-occurrence of the forecasted eruption were generally viewed as a needless, disruptive “false alarm”. As fallout from the unanticipated turn of events, the scientists of the RVO lost credibility with the emergency-response authorities, who in turn lost the confidence of the general population.

Ultimately however, the initially unsatisfactory outcome of the 1983–1985 unrest positively influenced the successful response to a later major eruption. After remaining at the normally low level of activity after for about ten years the mid-1980s crisis, Rabaul erupted suddenly on 19 September 1994 following only 27 hours of precur-



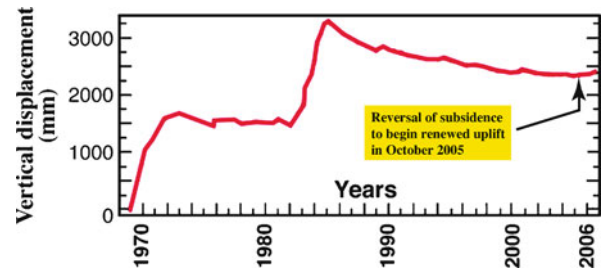
Volcanic Hazards and Early Warning, Figure 6  
View from space of the powerful explosive eruption of 19 September 1994 from Rabaul Caldera. (Photograph courtesy of astronauts aboard the NASA Shuttle)

sory seismicity [4]. This major eruption (Fig. 6) devastated nearly all of Rabaul City, but fortunately caused only 5 deaths. The lower-than-expected fatalities can in part be attributable to the fact that people largely “self-evacuated” without waiting for official advisories to do so. They apparently well remembered the lessons (e. g., enhanced awareness of volcano hazards, measures to take in case of eruption, evacuation drills) learned from the volcanic crisis ten years earlier.

### Ongoing Irregular Long-Duration Volcano Unrest but no Clear Indication of Possible Eruption

Recent volcano unrest at several calderas, which are well studied and systematically monitored in real time or near-real time by geophysical and geochemical networks, provide good examples of the difficulty in determining the most likely outcomes of unrest. Because of the abundance of historical observations and volcano-monitoring data, Campi Flegrei Caldera, Italy, located adjacent to Vesuvius Volcano in the Naples metropolitan area, is a particularly instructive example. During the early 1970s and again during 1983–1985, the caldera exhibited pronounced episodes of volcano unrest (Fig. 7) that generated scientific and public concern. Neither of these unrest episodes culminated in eruption, and each was followed by longer intervals of relatively low, fluctuating activity or net subsidence, hinting of an outcome whereby the restless volcano would return to dormancy.

From a recent study by Troise et al. [42], however, the post-1985 subsidence trend reversed in late 2005, and the



Volcanic Hazards and Early Warning, Figure 7  
Uplift and subsidence at Campi Flegrei Caldera, Italy, during the period 1969–2006 (modified from Fig. 1b in [42]). Vertical displacements measured by precise leveling, together with associated horizontal displacements and increased seismicity, indicate two episodes of caldera uplift, one during 1969–1973 and another during 1982–1985. Neither episode culminated in eruption (see text). Note that the current “baseline” level of activity is higher than that prior to 1982

caldera since has risen by a few centimeters (Fig. 7). Historical records, extending back in time for more than two millennia (see Fig. 1a in [42]), demonstrate a pattern of relatively short-duration uplifts (lasting years to decades) interspersed with long-lived gradual subsidence (lasting centuries). Of these several ups and downs of Campi Flegrei caldera over the past two millennia, only one episode of volcano unrest culminated in eruptive activity – the Monte Nuovo eruption in 1538 [9]. The historical behavior of Campi Flegrei emphasizes the challenging task of attempting to determine the outcome of volcano unrest with only a few years or decades of volcano-monitoring data.

Similarly, in recent years, ground-based and space-based geodetic data [e. g., Interferometric Synthetic Aperture Radar (InSAR) and GPS], combined with seismic monitoring and other geophysical and geochemical investigations, have documented uplift and (or) subsidence at several other well-monitored long-dormant volcanic systems (e. g., Three Sisters volcanic center, Oregon; Long Valley Caldera, California; and Yellowstone Caldera, Wyoming–Montana). For more information concerning these systems as well as other volcanic systems, the interested reader is referred to the following works: Wicks et al. [48,49]; Battaglia et al. [2]; Hill [12]; and Dzurisin [10 and chapters therein]. It is important to stress that, while the level and intensity at these restless volcanic systems have varied within the periods of systematic instrumental monitoring (only a few decades at most), the available datasets and interpretations are inadequate to determine whether or not the ongoing unrest will culminate in eruption.

### Kilauea Volcano, Hawaii (USA)

Brief mention should be made of frequently active volcanic systems – typically basaltic in composition – that are characterized by conspicuous and measurable unrest occurring nearly continuously over long time periods. Perhaps the best-studied volcanic systems exhibiting such behavior are those of Kilauea and Mauna Loa volcanoes on the Island of Hawaii (see [8] and chapters therein). These two volcanoes are among the most active in the world and have been well monitored since the establishment in 1912 of the Hawaiian Volcano Observatory, operated continuously by the USGS since 1948. Abundant “baseline” monitoring data, particularly for Kilauea, demonstrate well-defined patterns and outcomes of unrest (e. g., Fig. 5 in [41]; Fig. 8 in [37]). However, it must be emphasized that the behavior and outcomes of unrest observed for Kilauea – culminating either in subsurface intrusion only or in eruption – occur over much shorter time scales than those for the examples of explosive volcanoes discussed in this paper (e. g., Figs. 2, 3, and 7).

For Kilauea, because of its high eruption frequency and, at times, nearly continuous activity (e. g., [11]), the outcomes of its volcano unrest and the need to issue early warnings are manageable for the civil authorities with little or no anxiety. This fortunate situation reflects a positive circumstance: because of the volcano’s frequent and, at times, continuous eruptive activity, the dissemination of hazards information and early warnings is easily and routinely accomplished among the scientists, emergency-management officials, news media, and the affected public. Moreover, analyses of data for well-monitored, frequently active volcanoes, such as Kilauea, provide many case histories that can serve to test and refine volcano-monitoring techniques as well as eruption-forecasting methodologies. Nonetheless, it should be noted that eruptions at Kilauea and other basaltic volcanoes differ fundamentally from those of more explosive – hence, potentially more dangerous – stratovolcanoes that erupt more viscous materials of intermediate composition (e. g., Mount St. Helens, Mount Pinatubo).

### Challenges in Achieving Refined Predictive Capability

The discussion in the preceding sections paints a somewhat discouraging picture: namely, possibly with very rare exceptions (e. g., the Mount St. Helens and Pinatubo case histories), to date scientists still lack a reliable capability for precisely *and* accurately predicting the outcomes of unrest at explosive volcanoes with low eruption frequency. This unsatisfactory state of affairs does not well serve so-

cietal needs to reduce the risks of volcanic hazards and demands improvement. Considered below are some elements that I deem to be essential to achieve refined predictive capability and early warnings of possible eruptive activity.

### More Geologic Mapping and Dating Studies of Volcanoes

Of the world’s active and potentially active volcanoes, only a handful (mostly in the developed countries) have been studied sufficiently for the detailed and complete reconstruction of their past behavior, eruption frequencies, and associated hazards. Specifically, geologic mapping (at a scale of 1:25,000 or better) and dating studies of stratigraphically well-controlled eruptive products have been done for only a few volcanoes. Recorded historical eruptions for most volcanoes – especially those in the New World – rarely extend back in time for more than five centuries. Their eruptive histories generally are too short to serve as a comprehensive time-series dataset for quantitative statistical analysis of eruption frequency or repose intervals. Thus, it is essential to expand the time-series dataset by obtaining radiometric ages of prehistoric eruptive products. Moreover, in studies of eruption frequency, no matter how abundant the data used, the calculated frequencies obtained are necessarily *minima*, because the products of small eruptive events are lost to erosion and not preserved in the geologic record.

### More Volcano Monitoring at More Volcanoes: Importance of Establishing Long-Term “Baseline” Behavior

As noted by Tilling (see p. 9 in [40]), most volcanologists believe that “... if a volcano is monitored extensively in real- or near-real time by well-deployed instrumental networks, it should be possible to make much better forecasts of the outcomes of volcano unrest.” Unfortunately, many of the world’s active and potentially active volcanoes are poorly monitored, and some are not monitored at all. Thus, to advance our understanding of eruptive dynamics in order to make more precise early warnings, initiating or expanding volcano monitoring at many more volcanoes is imperative. Ideally, volcano monitoring should be conducted in real- or near-real time, and it should involve a combination of monitoring techniques rather than reliance on any single one [36]. Moreover, the longer the time span of volcano monitoring, the more diagnostic is the relevant database needed for detecting and interpreting any significant departure from the overall variation in the long-term “baseline” behavior of the volcano.

### The Dilemma of “False Alarms”

With the current state-of-the-art in volcanology and monitoring techniques, it is difficult, and generally impossible, for volcanologists to determine with any certainty: (1) the most likely outcome of onset of unrest at a long-dormant volcano; and (2) the eventual possible occurrence of a major eruption in the course of long-duration, but relatively low-level ongoing unrest. Nonetheless, scientists still have an obligation to emergency managers and the threatened populations to provide the best-possible hazards information, forecasts or predictions, and warnings within the limitations of available knowledge, technology, and volcano-monitoring techniques. In so doing, the scientists run the unavoidable risk of raising a “false alarm” and the associated criticism and loss of credibility when the forecasted event fails to materialize. On the other hand, emergency-management officials as well as the general public must be prepared to accept the disruptive socio-economic consequences and costs of occasional “false alarms”. The Rabaul case history (previously discussed in Sect. “[Rabaul Caldera \(Papua New Guinea\)](#)”) underscores the sentiment of Banks et al. (see p. 78 in [1]): “... False alarms themselves can provide, through objective assessment of the scientific and public response to a volcanic crisis that ended without eruption, valuable lessons useful ... for the next crisis, which could culminate in an eruption.” Even so, every attempt should be made to minimize “false alarms” to the extent possible, in order to maintain scientific credibility and the confidence of the emergency-management authorities and the populace at risk.

### Future Directions

An obvious critical need is the comprehensive geoscience studies of more volcanoes, to reconstruct their prehistoric eruptive histories, thereby extending the time-series database needed for quantitative analysis and interpretation. In addition, volcano monitoring, even if rudimentary – deploying only 1 to 3 seismometers and making simple field observations and measurements [34] – is needed for many more volcanoes, both to establish “baselines” and to provide early warning. Obviously, future research necessarily should also be focused on methodologies to make more robust the acquisition and processing of real- or near-real-time volcano-monitoring data, using or refining *existing* instrumentation and techniques, volcano databases, analytical tools, and empirical and theoretical models. An especially promising avenue of future research would require the wider utilization of broadband instruments in seismic-monitoring networks at active and potentially volcanoes. This would enable systematic and

quantitative analysis and interpretation of Long-Period (LP) and Very-Long-Period (VLP) seismicity, which from worldwide experience precedes and accompany nearly all eruptions ► [Volcanoes, Non-linear Processes in](#).

A long-standing and vexing problem for volcanologists is that, at present, no geophysical or geochemical criteria exist to identify the distinguishing characteristics between volcano unrest that merely ends in a subsurface intrusion of magma and unrest that culminates with magma breaking the surface to produce an eruption. From the vantage point of the emergency-management officials, subsurface intrusion or movement of magma, while of great academic interest to volcanologists in deciphering volcano dynamics, poses no threat to public safety. In contrast, the threat of an imminent eruption requires timely decisions and mitigative actions to ensure public safety. Thus, the future development, if possible, of quantitative criteria to distinguish the precursory pattern of a subsurface intrusion from that for an eruption would mark a quantum leap in the young science of volcanology. In this regard, an improved understanding of magma intrusion vs. eruptive dynamics might provide scientists more powerful tools in choosing between the various paths shown in Fig. 3 as the most likely outcome of volcano unrest.

In attempting to quantify estimates of the probabilities of specified hazardous volcanic events within specified timeframes (i.e., eruption predictions), scientists have made increasing use of “materials-failure” models (e.g., [25,45,46]), “event trees” (e.g., [27]), and more statistically rigorous variants of the “event-tree”, “decision-tree”, “pattern-recognition”, and “occurrence-pattern” approaches (e.g., [17,18,19,24,29] ► [Volcanic Eruptions: Stochastic Models of Occurrence Patterns](#)). Depending on the particular circumstances of the volcano(es) and eruptions under study, these and other statistical tools have their individual strengths and weaknesses. Taken together, however, the utility and potential successful application of such approaches are inherently hampered by the incompleteness and (or) too-short time span of the eruptive-history or volcano-monitoring datasets analyzed. Thus, to become more robust, the probabilistic methodologies must employ larger and more complete datasets. In any case, while useful, the existing techniques or models cannot yield results with the precision and accuracy that emergency-managers demand and need as they confront a “decision window” during a volcanic crisis. Simply stated, even if the calculated results prove to be accurate, their error bars generally exceed the time span bracketed by the “decision window”.

In recent years, the World Organization of Volcano Observatories (WOVO) – a Commission of the In-

ternational Association of Volcanology and Chemistry of the Earth's Interior – has launched an ambitious but poorly funded effort, called *WOVOdat*, to construct a global database of empirical data on volcanic unrest. This database is being developed as a user-friendly Internet-based system with query and analysis tools [30]. Even though a basic design and schema now exist for the construction of *WOVOdat* [44], the actual compilation and integration of data are just barely beginning. The *WOVOdat* project recently received major funding from the Singapore Government and Nanyang Technological University [28]. This favorable development brightens the prospects for it ultimately to become a powerful tool in evaluating patterns and possible outcomes of volcanic unrest. Nonetheless, until *WOVOdat* becomes fully operational, we must continue to rely on the available datasets – however imperfect – and to employ and refine the existing analytical methodologies and statistical models.

### Acknowledgments

This article has benefited from constructive reviews and helpful suggestions by L. J. Patrick Muffler and Fred Klein (both of the US Geological Survey, Menlo Park) on an earlier draft. To them, I offer them my sincere thanks. The views expressed in this article have been shaped by my personal involvement in responses to several of the volcanic crises in recent decades, and by enlightening and instructive interactions and discussions with many colleagues in the global volcanological community.

### Bibliography

#### Primary Literature

1. Banks NG, Tilling RI, Harlow DH, Ewert JW (1989) Volcano monitoring and short-term forecasts. In: Tilling RI (ed) *Short Courses in Geology*, vol 1. Volcanic Hazards, American Geophysical Union, Washington, DC, chap 4, pp 51–80
2. Battaglia M, Roberts C, Segall P (2003) The mechanics of unrest at Long Valley caldera, California: 2. Constraining the nature of the source using geodetic and micro-gravity data. *J Volcanol Geotherm Res* 127:219–245
3. Berrino G, Corrado G, Luongo G, Toro B (1984) Ground deformation and gravity changes accompanying the 1982 Pozzuoli uplift. *Bull Volcanol* 47:187–200
4. Blong R, McKee C (1995) *The Rabaul eruption 1994: Destruction of a Town*. Natural Hazards Research Centre, Macquarie University, Sydney, 52 pp
5. Casadevall TJ (ed) (1994) *Volcanic ash and aviation safety: Proceedings of the First International Symposium on Volcanic Ash and Aviation Safety*. US Geological Survey Bulletin 2047. Government Printing Office, Washington, 450 pp
6. Chouet B (2004) Volcano seismology. *Pure Appl Geophys* 160:739–788
7. Crandell DR, Mullineaux DR, Rubin M (1975) Mount St. Helens volcano: Recent and future behaviour. *Science* 187:438–441
8. Decker RW, Wright TL, Stauffer PH (eds) (1987) *Volcanism in Hawaii*. US Geological Survey Professional Paper 1350. vol 1 and 2. U.S. Government Printing Office, Washington, 1667 pp
9. De Vito M, Lirer L, Mastrolorenzo G, Rolandi G (1987) The 1538 Monte Nuovo eruption (Campi Flegrei, Italy). *Bull Volcanol* 49:608–615
10. Dzurisin D (ed) (2006) *Volcano deformation: Geodetic Monitoring Techniques*. Springer-Praxis, Berlin, 441 pp
11. Heliker C, Swanson DA, Takahashi TJ (eds) (2003) *The Pu'u O'o-Kupaianaha Eruption of Kilauea Volcano, Hawaii: The first 20 years*. US Geological Survey Professional Paper 1676. U.S. Geological Survey, Reston, 206 pp
12. Hill DP (2006) *Unrest in Long Valley Caldera, California, 1978–2004*. *Geol Soc London* 269:1–24
13. ISDR (2004) *Living with Risk: International Strategy for Disaster Reduction (ISDR)*. United Nations, New York and Geneva vol 1, 431 pp and vol 2, 126 pp
14. Klein F (1982) Patterns of historical eruptions at Hawaiian volcanoes. *J Volcanol Geotherm Res* 12:1–35
15. Klein F (1984) Eruption forecasting at Kilauea Volcano, Hawaii. *J Geophys Res* 89(B5):3059–3073
16. Lipman PW, Mullineaux DR (eds) (1981) *The 1980 eruptions of Mount St. Helens, Washington*. US Geological Survey Professional Paper 1250. U.S. Government Printing Office, Washington, 844 pp
17. Marzocchi W (1996) Chaos and stochasticity in volcanic eruptions the case of Mount Etna and Vesuvius. *J Volcanol Geotherm Res* 70:205–212
18. Marzocchi W, Sandri L, Gasparini P, Newhall C, Boschi E (2004) Quantifying probabilities of volcanic events: The example of volcanic hazards at Mt. Vesuvius. *J Geophys Res* 109:B11201. doi:10.1029/2004JB003155
19. Marzocchi W, Sandri L, Selva J (2008) BET\_EF: A probabilistic tool for long- and short-term eruption forecasting. *Bull Volcanol* 70(5):623–632. <http://dx.doi.org/>
20. McKee CO, Johnson RW, Lowenstein PL, Riley SJ, Blong RJ, de St. Ours P, Talai B (1985) Rabaul caldera, Papua New Guinea: Volcanic hazards, surveillance, and eruption contingency planning. *J Volcanol Geotherm Res* 23:195–237
21. McNutt SR (1996) Seismic monitoring and eruption forecasting of volcanoes: A review of the state of the art and case histories. In: Scarpa R, Tilling RI (eds) *Monitoring and Mitigation of Volcanic Hazards*. Springer, Heidelberg, pp 99–146
22. McNutt SR (2000) Seismic monitoring. In: Sigurdsson H, Houghton B, McNutt SR, Rymer H, Stix J (eds) *Encyclopedia of Volcanoes*. Academic Press, San Diego, chap 68, pp 1095–1119
23. McNutt SR (2000) Synthesis of volcano monitoring. In: Sigurdsson H, Houghton B, McNutt SR, Rymer H, Stix J (eds) *Encyclopedia of Volcanoes*. Academic Press, San Diego, chap 71, pp 1167–1185
24. Mulargia F, Gasparini P, Marzocchi W (1991) Pattern recognition applied to volcanic activity: Identification of the precursory patterns to Etna recent flank eruptions and periods of rest. *J Volcanol Geotherm Res* 45:187–196
25. Murray JB, Ramirez Ruiz JJ (2002) Long-term predictions of the time of eruptions using remote distance measurement at Volcán de Colima, México. *J Volcanol Geotherm Res* 117(1–2): 79–89

26. Newhall CG, Punongbayan RS (eds) (1996) Fire and Mud: Eruptions and Lahars of the Mount Pinatubo, Philippines. Philippine Institute of Volcanology and Seismology, Quezon City, and University of Washington Press, Seattle, 1126 pp
27. Newhall CG, Hoblitt RP (2002) Constructing event trees for volcanic crises. *Bull Volcanol* 64:3–20
28. Newhall CG (2008) written communication, 11 February 2008
29. Sandri L, Marzocchi W, Zaccarelli L (2004) A new perspective in identifying the precursory patterns of eruptions. *Bull Volcanol* 66:263–275
30. Schwandner FM, Newhall CG (2005) *WOVOdat*: The World Organization of Volcano Observatories database of volcanic unrest. European Geosciences Union, Geophysical Research Abstracts, vol 7, abstract # 05-J-09267 (extended abstract), 2 pp
31. Sigurdsson H, Houghton B, McNutt SR, Rymer H, Stix J (eds) (2000) *Encyclopedia of Volcanoes* (and chapters therein). Academic Press, San Diego, 1417 pp
32. Simkin T, Siebert L (1994) *Volcanoes of the World: A Regional Directory, Gazetteer, and Chronology of Volcanism During the Last 10,000 Years*, 2nd edn. Smithsonian Institution, Washington and Geoscience Press, Inc., Tucson, Arizona, 349 pp
33. Sparks RSJ (2003) *Frontiers: Forecasting volcanic eruptions*. *Earth Planet Sci Lett* 210(1–2):1–15
34. Swanson DA (1992) The importance of field observations for monitoring volcanoes, and the approach of “Keeping Monitoring as Simple as Practical”. In: Ewert JW, Swanson DA (eds) *Monitoring volcanoes: Techniques and Strategies used by the staff of the Cascades Volcano Observatory, 1980–90*. US Geological Survey Bulletin, vol 1966. U.S. Government Printing Office, Washington, 219–223
35. Swanson DA, Casadevall TJ, Dzurisin D, Holcomb RT, Newhall CG, Malone SD, Weaver CS (1985) Forecasts and predictions of eruptive activity at Mount St. Helens, USA: 1975–1984. *J Geodyn* 3:397–423
36. Tilling RI (1995) The role of monitoring in forecasting volcanic events. In: McGuire WJ, Kilburn CRJ, Murray JB (eds) *Monitoring Active Volcanoes: Strategies, Procedures, and Techniques*. UCL Press, London, pp 369–402
37. Tilling RI (2002) Volcanic Hazards. In: Meyer RA (ed) *Encyclopedia of Physical Science and Technology*, vol 17, 3rd edn. Academic Press, San Diego, pp 559–577
38. Tilling RI (2003) Volcano monitoring and eruption warnings. In: Zschau J, Küppers AN (eds) *Early Warning Systems for Natural Disaster Reduction*. Springer, Berlin, chap 5, pp 505–510
39. Tilling RI (2005) Volcano Hazards. In: Martí J, Ernst G (eds) *Volcanoes and the Environment*. Cambridge University Press, Cambridge, chap 2, pp 56–89
40. Tilling RI (2008) The critical role of volcano monitoring in risk reduction. *Adv Geosci* 14:3–11
41. Tilling RI, Dvorak JJ (1993) The anatomy of a basaltic volcano. *Nature* 363:125–133
42. Troise C, De Natale G, Pingue F, Obrizzo F, De Martino P, Tammaro U, Boschi E (2007) Renewed ground uplift at Campi Flegrei caldera (Italy): New insight on magmatic processes and forecast. *Geophys Res Lett* 34:L03301. doi:10.1029/2006GL28545
43. Turner MB, Cronin SJ, Bebbington MS, Platz T (2008) Developing probabilistic eruption forecasts for dormant volcanoes: A case study from Mt. Taranaki, New Zealand. *Bull Volcanol* 70:507–515
44. Venezky DY, Newhall CG (2007) *WOVOdat* design document: The schema, table descriptions, and create table statements for the database of worldwide volcanic unrest (*WOVOdat* version 1.0). US Geological Survey Open-File Report 2007-1117, 184 pp
45. Voight B (1988) A method for prediction of volcanic eruptions. *Nature* 332:125–130
46. Voight B, Cornelius RR (1991) Prospects for eruption prediction in near real-time. *Nature* 350:695–698
47. WG99, Working Group on California Earthquake Probabilities (1999) Earthquake probabilities in the San Francisco Bay region: 2000 to 2030 – A summary of findings. US Geological Survey Open-File Report 99-517, 60 pp
48. Wicks C, Dzurisin D, Ingebritsen S, Thatcher W, Lu Z, Iverson RM (2002) Magmatic activity beneath the quiescent Three Sisters volcanic center, central Oregon Cascade Range, USA. *Geophys Res Lett* 29(7):26–751,26–754. doi:10.1029/2001GL014205
49. Wicks C, Thatcher W, Dzurisin D, Svarc J (2006) Uplift, thermal unrest, and magmatic intrusion at Yellowstone Caldera. *Nature* 440:72–75

### Books and Reviews

- Chester D (1993) *Volcanoes and Society*. Edward Arnold (a Division of Hodder & Stoughton), London, 351 pp
- Martí J, Ernst G (eds) (2005) *Volcanoes and the Environment*. Cambridge University Press, Cambridge, 471 pp
- Scarpa R, Tilling RI (eds) (1996) *Monitoring and Mitigation of Volcano Hazards*. Springer, Heidelberg, 841 pp



# Volcano Seismic Signals, Source Quantification of

HIROYUKI KUMAGAI<sup>1,2</sup>

<sup>1</sup> National Research Institute for Earth Science and Disaster Prevention, Tsukuba, Japan

<sup>2</sup> IAVCEI/IASPEI Joint Commission on Volcano Seismology, Tsukuba, Japan

## Article Outline

Glossary

Definition of the Subject

Introduction

Phenomenological Representation of Seismic Sources

Waveform Inversion

Spectral Analysis

Fluid-Solid Interactions

Future Directions

Acknowledgment

Appendix A: Green's Functions

Appendix B: Moment Tensor for a Spherical Source

Appendix C: Moment Tensor for a Cylindrical Source

Bibliography

## Glossary

**Moment tensor** A point seismic source representation defined by the first-order moment of the equivalent body force or the stress glut. Slip on a fault as well as volumetric changes such as an isotropic expansion and tensile crack can be represented by the moment tensor.

**Waveform inversion** An approach to estimate source mechanisms and locations of seismic events by finding the best fits between observed and synthesized seismograms.

**Autoregressive equation** A difference form of the equation of motion of a linear dynamic system, which is a basic equation to determine the complex frequencies (frequencies and Q factors) of decaying harmonic oscillations in observed signals.

**Crack wave** A dispersive wave generated by fluid-solid interactions in a crack. The phase velocity of the crack wave is smaller than the acoustic velocity of the fluid in the crack.

## Definition of the Subject

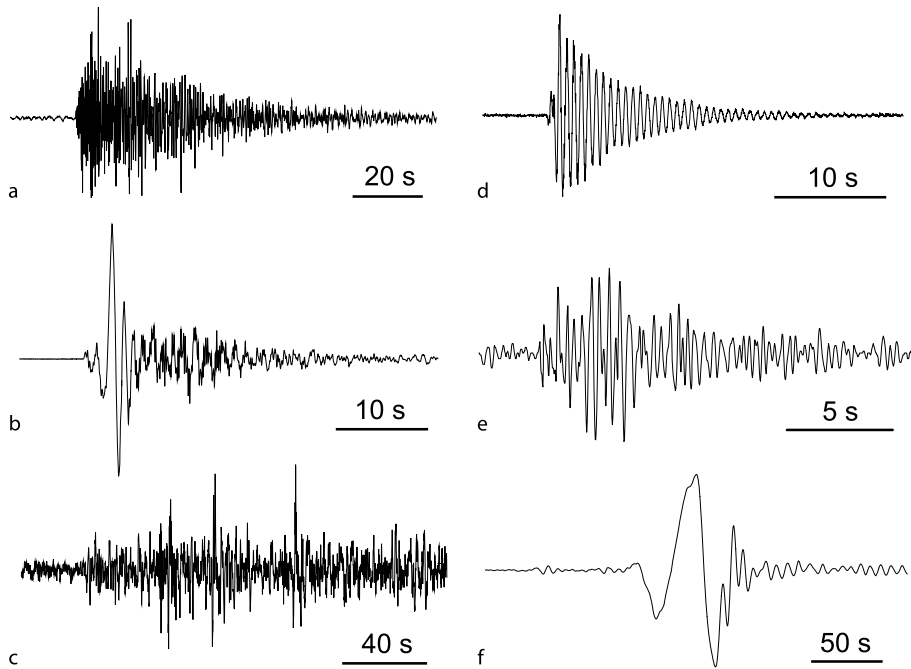
Volcano seismicity produces a wide variety of seismic signals that provide glimpses of the internal dynamics of volcanic systems. Quantitative approaches to analyze and in-

terpret volcano-seismic signals have been developed since the late 1970s. The availability of seismic equipments with wide frequency and dynamic ranges since the early 1990s revealed a further variety of volcano-seismic signals in oscillation periods longer than a few seconds. Quantification of the sources of volcano-seismic signals is crucial to achieving a better understanding of the physical states and dynamics of magmatic and hydrothermal systems.

## Introduction

Volcanoes occur in tectonically active regions of the earth where magmatic and hydrothermal fluids display complex interactions with volcanic rocks and the atmosphere. Volcano seismicity is the manifestation of such complex interactions occurring inside of volcanic edifices. Volcano seismology aims at achieving a better understanding of the physical states and dynamics of magmatic and hydrothermal systems beneath active volcanoes through observation, analysis, and interpretation of volcano-seismic signals. It is well known that volcano-seismic signals have a wide variety of signatures (Fig. 1), which have been classified by various schemes e. g., [25,61,80,84]. The volcano-tectonic (VT) earthquake, low-frequency (LF) or long-period (LP) event, tremor, very-long-period (VLP) event, hybrid event, and explosion earthquake have been traditionally used to specify volcano-seismic signals. Here, LF denotes frequencies between 0.5 and 5 Hz [61], and LP and VLP denote periods between 0.2 and 2 s and longer than a few seconds, respectively [25].

In contrast to the various classification schemes, source processes of volcano-seismic signals may be classified into three main families: (i) brittle failure in volcanic rock triggered by fluid movement, (ii) transient pressure disturbance caused by mass transport and/or volumetric change, and (iii) resonance of a fluid-filled resonator excited by a pressure disturbance. These three processes may be linked to the signals as follows: (i) VT earthquake, (ii) VLP event and explosion earthquake, (iii) LF or LP event, tremor, and VLP event, and (i)+(iii) hybrid event. VT earthquakes, sometimes called high-frequency (HF) earthquakes, are indistinguishable from ordinary tectonic earthquakes, but occur in volcanic regions due to process (i). LF or LP events have signals characterized by decaying oscillations, whereas the signature of tremor shows continuous oscillations. These oscillatory signatures can be interpreted as process (iii), in which the excitations are impulsive for LF or LP events and successive for tremor. Most VLP events are characterized by impulsive signatures, which may be generated by process (ii). However, some VLP events show sustained oscillations or decay-



Volcano Seismic Signals, Source Quantification of, Figure 1

Various volcano-seismic signals: **a** volcano-tectonic earthquake that occurred between Miyakejima and Kozujima, Japan; **b** explosion earthquake observed at Asama volcano, Japan; **c** tremor that occurred beneath Mt. Fuji, Japan; **d** long-period event observed at Kusatsu-Shirane volcano, Japan; **e** long-period event observed at Guagua Pichincha volcano, Ecuador; **f** very-long-period event that occurred beneath Miyakejima, Japan (low-pass filtered at 20 s)

ing harmonic oscillations, which may reflect process (iii). Hybrid events show mixed characters of both VT earthquakes and LP events, suggesting that processes (i) and (iii) both take place at the source. Explosion earthquakes occur in association with eruptions in which process (ii) is dominant.

Recently, there have been remarkable advances in quantitative understanding of the sources of volcano-seismic signals. These advances were made possible by the development of source models and analysis techniques coupled with increased computer capacity and the availability of seismic observation equipment with wide frequency and dynamic ranges. Major studies that have contributed to these advances are summarized as follows: Detailed magma intrusion processes imaged by VT earthquakes e. g., [19,43,46,86,111,112,123]; Quantitative interpretations of oscillations in LP and VLP events and tremor to diagnose the states of fluids at the sources of these signals e. g., [2,7,21,22,23,24,28,29,38,39,40,41,42,44,52,53,65,66,67,70,71,73,77,78,85,87,88,90,91,92,93,94,95,98,134]; Tracking the sources of LP events and tremor e. g., [3,4,5,13,30,45,82,113,114]; Waveform analysis of

impulsive and oscillatory signatures in VLP events and explosion earthquakes to investigate source dynamics e. g., [6,8,9,10,11,31,32,33,34,36,47,48,49,51,54,55,56,57,58,59,60,64,69,72,76,97,102,103,105,107,108,110,120,122,126,132,133]. These volcano seismological studies have been reviewed by various authors [25,26,27,61,80,81,96].

The purpose of this article is to provide a systematic presentation of the theoretical basis for quantification of the sources of volcano-seismic signals suitable for readers, including graduate students and young researchers, who are new to the study of volcano seismology. I focus on four subjects: (1) phenomenological representation of seismic sources, (2) waveform inversion to estimate source mechanisms, (3) spectral analysis based on an autoregressive model, and (4) physical properties of fluid-solid coupled waves. Some additional notes are provided in Appendixes. These constitute the basic elements needed to quantify the sources of volcano-seismic signals. Applications of quantitative approaches are not fully reviewed in this manuscript. Readers are encouraged to consult both the original research papers and the review papers cited above.

## Phenomenological Representation of Seismic Sources

We first derive the phenomenological representation of seismic sources using single force and moment tensor. This representation provides the basis to study source processes of volcano seismicity. I follow the approach presented by Backus and Mulcahy [12] (hereafter referred to as BM) and Aki and Richards [1] for the moment tensor representation. The theory elaborated by Takei and Kumazawa [118,119] is used to introduce single force in seismic sources.

### Stress Glut

We consider indigenous seismic sources occurring within the closed earth system. This means that external bodies and detached mass are excluded. Let us assume an isotropic medium for the earth, which is represented by density  $\rho$  and Lamé's constants  $\lambda$  and  $\mu$ . The exact equation of motion of the earth is given by

$$\rho \frac{\partial^2 u_i}{\partial t^2} = \frac{\partial \sigma_{ni}^{\text{true}}}{\partial x_n}, \quad (1)$$

where  $u_i$  and  $\sigma_{ni}^{\text{true}}$  are the true displacement field and true stress tensor, respectively, actually produced in the earth. These are functions of space  $\mathbf{x} = (x_1, x_2, x_3)$  and time  $t$ . The summation convention for repeated subscripts is used throughout this paper unless otherwise stated. In BM's theory, the failure of the elastic stress-strain relationship in source processes, such as slip across a plane that can not be described by the linear elastodynamics, is represented by an equivalent body force. We use a model stress  $\sigma_{ni}^{\text{model}}$  given by the elastic stress-strain relationship as

$$\sigma_{ni}^{\text{model}} = \lambda \frac{\partial u_k}{\partial x_k} \delta_{ni} + \mu \left( \frac{\partial u_n}{\partial x_i} + \frac{\partial u_i}{\partial x_n} \right), \quad (2)$$

where  $\delta_{ni}$  is the Kronecker symbol ( $\delta_{ni} = 0$  for  $n \neq i$  and  $\delta_{ni} = 1$  for  $n = i$ ). By replacing the true stress by the model stress, Eq. (1) can be rewritten as follows:

$$\rho \frac{\partial^2 u_i}{\partial t^2} = \frac{\partial \sigma_{ni}^{\text{model}}}{\partial x_n} + f_i^S, \quad (3)$$

where  $f_i^S$  is the equivalent body force defined as

$$f_i^S = \frac{\partial}{\partial x_n} (\sigma_{ni}^{\text{true}} - \sigma_{ni}^{\text{model}}). \quad (4)$$

The difference between  $\sigma_{ni}^{\text{true}}$  and  $\sigma_{ni}^{\text{model}}$  is called the stress glut. The equivalent body force, which is the spatial derivative of the stress glut, describes the source and vanishes

outside the source region for motion in a completely elastic medium.

We introduce Green's function  $G_{ij}(\mathbf{x}, t; \boldsymbol{\eta}, \tau)$ , which is the  $i$ th component of displacement at  $(\mathbf{x}, t)$  excited by the unit impulse applied at  $\mathbf{x} = \boldsymbol{\eta}$  and  $t = \tau$  in the  $j$ -direction.  $G_{ij}(\mathbf{x}, t; \boldsymbol{\eta}, \tau)$  satisfies the following equation:

$$\rho \frac{\partial^2 G_{ij}}{\partial t^2} = \frac{\partial}{\partial x_n} \left\{ \lambda \frac{\partial G_{kj}}{\partial x_k} \delta_{ni} + \mu \left( \frac{\partial G_{nj}}{\partial x_i} + \frac{\partial G_{ij}}{\partial x_n} \right) \right\} + \delta(\mathbf{x} - \boldsymbol{\eta}) \delta(t - \tau) \delta_{ij}. \quad (5)$$

Note that  $G_{ij}(\mathbf{x}, t; \boldsymbol{\eta}, \tau) = G_{ij}(\mathbf{x}, t - \tau; \boldsymbol{\eta}, 0)$ . The displacement  $u_i$  can then be described by using the equivalent body force and Green's functions as follows ("Appendix A: Green's Functions"):

$$u_i(\mathbf{x}, t) = \int_{-\infty}^{\infty} \iiint_V f_j^S(\boldsymbol{\eta}, \tau) G_{ij}(\mathbf{x}, t - \tau; \boldsymbol{\eta}, 0) dV(\boldsymbol{\eta}) d\tau. \quad (6)$$

BM expanded  $f_j^S$  in terms of polynomial moments at a particular point  $\boldsymbol{\xi}_0$  as

$$\begin{aligned} f_j^S(\boldsymbol{\eta}, \tau) &= F_j^{S0}(\boldsymbol{\xi}_0, \tau) \delta(\boldsymbol{\eta} - \boldsymbol{\xi}_0) \\ &\quad + F_{jk}^{S1}(\boldsymbol{\xi}_0, \tau) \frac{\partial \delta(\boldsymbol{\eta} - \boldsymbol{\xi}_0)}{\partial \eta_k} + \dots \\ &= f_j^{S0}(\boldsymbol{\eta}, \tau) + f_j^{S1}(\boldsymbol{\eta}, \tau) + \dots \end{aligned} \quad (7)$$

This expansion is schematically shown in Fig. 2, which may help the reader to understand that  $F_j^{S0}$  and  $F_{jk}^{S1}$  have the dimensions of force and moment, respectively.

### Single Force

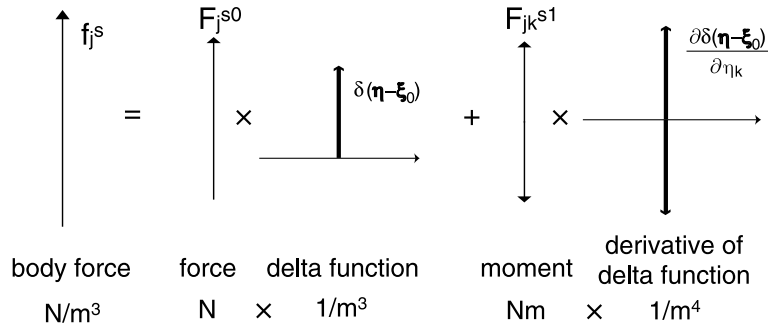
$F_j^{S0}$  is given as

$$F_j^{S0}(\boldsymbol{\xi}_0, \tau) = \iiint_V f_j^S(\boldsymbol{\eta}, \tau) dV(\boldsymbol{\eta}), \quad (8)$$

which is the total force. This must be zero because of the conservation of linear momentum (no force appears or disappears in the earth). Therefore,

$$\iiint_V f_j^S(\boldsymbol{\eta}, \tau) dV(\boldsymbol{\eta}) = 0. \quad (9)$$

This states that no single force exists in any indigenous seismic sources in the scheme presented by BM. However, we will see later that a single force exists in indigenous sources if the non-linear effect of mass advection is taken into account.



Volcano Seismic Signals, Source Quantification of, Figure 2  
Schematic diagram of the expansion of the equivalent body force  $f_j^s$  in terms of polynomial moments

**Moment Tensor**

We next examine the equivalent body force arising from the first-order moment,

$$f_j^{s1}(\eta, \tau) = F_{jk}^{s1}(\xi_0, \tau) \frac{\partial \delta(\eta - \xi_0)}{\partial \eta_k}. \tag{10}$$

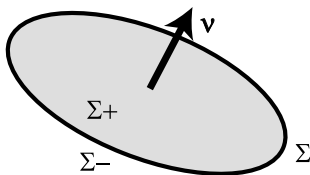
If we consider a displacement discontinuity across a surface  $\Sigma$ ,  $f_j^{s1}$  may be expressed as [1]

$$f_j^{s1}(\eta, \tau) = - \iint_{\Sigma} m_{jk}(\xi, \tau) \frac{\partial}{\partial \eta_k} \delta(\eta - \xi) d\Sigma(\xi), \tag{11}$$

where  $m_{jk}$  is the moment density tensor given as

$$m_{jk} = \lambda v_l [u_l] \delta_{jk} + \mu (v_j [u_k] + v_k [u_j]). \tag{12}$$

Here,  $[u]$  is the displacement discontinuity defined by  $[u] = u|_{\Sigma+} - u|_{\Sigma-}$  and  $v = (v_1, v_2, v_3)$  is the unit vector normal to  $\Sigma$ , where  $u|_{\Sigma+}$  and  $u|_{\Sigma-}$  are displacements on the  $\Sigma+$  and the  $\Sigma-$  sides of  $\Sigma$ , respectively (see Fig. 3). An explicit expression of  $m_{jk}$  is



Volcano Seismic Signals, Source Quantification of, Figure 3  
An internal surface  $\Sigma$  across which a displacement discontinuity occurs. The displacement discontinuity is denoted by  $[u] = u|_{\Sigma+} - u|_{\Sigma-}$ , where  $u|_{\Sigma+}$  and  $u|_{\Sigma-}$  are displacements on the  $\Sigma+$  and  $\Sigma-$  sides of  $\Sigma$ , respectively [1]

$$m = \begin{pmatrix} \lambda v_k [u_k] + 2\mu v_1 [u_1] & \mu (v_1 [u_2] + v_2 [u_1]) \\ \mu (v_2 [u_1] + v_1 [u_2]) & \lambda v_k [u_k] + 2\mu v_2 [u_2] \\ \mu (v_3 [u_1] + v_1 [u_3]) & \mu (v_3 [u_2] + v_2 [u_3]) \\ & \mu (v_1 [u_3] + v_3 [u_1]) \\ & \mu (v_2 [u_3] + v_3 [u_2]) \\ & \lambda v_k [u_k] + 2\mu v_3 [u_3] \end{pmatrix}. \tag{13}$$

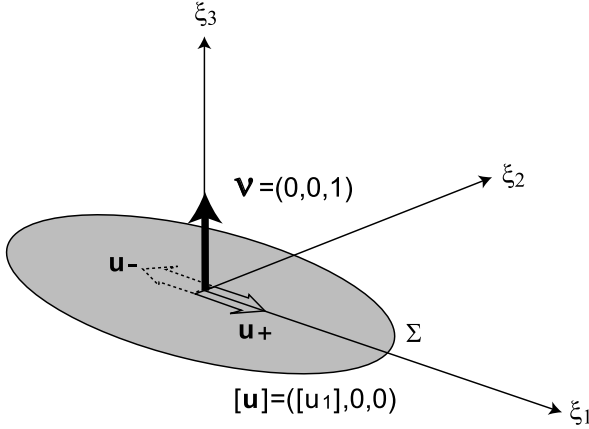
Therefore  $m_{jk} = m_{kj}$ . Using Green's functions, the displacement field due to  $f_j^{s1}$  is expressed as

$$\begin{aligned} u_i^{s1}(\mathbf{x}, t) &= \int_{-\infty}^{\infty} \iiint_V f_j^{s1}(\eta, \tau) G_{ij}(\mathbf{x}, t - \tau; \eta, 0) dV(\eta) d\tau \\ &= \int_{-\infty}^{\infty} \iiint_V \left[ - \iint_{\Sigma} m_{jk}(\xi, \tau) \frac{\partial}{\partial \eta_k} \delta(\eta - \xi) d\Sigma(\xi) \right] \\ &\quad \cdot G_{ij}(\mathbf{x}, t - \tau; \eta, 0) dV(\eta) d\tau \\ &= \int_{-\infty}^{\infty} \iint_{\Sigma} \left[ m_{jk}(\xi, \tau) \iiint_V - \frac{\partial}{\partial \eta_k} \delta(\eta - \xi) G_{ij}(\mathbf{x}, t - \tau; \eta, 0) dV(\eta) \right] d\Sigma(\xi) d\tau \\ &= \int_{-\infty}^{\infty} \iint_{\Sigma} m_{jk}(\xi, \tau) \frac{\partial}{\partial \xi_k} G_{ij}(\mathbf{x}, t - \tau; \xi, 0) d\Sigma(\xi) d\tau, \end{aligned} \tag{14}$$

where I used the following property:

$$\begin{aligned} \iiint_V \frac{\partial}{\partial \eta_k} \delta(\eta - \xi) G_{ij}(\mathbf{x}, t - \tau; \eta, 0) dV(\eta) \\ = - \frac{\partial}{\partial \xi_k} G_{ij}(\mathbf{x}, t - \tau; \xi, 0). \end{aligned} \tag{15}$$

We assume that the source is small compared to the wavelengths of seismic waves and approximate the source as



Volcano Seismic Signals, Source Quantification of, Figure 4  
Slip  $[u_1]$  in the plane  $\xi_3 = 0$

a point at  $\xi_0$ . Then, Eq. (14) can be written as

$$u_i^{S1}(\mathbf{x}, t) = \int_{-\infty}^{\infty} M_{jk}(\xi_0, \tau) \frac{\partial}{\partial \xi_k} G_{ij}(\mathbf{x}, t - \tau; \xi_0, 0) d\tau, \quad (16)$$

where  $M_{jk}$  is the moment tensor defined as

$$M_{jk}(\xi_0, \tau) = \iint_{\Sigma} m_{jk}(\xi, \tau) d\Sigma(\xi). \quad (17)$$

**Slip on a Fault** We consider slip  $[u_1]$  in the plane  $\xi_3 = 0$  (see Fig. 4). In this case,  $\mathbf{v} = (0, 0, 1)$  and  $[\mathbf{u}] = ([u_1], 0, 0)$ . Substituting  $\mathbf{v}$  and  $[\mathbf{u}]$  into Eq. (13), we obtain the moment density tensor  $m_{jk}$  as

$$\mathbf{m} = \begin{pmatrix} 0 & 0 & \mu[u_1(\xi, \tau)] \\ 0 & 0 & 0 \\ \mu[u_1(\xi, \tau)] & 0 & 0 \end{pmatrix}, \quad (18)$$

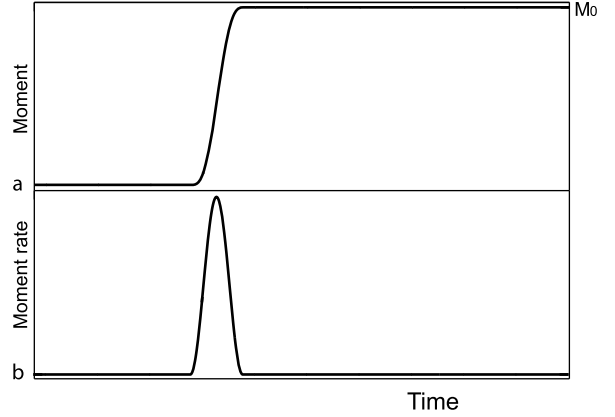
and the moment tensor  $M_{jk}$  as

$$\mathbf{M} = M(\xi_0, \tau) \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad (19)$$

where

$$M(\xi_0, \tau) = \mu \iint_{\Sigma} [u_1(\xi, \tau)] d\Sigma(\xi). \quad (20)$$

This is called the moment function or source-time function. The time history of slip across a plane is typically represented by a step-like function, and therefore  $M(\xi_0, \tau)$  displays a step-like function (Fig. 5a). We can also define the time derivative of  $M$  as



Volcano Seismic Signals, Source Quantification of, Figure 5  
Moment and moment-rate functions for slip across a plane in the form of a step-like function

$$\dot{M}(\xi_0, \tau) = \mu \frac{\partial}{\partial \tau} \iint_{\Sigma} [u_1(\xi, \tau)] d\Sigma(\xi). \quad (21)$$

This is called the moment-rate function, which is an impulsive function (Fig. 5b). After slip stops,  $M$  has a finite value  $M_0$  (Fig. 5a), which is given as

$$M_0 = \mu A \bar{u}_1. \quad (22)$$

$A$  and  $\bar{u}_1$  are the slip area and average slip, respectively, which are defined as

$$A = \iint_{\Sigma} d\Sigma, \quad (23)$$

$$\bar{u}_1 = \frac{\iint_{\Sigma} [u_1(\xi, \tau = \infty)] d\Sigma}{A}, \quad (24)$$

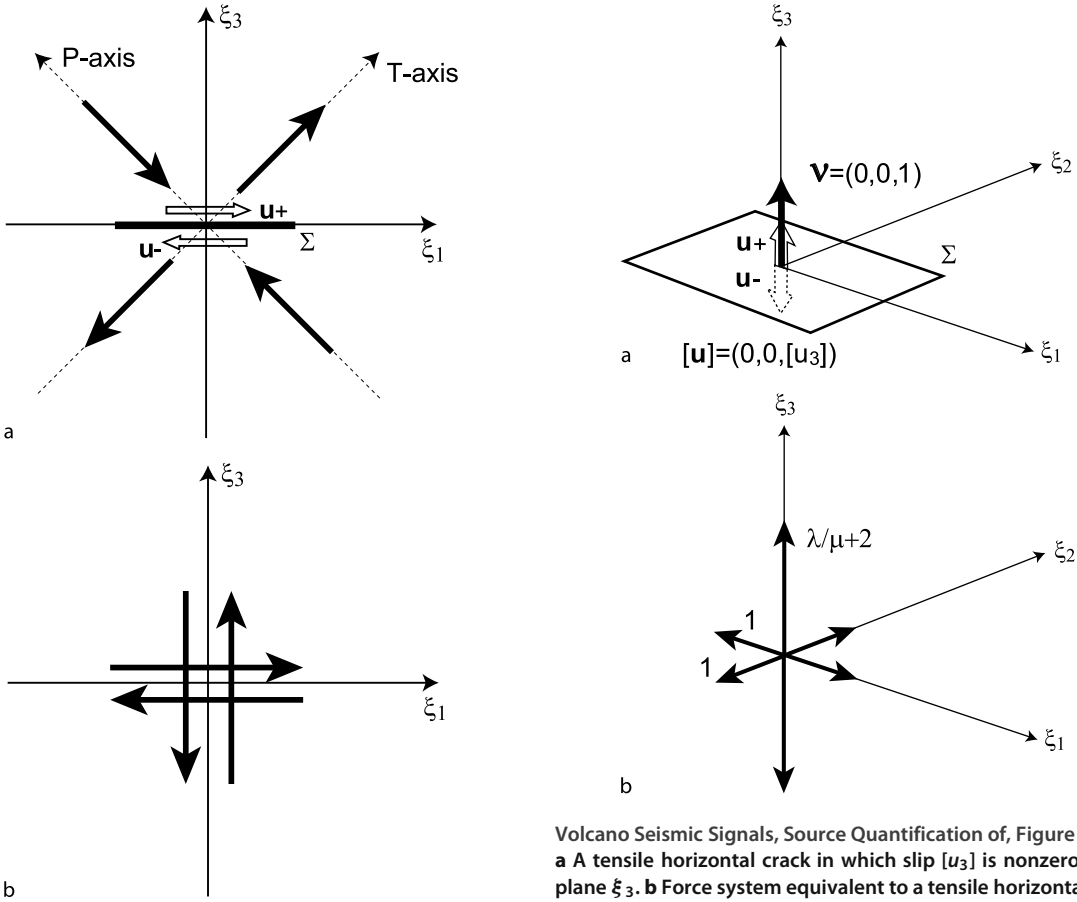
where  $\tau = \infty$  represents time after the termination of slip.  $M_0$  is called the seismic moment. When  $\tau = \infty$ ,  $M_{jk}$  becomes

$$\mathbf{M} = M_0 \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}. \quad (25)$$

This matrix can be diagonalized to

$$\mathbf{M}' = M_0 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix}, \quad (26)$$

where three eigenvalues (1, 0, and  $-1$ ) correspond to the T (maximum compression), B (null), and P (minimum compression) axes, respectively. As shown in Fig. 6, this



Volcano Seismic Signals, Source Quantification of, Figure 6  
**a** Force system for slip  $[u_1]$  in the plane  $\xi_3 = 0$ . **b** A double couple equivalent to the force system in Fig. 6a

system constitutes a double couple, which is equivalent to fault slip.

**Tensile Crack** We examine a crack such that slip  $[u_3]$  is nonzero in the plane  $\xi_3 = 0$  (a horizontal crack; see Fig. 7). In this case,  $\mathbf{v} = (0, 0, 1)$  and  $[\mathbf{u}] = (0, 0, [u_3])$ . Therefore, we find from Eq. (13) that

$$\mathbf{m} = \begin{pmatrix} \lambda[u_3(\xi, \tau)] & 0 & 0 \\ 0 & \lambda[u_3(\xi, \tau)] & 0 \\ 0 & 0 & (\lambda + 2\mu)[u_3(\xi, \tau)] \end{pmatrix}. \tag{27}$$

The moment tensor for a horizontal crack is given as

$$\mathbf{M} = M(\xi_0, \tau) \begin{pmatrix} \lambda/\mu & 0 & 0 \\ 0 & \lambda/\mu & 0 \\ 0 & 0 & \lambda/\mu + 2 \end{pmatrix}, \tag{28}$$

and

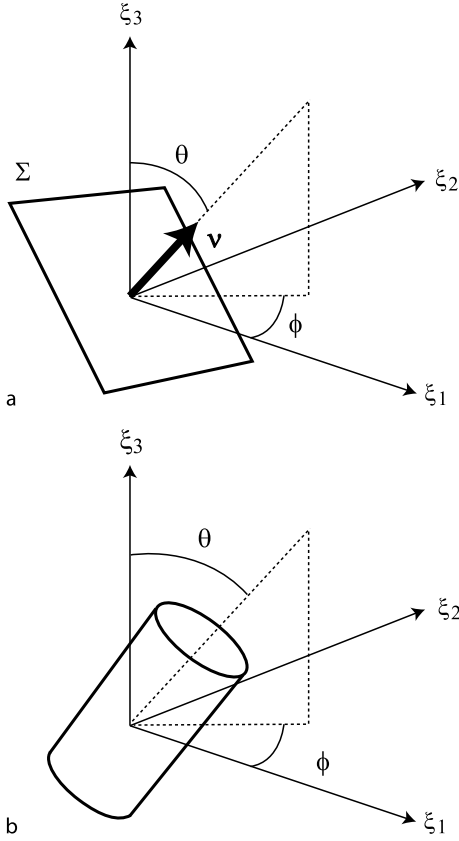
$$\begin{aligned} M(\xi_0, \tau) &= \mu \iint_{\Sigma} [u_3(\xi, \tau)] d\Sigma(\xi) \\ &= \mu A \bar{u}_3(\tau) \\ &= \mu \Delta V(\tau), \end{aligned} \tag{29}$$

where  $\Delta V$  represents the incremental volume change of the crack.

Figure 8a is a tensile crack with normal direction  $(\theta, \phi)$ , and we find that  $\mathbf{v} = (\sin \theta \cos \phi, \sin \theta \sin \phi, \cos \theta)$  and  $[\mathbf{u}] = [u](\sin \theta \cos \phi, \sin \theta \sin \phi, \cos \theta)$ . Using Eqs. (13) and (17), we obtain the moment tensor for the crack:

$$\mathbf{M} = \mu \Delta V \begin{pmatrix} \lambda/\mu + 2 \sin^2 \theta \cos^2 \phi & & \\ 2 \sin^2 \theta \sin \phi \cos \phi & & \\ 2 \sin \theta \cos \theta \cos \phi & & \\ 2 \sin^2 \theta \sin \phi \cos \phi & 2 \sin \theta \cos \theta \cos \phi & \\ \lambda/\mu + 2 \sin^2 \theta \sin^2 \phi & 2 \sin \theta \cos \theta \sin \phi & \\ 2 \sin \theta \cos \theta \sin \phi & \lambda/\mu + 2 \cos^2 \theta & \end{pmatrix}, \tag{30}$$

where  $\Delta V = \iint_{\Sigma} [u] d\Sigma$ .



Volcano Seismic Signals, Source Quantification of, Figure 8  
Source coordinates and geometries for **a** a crack and **b** a cylinder

For a vertical crack ( $\theta = \pi/2$ ),

$$\mathbf{M} = \mu \Delta V \begin{pmatrix} \lambda/\mu + 2 \cos^2 \phi & 2 \sin \phi \cos \phi & 0 \\ 2 \sin \phi \cos \phi & \lambda/\mu + 2 \sin^2 \phi & 0 \\ 0 & 0 & \lambda/\mu \end{pmatrix}. \quad (31)$$

A tensile crack is equivalent to a superposition of three vector dipoles with amplitude ratios 1:1:( $\lambda + 2\mu$ )/ $\lambda$  (Fig. 7). If  $\lambda = \mu$ , the amplitude ratios are represented by 1:1:3. If  $\lambda = 2\mu$ , which may be appropriate either for volcanic rock near liquidus temperature or for a highly heterogeneous medium, the ratios become 1:1:2.

**Spherical Source** The moment tensor for a spherical source is given as

$$\mathbf{M} = (\lambda + 2\mu) \Delta V_s \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad (32)$$

where  $\Delta V_s = 4\pi R^2 \Delta_s$  is the volume change caused by the displacement  $\Delta_s$  of a spherical surface of ra-

dius  $R$  [89] (see “Appendix B: Moment Tensor for a Spherical Source”). The moment tensor for a spherical source may be useful to describe the source of explosion earthquakes e. g., [56].

**Cylindrical Source** The moment tensor for a vertical cylinder is

$$\mathbf{M} = \frac{\lambda + 2\mu}{\lambda + \mu} \mu \Delta V_c \begin{pmatrix} \lambda/\mu + 1 & 0 & 0 \\ 0 & \lambda/\mu + 1 & 0 \\ 0 & 0 & \lambda/\mu \end{pmatrix}, \quad (33)$$

where  $\Delta V_c = 2\pi R L \Delta_c$  is the volume change caused by the radial displacement  $\Delta_c$  of the surface of a cylinder of radius  $R$  and length  $L$  [89,126] (see “Appendix C: Moment Tensor for a Cylindrical Source”).

The moment tensor for a cylinder with axis orientation angles  $(\theta, \phi)$  (Fig. 8b) is given as

$$\mathbf{M} = \frac{\lambda + 2\mu}{\lambda + \mu} \mu \Delta V_c \times \begin{pmatrix} \lambda/\mu + (\cos^2 \theta \cos^2 \phi + \sin^2 \phi) & & \\ \sin^2 \theta \sin \phi \cos \phi & & \\ \sin \theta \cos \theta \cos \phi & & \\ \sin^2 \theta \sin \phi \cos \phi & \sin \theta \cos \theta \cos \phi & \\ \lambda/\mu + (\cos^2 \theta \sin^2 \phi + \cos^2 \phi) & \sin \theta \cos \theta \sin \phi & \\ \sin \theta \cos \theta \sin \phi & \lambda/\mu + \sin^2 \theta & \end{pmatrix} \quad (34)$$

(see “Appendix C: Moment Tensor for a Cylindrical Source”). Radial volumetric changes of a pipe may occur in association with mass transport in a cylindrical conduit e. g., [126].

### Inertial Glut

In the scheme presented by BM, no single force exists in indigenous seismic sources because of the requirement for conservation of linear momentum. However, Takei and Kumazawa [118] presented a theoretical justification for the existence of a single force in indigenous sources, if the non-linear effect of mass advection is taken into account.

Following the theory of Takei and Kumazawa [118], we consider the exact equation of motion of the earth, given as

$$\rho^{\text{true}} \frac{D^2 u_i}{Dt^2} = \frac{\partial \sigma_{ni}^{\text{true}}}{\partial x_n}, \quad (35)$$

where  $\rho^{\text{true}}$  is true density as a function of both space and time.  $\frac{D}{Dt}$  is the particle derivative:

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + v_k \frac{\partial}{\partial x_k}, \quad (36)$$

where  $v_k = \frac{Du_k}{Dt}$  is particle velocity. Using model density  $\rho^{\text{model}}$  and model stress  $\sigma_{ni}^{\text{model}}$  from Eq. (2), Eq. (35) can be rewritten as follows:

$$\rho^{\text{model}} \frac{\partial^2 u_i}{\partial t^2} = \frac{\partial \sigma_{ni}^{\text{model}}}{\partial x_n} + f_i^V. \quad (37)$$

This equation is similar to Eq. (3), but the equivalent body force  $f_i^V$  is defined in this case as

$$\begin{aligned} f_i^V &= \left( \frac{\partial \sigma_{ni}^{\text{true}}}{\partial x_n} - \frac{\partial \sigma_{ni}^{\text{model}}}{\partial x_n} \right) \\ &+ \left( \rho^{\text{model}} \frac{\partial^2 u_i}{\partial t^2} - \rho^{\text{true}} \frac{D^2 u_i}{Dt^2} \right) \\ &= f_i^S + f_i^I. \end{aligned} \quad (38)$$

Here,  $f_i^V$  is  $f_i^S$ , arising from the stress glut, plus  $f_i^I$ , representing the difference between the inertia terms, which is called the inertial glut [118].

In the same way as presented by BM, we expand  $f_i^I$  in terms of polynomial moments at a particular point  $\xi_0$  as

$$f_i^I(\eta, \tau) = F_i^{I0}(\xi_0, \tau) \delta(\eta - \xi_0) + \text{higher-order terms}, \quad (39)$$

where

$$\begin{aligned} F_i^{I0} &= \iiint_V f_i^I(\eta, \tau) dV(\eta) \\ &= - \iiint_V \left( \rho^{\text{true}} \frac{D^2 u_i}{Dt^2} - \rho^{\text{model}} \frac{\partial^2 u_i}{\partial t^2} \right) dV(\eta). \end{aligned} \quad (40)$$

By using mass conservation

$$\frac{\partial \rho^{\text{true}}}{\partial t} + \frac{\partial}{\partial x_k} (\rho^{\text{true}} v_k) = 0 \quad (41)$$

and Gauss' theorem

$$\iiint_V \frac{\partial v_k}{\partial x_k} dV = \iint_{S_e} v_k v_k dS_e, \quad (42)$$

the first integral of Eq. (40) is expressed as

$$\begin{aligned} \iiint_V \rho^{\text{true}} \frac{D^2 u_i}{Dt^2} dV &= \frac{\partial}{\partial t} \iiint_V \rho^{\text{true}} v_i dV \\ &+ \iint_{S_e} (\rho^{\text{true}} v_i v_k) v_k dS_e, \end{aligned} \quad (43)$$

where  $S_e$  is the surface of the earth. The second term in the right-hand side of Eq. (43) vanishes because it is a surface integral of a non-linear term, which is infinitesimal. Therefore, we find that

$$\begin{aligned} \iiint_V \rho^{\text{true}} \frac{D^2 u_i}{Dt^2} dV &= \frac{\partial}{\partial t} \iiint_V \rho^{\text{true}} v_i dV \\ &= \dot{L}_i^t. \end{aligned} \quad (44)$$

Here,  $\dot{L}_i^t$  is the time derivative of the total linear momentum, which is zero at any given time because of conservation of linear momentum of the earth. The second integral in Eq. (40) is given as

$$\begin{aligned} \iiint_V \rho^{\text{model}} \frac{\partial^2 u_i}{\partial t^2} dV &= \frac{\partial}{\partial t} \iiint_V \rho^{\text{model}} \frac{\partial u_i}{\partial t} dV \\ &= \dot{L}_i^m, \end{aligned} \quad (45)$$

which is the time derivative of linear momentum defined by the model density. Therefore, Eq. (40) is written as

$$\begin{aligned} F_i^{I0} &= -(\dot{L}_i^t - \dot{L}_i^m) \\ &= \dot{L}_i^m, \end{aligned} \quad (46)$$

which shows that force  $F_i^{I0}$  exists as an apparent change of the total linear momentum of the earth referred to a biased model density structure.

Takei and Kumazawa [118] stated that  $F_i^{I0}$  originates from (a) the difference of density structure of the prescribed model from the actual value in the source region before the event, and from (b) a temporal change of the density structure in the source region caused by finite displacement of mass during the event. The latter is a non-linear effect caused by mass advection, which may occur by fluid flow, especially in volcanic regions.

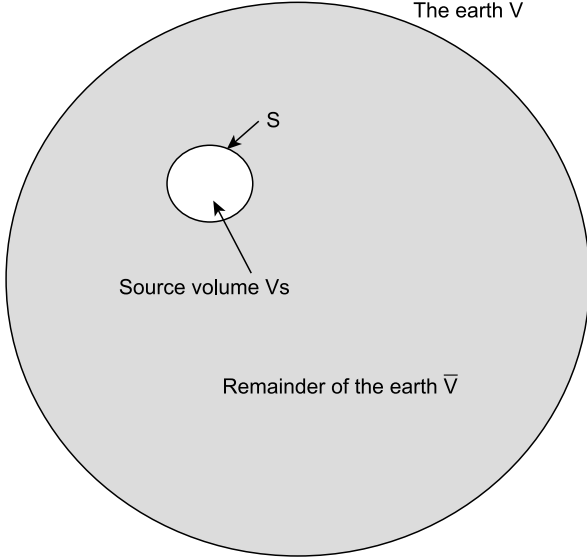
### Single Force Defined by Action and Reaction

We further consider the single force from the viewpoint of the interaction force between the source volume and the rest of the earth. Following Takei and Kumazawa [119], we define the source region  $V_s$  enclosed by a surface  $S$  and the remaining part of the earth  $\bar{V}$  (Fig. 9). Using Eqs. (38), (9), (44), and (45), we obtain

$$\begin{aligned} \iiint_V f_i^V dV &= \iiint_V (f_i^S + f_i^I) dV \\ &= \iiint_V f_i^I dV \\ &= -\frac{\partial}{\partial t} \iiint_V \left( \rho^{\text{true}} v_i - \rho^{\text{model}} \frac{\partial u_i}{\partial t} \right) dV \\ &= -\frac{\partial}{\partial t} \left[ \iiint_{V_s} \left( \rho^{\text{true}} v_i - \rho^{\text{model}} \frac{\partial u_i}{\partial t} \right) dV \right. \\ &\quad \left. - \iiint_{\bar{V}} \left( \rho^{\text{true}} v_i - \rho^{\text{model}} \frac{\partial u_i}{\partial t} \right) dV \right], \end{aligned} \quad (47)$$

where infinitesimal surface integral terms are omitted. The second integral in the right-hand side of Eq. (47) vanishes because  $\rho^{\text{true}} = \rho^{\text{model}}$  and  $v_i = \partial u_i / \partial t$  in  $\bar{V}$ .  $f_i^V$  exists





Volcano Seismic Signals, Source Quantification of, Figure 9  
The earth  $V$  divided into a source volume  $V_s$  enclosed by a surface  $S$  and the remainder of the earth  $\bar{V}$

only in the source region, and thus the volume integral of  $f_i^V$  in the left hand side of Eq. (47) can be limited to  $V_s$ . Therefore, Eq. (47) is written as

$$\iiint_{V_s} f_i^V dV = -\frac{\partial}{\partial t} \iiint_{V_s} \left( \rho^{\text{true}} v_i - \rho^{\text{model}} \frac{\partial u_i}{\partial t} \right) dV. \quad (48)$$

This equation is then rewritten as

$$\begin{aligned} \frac{\partial}{\partial t} \iiint_{V_s} \rho^{\text{true}} v_i dV &= \frac{\partial}{\partial t} \iiint_{V_s} \rho^{\text{model}} \frac{\partial u_i}{\partial t} dV \\ &\quad - \iiint_{V_s} f_i^V dV \\ &= \iiint_{V_s} \left( \rho^{\text{model}} \frac{\partial^2 u_i}{\partial t^2} - f_i^V \right) dV. \end{aligned} \quad (49)$$

From Eq. (37), we find that the integrand of the right-hand side of Eq. (49) equals  $\partial \sigma_{ni}^{\text{model}} / \partial x_n$ . Therefore, we obtain

$$\begin{aligned} \frac{\partial}{\partial t} \iiint_{V_s} \rho^{\text{true}} v_i dV &= \iiint_{V_s} \frac{\partial \sigma_{ni}^{\text{model}}}{\partial x_n} dV \\ &= \iint_S \sigma_{ni}^{\text{model}} v_n dS. \end{aligned} \quad (50)$$

This indicates that a temporal change of linear momentum in the source region is equivalent to the total force exerted on the boundary  $S$ . Therefore,  $\iint_S \sigma_{ni}^{\text{model}} v_n dS$  can be regarded as the action force originating from non-linear

processes in the source volume  $V_s$ . In  $\bar{V}$ , the reaction force opposite to  $\iint_S \sigma_{ni}^{\text{model}} v_n dS$  acts through the boundary  $S$ . We define traction  $P_i$  on  $S$ , which satisfies the following condition:

$$\sigma_{ni}^{\text{model}} v_n = -P_i \quad \text{on } S. \quad (51)$$

Then, we find that

$$\begin{aligned} \iint_S P_i dS &= -\iint_S \sigma_{ni}^{\text{model}} v_n dS \\ &= -\frac{\partial}{\partial t} \iiint_{V_s} \rho^{\text{true}} v_i dV, \end{aligned} \quad (52)$$

which is the reaction force opposite to the action force.

If the source is small compared to the wavelengths of seismic waves, the action force originating in a closed small source volume  $V_s$  may not be observable outside the source volume  $\bar{V}$ . Instead, the traction  $P_j$  exerted on  $S$  may be viewed in  $\bar{V}$  as a traction discontinuity on  $S$ . As shown by Aki and Richards [1], a traction discontinuity is the source of seismic waves represented by an equivalent body force. We consider seismic waves in  $\bar{V}$ , and the equation of motion in  $\bar{V}$  may be expressed as

$$\rho^{\text{model}} \frac{\partial^2 u_i}{\partial t^2} = \frac{\partial \sigma_{ni}^{\text{model}}}{\partial x_n} + f_i \quad \text{in } \bar{V}, \quad (53)$$

where  $f_i$  is the equivalent body force due to the traction discontinuity on  $S$ , given as

$$f_i(\boldsymbol{\eta}, \tau) = \iint_S P_i(\boldsymbol{\xi}, \tau) \delta(\boldsymbol{\eta} - \boldsymbol{\xi}) dS(\boldsymbol{\xi}). \quad (54)$$

Note that

$$\iiint_{\bar{V}} f_i(\boldsymbol{\eta}, \tau) dV(\boldsymbol{\eta}) = \iint_S P_i(\boldsymbol{\xi}, \tau) dS(\boldsymbol{\xi}), \quad (55)$$

which is the total force acting through  $S$  in  $\bar{V}$ .

We may use Green's functions estimated from the model structure because the effect of the true structure in the source volume is negligible in the point source approximation. Therefore, by using Eq. (54), the displacement field  $u_i^f$  due to the single force may be described as

$$\begin{aligned} u_i^f(\mathbf{x}, t) &= \int_{-\infty}^{\infty} \iiint_{\bar{V}} f_j(\boldsymbol{\eta}, \tau) G_{ij}(\mathbf{x}, t - \tau; \boldsymbol{\eta}, 0) dV(\boldsymbol{\eta}) d\tau \\ &= \int_{-\infty}^{\infty} \iiint_{\bar{V}} \left[ \iint_S P_j(\boldsymbol{\xi}, \tau) \delta(\boldsymbol{\eta} - \boldsymbol{\xi}) dS(\boldsymbol{\xi}) \right] \\ &\quad \cdot G_{ij}(\mathbf{x}, t - \tau; \boldsymbol{\eta}, 0) dV(\boldsymbol{\eta}) d\tau \\ &= \int_{-\infty}^{\infty} \iint_S P_j(\boldsymbol{\xi}, \tau) G_{ij}(\mathbf{x}, t - \tau; \boldsymbol{\xi}, 0) dS(\boldsymbol{\xi}) d\tau, \end{aligned} \quad (56)$$

or, equivalently,

$$u_i^f(\mathbf{x}, t) = \int_{-\infty}^{\infty} F_j(\xi_0, \tau) G_{ij}(\mathbf{x}, t - \tau; \xi_0, 0) d\tau, \quad (57)$$

where

$$\begin{aligned} F_j(\xi_0, \tau) &= \iint_S P_j(\xi, \tau) dS(\xi) \\ &= -\frac{\partial}{\partial t} \iiint_{V_s} \rho^{\text{true}} v_i dV \\ &= -\dot{L}_j^s. \end{aligned} \quad (58)$$

Here,  $L_j^s$  is the total linear momentum inside the source volume. This single force representation in terms of the traction  $P_j$  is consistent with the relationship between the traction modes and equivalent body force given by Takei and Kumazawa [119] in small-source and long-wavelength approximations.

$F_j$  is the reaction force opposite to the action force generated by non-linear processes in the source volume. A simple example is dense materials falling in a magma chamber (Fig. 10a). During the acceleration stage of falling dense materials,  $\dot{L}_j^s$  is a downward force and we therefore observe an upward single force as the reaction force; during the deceleration stage, we observe a downward single force (Fig. 10a). In the case of lighter materials ascending

in a magma chamber, we observe a downward single force during the acceleration stage of ascending lighter materials, and an upward single force is observed during the deceleration stage of ascending lighter materials (Fig. 10b).

**Summary**

Here, we summarize the phenomenological representation of indigenous seismic sources. For point sources, the displacement field  $u_i$  can be expressed by the superposition of the displacement fields excited by single force  $u_i^f$  of Eq. (57) and moment tensor  $u_i^{S1}$  of Eq. (16):

$$\begin{aligned} u_i(\mathbf{x}, t) &= \int_{-\infty}^{\infty} F_j(\xi_0, \tau) G_{ij}(\mathbf{x}, t - \tau; \xi_0, 0) d\tau \\ &+ \int_{-\infty}^{\infty} M_{jk}(\xi_0, \tau) \frac{\partial}{\partial \xi_k} G_{ij}(\mathbf{x}, t - \tau; \xi_0, 0) d\tau, \end{aligned} \quad (59)$$

where

$$F_j(\xi_0, \tau) = \iint_S P_j(\xi, \tau) dS(\xi) \quad (60)$$

and

$$M_{jk}(\xi_0, \tau) = \iint_{\Sigma} m_{jk}(\xi, \tau) d\Sigma(\xi). \quad (61)$$

Here,  $P_j$  is the traction exerted on the boundary  $S$  between the source volume and the rest of the earth and  $m_{jk}$  is the moment density tensor defined by the displacement discontinuity across the surface  $\Sigma$ .  $F_j$  is related to the total linear momentum within the source volume  $L_j^s$  as

$$F_j = -\dot{L}_j^s \quad (j = 1, 2, 3). \quad (62)$$

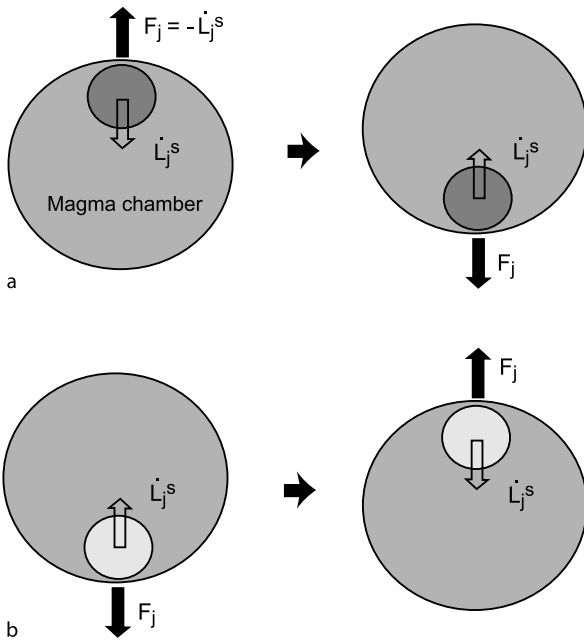
$M_{jk}$  is the moment tensor consisting of six independent components:

$$\mathbf{M} = \begin{pmatrix} M_{11} & M_{12} & M_{13} \\ M_{12} & M_{22} & M_{23} \\ M_{13} & M_{23} & M_{33} \end{pmatrix}. \quad (63)$$

**Waveform Inversion**

Waveform inversion is a quantitative tool used to estimate the source mechanisms and locations of volcano-seismic signals. The source parameters are determined by finding the best fits between observed and synthetic seismic waveforms. This can be performed by solving an inverse problem by using the concept of linear inverse theory e. g., [83].

We may reasonably assume that the source of a tectonic earthquake is represented by the moment tensor corresponding to a double couple. However, volcano-seismic



Volcano Seismic Signals, Source Quantification of, Figure 10 Schematic illustration of the reaction force ( $F_j$ ) opposite to the action force ( $\dot{L}_j^s$ ) generated by nonlinear processes in a source volume: a dense materials falling in a magma chamber and b light materials ascending in a magma chamber

signals may be generated by complex source processes, requiring a general source representation consisting of single force and moment tensor. Furthermore, while we can assume a step-like function for the source-time function of a tectonic earthquake, we must determine the complex time history at the source of volcano-seismic signals, which involves more unknowns. Waveform inversion of volcano-seismic signals is thus a more complicated task than the inversion of a tectonic earthquake.

We assume a single force and moment tensor at a particular source location, and obtain from Eq. (59) the following relationship:

$$u_i(t) = \int_{-\infty}^{\infty} [F_j(\tau)G_{ij}(t-\tau) + M_{jk}(\tau)G_{ij,k}(t-\tau)] d\tau, \tag{64}$$

where  $G_{ij,k} = \partial G_{ij} / \partial \xi_k$ . We rewrite Eq. (64) in the following form:

$$u_i(t) = \sum_{j=1}^9 \int_{-\infty}^{\infty} m_j(\tau) \Gamma_{ij}(t-\tau) d\tau, \tag{65}$$

where

$$m_j = [F_1, F_2, F_3, M_{11}, M_{22}, M_{33}, M_{12}, M_{23}, M_{13}] \tag{66}$$

and

$$\Gamma_{ij} = [G_{i1}, G_{i2}, G_{i3}, G_{i1,1}, G_{i2,2}, G_{i3,3}, G_{i1,2} + G_{i2,1}, G_{i2,3} + G_{i3,2}, G_{i1,3} + G_{i3,1}]. \tag{67}$$

We discretize  $t$  and  $\tau$  as  $t = k\Delta t$  ( $k = 0, \dots, N_t - 1$ ) and  $\tau = l\Delta t$  ( $l = 0, \dots, N_t - 1$ ), respectively, where  $\Delta t$  is the sampling interval. This yields the discrete form of Eq. (65) given as

$$u_i(k\Delta t) = \sum_{j=1}^9 \sum_{l=0}^{N_t-1} m_j(l\Delta t) \Gamma_{ij}(k\Delta t - l\Delta t) \Delta t. \tag{68}$$

If we have  $N_c$  traces of  $u_i$ , Eq. (68) is written in matrix form

$$\mathbf{u} = \mathbf{\Gamma} \mathbf{m}, \tag{69}$$

where

$$\mathbf{u} = [u_1(0), \dots, u_1((N_t - 1)\Delta t), \dots, u_{N_c}(0), \dots, u_{N_c}((N_t - 1)\Delta t)]^T, \tag{70}$$

$$\mathbf{m} = [m_1(0), \dots, m_1((N_t - 1)\Delta t), \dots, m_9(0), \dots, m_9((N_t - 1)\Delta t)]^T, \tag{71}$$

$$\mathbf{\Gamma} = \begin{bmatrix} \Gamma_{11}(0) & \dots & \Gamma_{11}((1 - N_t)\Delta t) & \dots \\ \vdots & \vdots & \vdots & \vdots \\ \Gamma_{11}((N_t - 1)\Delta t) & \dots & \Gamma_{11}(0) & \dots \\ \vdots & \vdots & \vdots & \vdots \\ \Gamma_{N_{c1}}(0) & \dots & \Gamma_{N_{c1}}((1 - N_t)\Delta t) & \dots \\ \vdots & \vdots & \vdots & \vdots \\ \Gamma_{N_{c1}}((N_t - 1)\Delta t) & \dots & \Gamma_{N_{c1}}(0) & \dots \\ \dots & \Gamma_{19}(0) & \dots & \Gamma_{19}((1 - N_t)\Delta t) \\ \vdots & \vdots & \vdots & \vdots \\ \dots & \Gamma_{19}((N_t - 1)\Delta t) & \dots & \Gamma_{19}(0) \\ \vdots & \vdots & \vdots & \vdots \\ \dots & \Gamma_{N_{c9}}(0) & \dots & \Gamma_{N_{c9}}((1 - N_t)\Delta t) \\ \vdots & \vdots & \vdots & \vdots \\ \dots & \Gamma_{N_{c9}}((N_t - 1)\Delta t) & \dots & \Gamma_{N_{c9}}(0) \end{bmatrix} \Delta t. \tag{72}$$

Here  $T$  denotes transpose.  $\mathbf{\Gamma}$  is called the data kernel, which is an  $N \times M$  matrix, and  $\mathbf{u}$  and  $\mathbf{m}$  are vectors of lengths  $N$  and  $M$ , respectively, where  $N = N_c N_t$  and  $M = 9N_t$ .

We denote observed waveforms as  $\mathbf{u}^{obs}$ , which satisfy the following equation:

$$\mathbf{u}^{obs} - \mathbf{\Gamma} \mathbf{m} = \mathbf{e}, \tag{73}$$

where  $\mathbf{e}$  is the prediction error or misfit. In the case of  $N > M$ , we estimate  $\mathbf{m}$  by minimizing the residual  $R$ , defined as

$$R = \frac{\mathbf{e}^T \mathbf{e}}{\mathbf{u}^{obs T} \mathbf{u}^{obs}} = \frac{\sum_{i=1}^N e_i^2}{\sum_{i=1}^N u_i^{obs}}, \tag{74}$$

where

$$\mathbf{e}^T \mathbf{e} = (\mathbf{u}^{obs} - \mathbf{\Gamma} \mathbf{m})^T (\mathbf{u}^{obs} - \mathbf{\Gamma} \mathbf{m}). \tag{75}$$

To minimize  $R$  is to find values of  $\mathbf{m}$  that satisfy the following equations:

$$\frac{\partial R}{\partial m_q} = 0 \quad (q = 1, \dots, M). \tag{76}$$

This leads to the following relationship (the normal equation)

$$\mathbf{\Gamma}^T \mathbf{\Gamma} \mathbf{m} - \mathbf{\Gamma}^T \mathbf{u}^{obs} = 0. \tag{77}$$

Therefore, we have the following solution:

$$\mathbf{m}^{est} = [\mathbf{\Gamma}^T \mathbf{\Gamma}]^{-1} \mathbf{\Gamma}^T \mathbf{u}^{obs}, \tag{78}$$

which is the least-squares solution. This least-squares approach in the time domain has been used in many volcano

seismological studies e. g., [32,33,48,69,71,72,73,92,94,102, 105,107,108,120,122].

As can be seen from Eq. (68),  $u_i$  is calculated by the convolution of  $m_j$  with  $\Gamma_{ij}$ . This means that the displacement  $u_i(k\Delta t)$  includes the entire effect of the time history of the single force and moment tensor as well as Green's functions before time  $k\Delta t$ . This may be easily understood if we consider  $u_i((N_t - 1)\Delta t)$  in Eq. (68):

$$u_i((N_t - 1)\Delta t) = \Delta t \sum_{j=1}^9 [m_j(0)\Gamma_{ij}((N_t - 1)\Delta t) + m_j(\Delta t)\Gamma_{ij}((N_t - 2)\Delta t) + \dots + m_j((N_t - 2)\Delta t)\Gamma_{ij}(\Delta t) + m_j((N_t - 1)\Delta t)\Gamma_{ij}(0)]. \tag{79}$$

The convolution relationship leads to the large matrix of  $\mathbf{F}$  of Eq. (69), which requires long computation time to obtain the least-squares solution. A better alternative is to solve the inverse problem in the frequency domain [11,95, 116]. The Fourier transform of Eq. (65) is given by

$$\bar{u}_i(\omega) = \sum_{j=1}^9 \bar{m}_j(\omega)\bar{\Gamma}_{ij}(\omega), \tag{80}$$

where  $\omega$  is the angular frequency and  $\bar{u}_i, \bar{m}_j$ , and  $\bar{\Gamma}_{ij}$  are the Fourier transforms of  $u_i, m_j$ , and  $\Gamma_{ij}$ . This frequency domain equation is discretized as follows:

$$\bar{u}_i(k\Delta\omega) = \sum_{j=1}^9 \bar{m}_j(k\Delta\omega)\bar{\Gamma}_{ij}(k\Delta\omega), \tag{81}$$

where  $\Delta\omega = 2\pi/(N_t\Delta t)$  and  $k = 0, \dots, N_t/2$ . Note that the Fourier transformed components are complex and we have a set of the equations corresponding to real and imaginary parts; therefore, Eq. (81) consists of  $N_t$  equations. We further note that  $\bar{u}_i, \bar{m}_j$ , and  $\bar{\Gamma}_{ij}$  in Eq. (81) depend only on  $k\Delta\omega$ . This indicates that individual frequency components of  $\bar{u}_i, \bar{m}_j$ , and  $\bar{\Gamma}_{ij}$  are independent of each other and Eq. (81) can be solved separately. This is different from the time-domain Eq. (65):  $u_i$  is a function of  $k\Delta t$ , whereas  $m_j$  is a function of  $l\Delta t$ , and  $\Gamma_{ij}$  is a function of  $k\Delta t$  and  $l\Delta t$ , so we cannot separate the time-domain equation in individual values of  $k\Delta t$ . The use of the frequency-domain equation results in a smaller size of the data kernel, as shown below.

If we have  $N_c$  traces of  $\bar{u}_i$ , Eq. (81) is written in matrix form

$$\bar{\mathbf{u}} = \bar{\mathbf{F}} \bar{\mathbf{m}}. \tag{82}$$

Here,  $\bar{\mathbf{u}}$  is a vector consisting of  $\bar{u}_i$ , which is arranged in individual frequency components as follows:

$$\bar{\mathbf{u}} = [\bar{u}_1(0), \bar{u}_2(0), \dots, \bar{u}_{N_c}(0), \bar{u}_1(\Delta\omega), \bar{u}_2(\Delta\omega), \dots, \bar{u}_{N_c}(\Delta\omega), \dots, \bar{u}_1((N_t/2)\Delta\omega), \bar{u}_2((N_t/2)\Delta\omega), \dots, \bar{u}_{N_c}((N_t/2)\Delta\omega)]^T \tag{83}$$

or, equivalently,

$$\bar{\mathbf{u}} = [\bar{\mathbf{u}}(0), \bar{\mathbf{u}}(\Delta\omega), \dots, \bar{\mathbf{u}}(k\Delta\omega), \dots, \bar{\mathbf{u}}(N_t/2\Delta\omega)]^T, \tag{84}$$

where

$$\bar{\mathbf{u}}(k\Delta\omega) = [\bar{u}_1(k\Delta\omega), \bar{u}_2(k\Delta\omega), \dots, \bar{u}_{N_c}(k\Delta\omega)]^T. \tag{85}$$

Similarly,  $\bar{\mathbf{m}}$  is a vector consisting of  $\bar{m}_j$ , which is arranged in individual frequency components as

$$\bar{\mathbf{m}} = [\bar{m}_1(0), \bar{m}_2(0), \dots, \bar{m}_9(0), \bar{m}_1(\Delta\omega), \bar{m}_2(\Delta\omega), \dots, \bar{m}_9(\Delta\omega), \dots, \bar{m}_1((N_t/2)\Delta\omega), \bar{m}_2((N_t/2)\Delta\omega), \dots, \bar{m}_9((N_t/2)\Delta\omega)]^T \tag{86}$$

or

$$\bar{\mathbf{m}} = [\bar{\mathbf{m}}(0), \bar{\mathbf{m}}(\Delta\omega), \dots, \bar{\mathbf{m}}(k\Delta\omega), \dots, \bar{\mathbf{m}}((N_t/2)\Delta\omega)]^T, \tag{87}$$

where

$$\bar{\mathbf{m}}(k\Delta\omega) = [\bar{m}_1(k\Delta\omega), \bar{m}_2(k\Delta\omega), \dots, \bar{m}_9(k\Delta\omega)]^T. \tag{88}$$

Accordingly,  $\bar{\mathbf{F}}$  is given as

$$\bar{\mathbf{F}} = \begin{bmatrix} \bar{\Gamma}^{\bar{}}(0) & & & & & & & & \\ & \bar{\Gamma}^{\bar{}}(\Delta\omega) & & & & & & & 0 \\ & & \ddots & & & & & & \\ & & & \bar{\Gamma}^{\bar{}}(k\Delta\omega) & & & & & \\ & & & & \ddots & & & & \\ & & 0 & & & \ddots & & & \\ & & & & & & \bar{\Gamma}^{\bar{}}((N_t/2)\Delta\omega) & & \end{bmatrix}, \tag{89}$$

where  $\bar{\mathbf{F}}(k\Delta\omega)$  is

$$\bar{\mathbf{F}}(k\Delta\omega) = \begin{bmatrix} \bar{\Gamma}_{11}^{\bar{}}(k\Delta\omega) & \cdots & \bar{\Gamma}_{19}^{\bar{}}(k\Delta\omega) \\ \vdots & \ddots & \vdots \\ \bar{\Gamma}_{N_c1}^{\bar{}}(k\Delta\omega) & \cdots & \bar{\Gamma}_{N_c9}^{\bar{}}(k\Delta\omega) \end{bmatrix}. \tag{90}$$

Since  $\bar{\mathbf{F}}$  is the block diagonal matrix, Eq. (82) can be reduced to

$$\bar{\mathbf{u}}(k\Delta\omega) = \bar{\mathbf{F}}(k\Delta\omega)\bar{\mathbf{m}}(k\Delta\omega) \quad (k = 0, \dots, N_t/2). \tag{91}$$

The dimension of matrix  $\bar{\mathbf{F}}(k\Delta\omega)$  is  $N_c \times 9$ , which is  $N_t^2$  times smaller than that of  $\mathbf{F}$  in the time domain (Eq. (72)).

This is because convolution in the time domain is equivalent to the multiplication in the frequency domain. The smaller size of  $\tilde{\mathbf{F}}(k\Delta\omega)$  contributes to faster computation of the least-squares solution. We denote the  $k\Delta\omega$  component of the Fourier transform of observed seismograms as

$$\tilde{\mathbf{u}}^{\text{obs}}(k\Delta\omega) = [\tilde{u}_1^{\text{obs}}(k\Delta\omega), \tilde{u}_2^{\text{obs}}(k\Delta\omega), \dots, \tilde{u}_{N_c}^{\text{obs}}(k\Delta\omega)]^T. \quad (92)$$

The least-squares solution in the frequency domain for the  $k\Delta\omega$  component is then given by

$$\tilde{\mathbf{m}}^{\text{est}}(k\Delta\omega) = [\tilde{\mathbf{F}}^H(k\Delta\omega)\tilde{\mathbf{F}}(k\Delta\omega)]^{-1} \cdot \tilde{\mathbf{F}}^H(k\Delta\omega)\tilde{\mathbf{u}}^{\text{obs}}(k\Delta\omega), \quad (93)$$

where  $H$  denotes the adjoint (Hermitian conjugate). We perform the inverse Fourier transform of  $\tilde{\mathbf{m}}^{\text{est}}(k\Delta\omega)$  ( $k = 0, \dots, N_t/2$ ) to obtain the solution in the time domain.

The use of appropriate Green's functions is critically important to obtain the correct solution  $\mathbf{m}^{\text{est}}$  in either the time or frequency domains. If the source is located in a medium that can be approximated by a homogeneous or layered structure with a flat surface, Green's functions can be calculated by using propagator matrices e.g., [63] and the discrete wavenumber method e.g., [16,17,20]. However, this situation is not common in volcanic regions, and we need to take into account the steep topographies and heterogeneous structures of volcanoes. The boundary integral equation method e.g., [18] and boundary element method e.g., [62] can be used to quantify the effect of three-dimensional topography in a homogeneous or layered structure. To treat both topography and structural heterogeneity simultaneously, the finite-difference method e.g., [106,109] and the discrete lattice method [104] are suitable. Once Green's functions (and their spatial derivatives) are calculated, we can obtain the least-squares solution  $\mathbf{m}^{\text{est}}$  through Eq. (78) or (93). This solution is, however, obtained for a particular source point and a grid search in space is required to find the best-fit source location.

An example of waveform inversion is provided in Figs. 11 and 12. The source mechanism of VLP signals observed by a dense broadband seismic network at Stromboli volcano, Italy, was estimated by using the waveform inversion method in the time domain [32]. In the inversion, VLP waveforms from 18 stations (Fig. 11) were used, and six moment tensor and three single force components for a point source were assumed. Green's functions were calculated by the finite-difference method [106] using the topography of the volcano. The inversion result (Fig. 12)

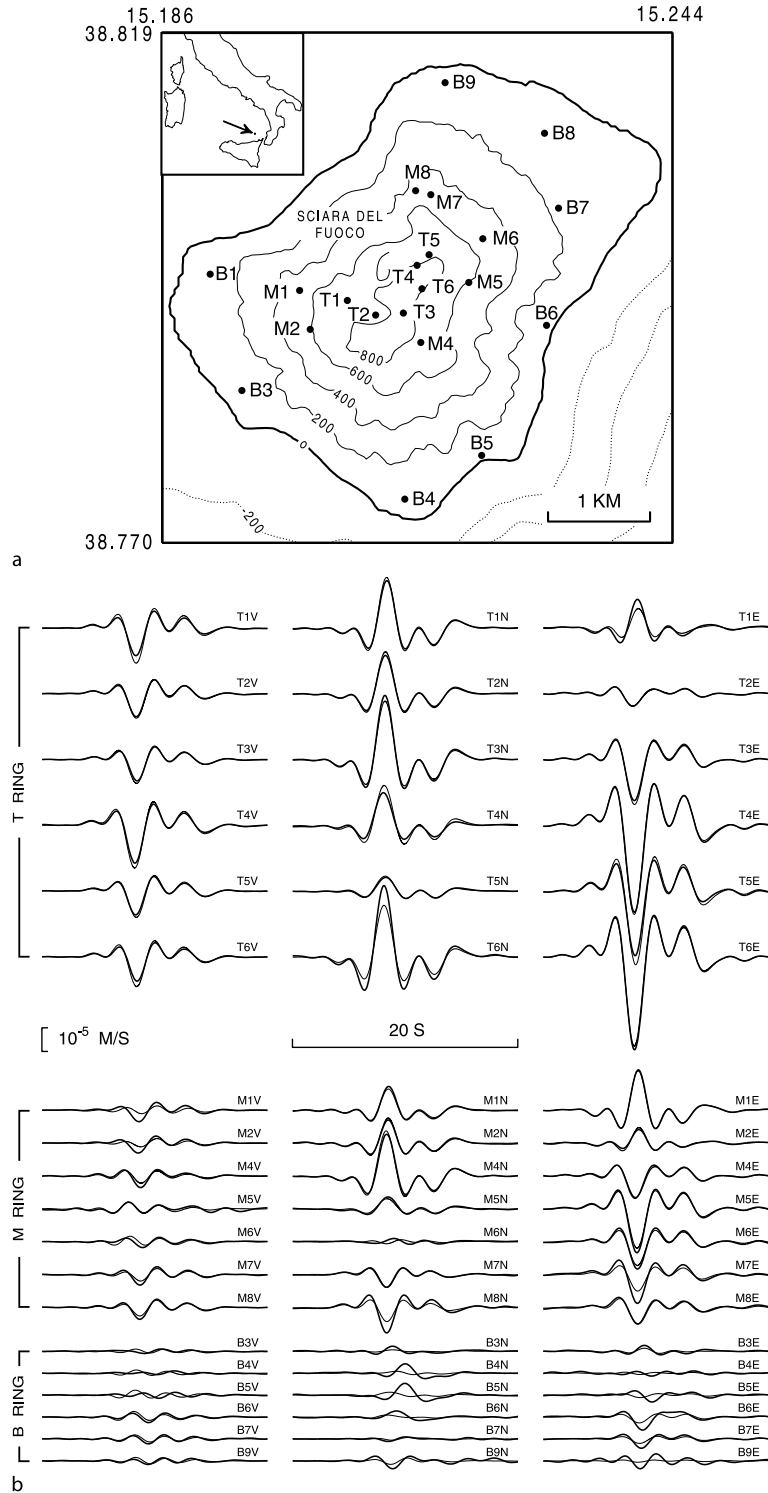
points to volumetric changes accompanying a dominantly vertical single force at the source of the VLP signals. The mechanism is interpreted as the movement of a magma column in response to the piston-like rise of a slug of gas in an inclined crack-like conduit of Stromboli volcano [32].

The waveform inversion approach discussed above might be successful if we have many (more than 10) stations covering the entire edifice of a volcano. However, such dense seismic networks exist at only a limited number of volcanoes. Many volcanoes are monitored by networks of just a few seismic stations. It is also possible that signals would be detected by only a few stations close to the source even in a dense seismic network. Nakano and Kumagai [92] proposed a practical approach for quantifying the source of volcano-seismic signals observed by a small number of seismic stations. In this approach, possible source geometries such as a crack and pipe, are assumed as a priori information in waveform inversion. This assumption reduces the number of free parameters and enables us to estimate the source mechanism and location with waveform data from 2 to 3 three-component seismic stations. Their approach is summarized as follows. The moment tensors for a crack and pipe are given in Eqs. (30) and (34), respectively. If we assume a value for  $\lambda/\mu$ , free parameters in the moment tensors are the source-time function,  $\theta$ , and  $\phi$ . We can determine the source-time function by the waveform inversion discussed above, in which a grid search with respect to  $\theta$  and  $\phi$  is conducted to find the best-fit angles at a fixed location. A grid search in space is also conducted to find the combination of the best-fit angles and source location, and their residuals are compared to adopt the best model. The applicability of this simple approach has been demonstrated by synthetic tests and observed seismograms [92].

## Spectral Analysis

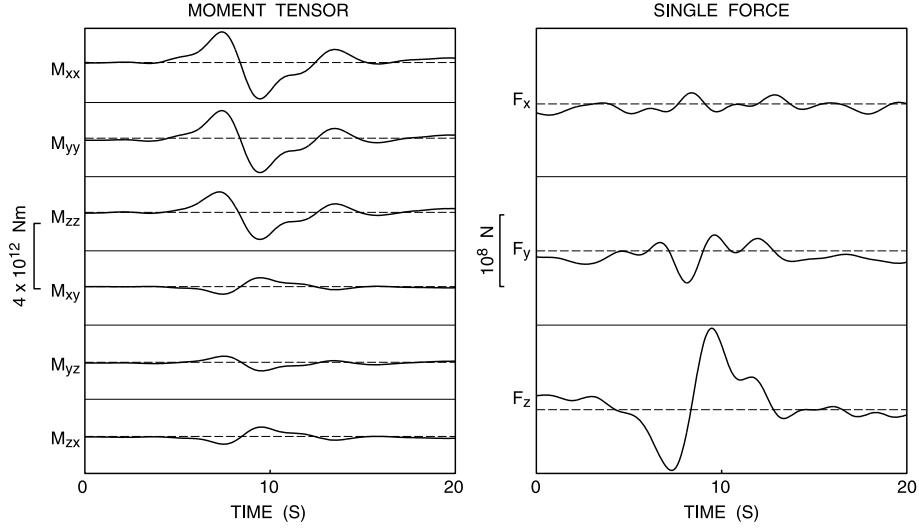
Volcano-seismic signals often show harmonic oscillatory signatures, which are interpreted as oscillations of fluid-filled resonators at the source of these signals. To determine the frequency characteristics of observed oscillations is essential to quantify the characteristic properties of the resonator systems. The discrete Fourier transform (DFT) has traditionally been used for this purpose. In DFT, the discrete time series  $u(k\Delta t)$  ( $k = 0, \dots, N_t - 1$ ) is transformed into the discrete frequency series

$$\tilde{u}(l\Delta\omega) = \Delta t \sum_{k=0}^{N_t-1} u(k\Delta t) e^{2\pi i l k / N_t} \quad (l = 0, \dots, N_t-1), \quad (94)$$



Volcano Seismic Signals, Source Quantification of, Figure 11

**a** Locations of broadband seismic stations at Stromboli Volcano, Italy. **b** Observed VLP velocity waveforms (*thick lines*) bandpassed between 2 and 20 s and the best-fit synthetic seismograms (*thin lines*) obtained from waveform inversion assuming six moment tensor and three single force components for the source mechanism [32]



Volcano Seismic Signals, Source Quantification of, Figure 12  
 Source-time functions of six moment tensor and three single force components estimated by waveform inversion for the VLP signals shown in Fig. 11b [32]

where  $\Delta\omega = 2\pi/(N_t\Delta t)$  and  $i = \sqrt{-1}$ . DFT is the decomposition of time series into a linear combination of the orthogonal basis functions  $e^{2\pi i k/N_t}$  in the frequency domain. Because of the stationary nature of the basis functions, DFT is especially useful to analyze stationary periodic signals. However, volcano-seismic signals often show decaying harmonic oscillations. These may be viewed as the impulsive response of a resonator system, which inherently contains dissipation mechanisms. A decaying harmonic oscillation can be represented by the complex frequency, which cannot be directly estimated by DFT based on the real frequency. The complex frequencies can be determined by solving an inverse problem for parameters in an appropriate model of a given time series. The Sompi method [75] is one such approach that is based on an autoregressive (AR) model of a linear dynamic system. This method has been successfully used to determine the complex frequencies of decaying harmonic oscillations in LP and VLP events. We follow the theory of Sompi presented by Kumazawa et al. [75] and Hori et al. [50].

Let us consider the equation of motion for a linear dynamic system consisting of a block and spring:

$$b \frac{d^2}{dt^2}x(t) + v \frac{d}{dt}x(t) + \kappa x(t) = 0, \tag{95}$$

where  $b$  is the mass of a block,  $\kappa$  is the spring constant, and  $v(dx(t)/dt)$  represents a dissipation term caused by friction, which is proportional to velocity. We set  $x(t) = e^{i\omega t}$  in Eq. (95), which is an eigenfunction of this system. We

then obtain

$$b(i\omega)^2 + v(i\omega) + \kappa = 0. \tag{96}$$

This is the characteristic equation of the system. If we consider oscillation solutions ( $v^2 - 4bk < 0$ ), Eq. (96) has two characteristic solutions:

$$i\omega_1 = \frac{-v + i\sqrt{4bk - v^2}}{2b}, \quad i\omega_2 = i\omega_1^*, \tag{97}$$

where  $*$  denotes the complex conjugate. If we set  $\omega = 2\pi(f - ig)$ , we obtain

$$f_1 = -f_2 = \frac{1}{2\pi} \sqrt{\frac{\kappa}{b} - \frac{v^2}{4b^2}}, \tag{98}$$

$$g_1 = g_2 = -\frac{1}{2\pi} \frac{v}{2b}. \tag{99}$$

Let us consider the following difference equation:

$$\sum_{j=0}^2 a(j)x((k-j)\Delta t) = 0 \tag{100}$$

or

$$a(0)x(k\Delta t) + a(1)x(k\Delta t - \Delta t) + a(2)x(k\Delta t - 2\Delta t) = 0. \tag{101}$$

If we set  $x(k\Delta t) = z^k$ , where  $z = e^{i\omega\Delta t}$ , Eq. (100) is given as

$$a(0)z^k + a(1)z^{k-1} + a(2)z^{k-2} = 0 \tag{102}$$

or, equivalently,

$$a(0)z^2 + a(1)z + a(2) = 0. \quad (103)$$

Equation (103) is the characteristic equation of the difference equation. If we consider oscillation solutions ( $a^2(1) - 4a(0)a(2) < 0$ ), Eq. (103) has the following two characteristic solutions:

$$z_1 = \frac{-a(1) + i\sqrt{4a(0)a(2) - a^2(1)}}{2a(0)}, \quad z_2 = z_1^*. \quad (104)$$

These two solutions constitute complex conjugate pairs. By setting  $\omega = 2\pi(f - ig)$ , we obtain

$$f_1 = -f_2 = \frac{1}{2\pi\Delta t} \tan^{-1} \left( -\sqrt{\frac{4a(0)a(2)}{a^2(1)} - 1} \right), \quad (105)$$

$$g_1 = g_2 = \frac{1}{2\pi\Delta t} \ln \sqrt{\frac{a(2)}{a(0)}}. \quad (106)$$

These solutions are equivalent to the solutions of the linear dynamic system given in Eqs. (98) and (99). Therefore, Eq. (100) is a difference form of the equation of motion (95).

We further consider a dynamic system consisting of two block-spring subsystems, which are coupled to each other. The equations of motion for this system are written as

$$M_1 \frac{d^2}{dt^2} x_1(t) + L_1 \frac{d}{dt} x_1(t) + K_1 x_1(t) = C(x_2 - x_1), \quad (107)$$

$$M_2 \frac{d^2}{dt^2} x_2(t) + L_2 \frac{d}{dt} x_2(t) + K_2 x_2(t) = C(x_1 - x_2), \quad (108)$$

where  $M_i$ ,  $L_i$ , and  $K_i$  are the mass, dissipation coefficient, and spring constant of the  $i$ th block and  $C$  is the spring constant between the blocks. The set of the equations of motion (107) and (108) can be equivalently given in the following form:

$$\begin{pmatrix} M_1 \frac{d^2}{dt^2} + L_1 \frac{d}{dt} + K_1 + C & -C \\ -C & M_2 \frac{d^2}{dt^2} + L_2 \frac{d}{dt} + K_2 + C \end{pmatrix} \cdot \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} = 0. \quad (109)$$

These equations have non-trivial solutions of  $x_i(t)$  if, and only if,

$$\det \begin{vmatrix} M_1 \frac{d^2}{dt^2} + L_1 \frac{d}{dt} + K_1 + C & -C \\ -C & M_2 \frac{d^2}{dt^2} + L_2 \frac{d}{dt} + K_2 + C \end{vmatrix} = 0. \quad (110)$$

This leads to the following 4th-order differential equation:

$$\left[ \left( M_1 \frac{d^2}{dt^2} + L_1 \frac{d}{dt} + K_1 + C \right) \cdot \left( M_2 \frac{d^2}{dt^2} + L_2 \frac{d}{dt} + K_2 + C \right) - C^2 \right] x_i(t) = 0, \quad (111)$$

where  $i = 1, 2$ . Each 4th-order differential equation has four complex frequencies, of which two are complex conjugates. Therefore, each block has two independent complex frequencies in this dynamic system. Similarly, a dynamic system consisting of  $m$  block-spring subsystems, which are all connected to each other, can be described by the following  $2m$ -order differential equation:

$$\sum_{j=0}^{2m} b_j \frac{d^j}{dt^j} x_i(t) = 0 \quad (i = 1, 2, \dots, m), \quad (112)$$

where  $b_j$  denotes the coefficients determined by the characteristic properties of the system. Each block has  $m$  independent complex frequencies in this case. A difference form of Eq. (112) is

$$\sum_{j=0}^{2m} a(j) x_i(k\Delta t - j\Delta t) = 0, \quad (113)$$

which is a homogeneous  $2m$ -order AR equation. Although we have considered block-spring systems as an example, Eq. (112) is a general equation to describe the free oscillations of any linear systems. We use its difference form, Eq. (113), as the basic equation to determine the characteristic complex frequencies of linear systems from time series records.

Hereafter, I set  $\Delta t = 1$  to simplify the following notations. We assume that observed discrete time series  $u(k)$  consists of signal  $x(k)$  and white noise  $e(k)$ :

$$u(k) = x(k) + e(k) \quad (k = 0, 1, \dots, N_t - 1), \quad (114)$$

and  $x(k)$  satisfies the following homogeneous AR equation:

$$\sum_{j=0}^{2m} a(j) x(k - j) = 0. \quad (115)$$

To determine the  $2m + 1$  AR coefficients, we minimize the residual  $R$ :

$$R = \frac{1}{N_t - 2m} \sum_{k=2m}^{N_t-1} \left[ \sum_{j=0}^{2m} a(j) u(k - j) \right]^2. \quad (116)$$



Since Eq. (115) is homogeneous, we cannot determine absolute values of the AR coefficients. To avoid the trivial solution  $a(j) = 0$  ( $j = 0, 1, \dots, 2m$ ), we impose the condition

$$\sum_{j=0}^{2m} a^2(j) = 1. \tag{117}$$

To minimize the residual  $R$  under the condition of Eq. (117) is a conditional minimization problem. This problem can be solved by minimizing the residual  $R'$ :

$$R' = R - \beta \left( \sum_{j=0}^{2m} a^2(j) - 1 \right), \tag{118}$$

where  $\beta$  is a constant. Consequently, we solve the following equation:

$$\frac{\partial}{\partial a(l)} R' = 0 \quad (l = 0, 1, \dots, 2m). \tag{119}$$

This equation leads to an eigenvalue problem

$$\mathbf{P}\mathbf{a} = \beta\mathbf{a}, \tag{120}$$

where  $\mathbf{a} = (a(0), a(1), \dots, a(2m))^T$  and  $\mathbf{P}$  is a non-Toeplitz matrix given by

$$\mathbf{P} = \frac{1}{N_t - 2m} \begin{bmatrix} \sum_{n=2m}^{N_t-1} u(n)u(n) & \cdots & \sum_{n=2m}^{N_t-1} u(n-2m)u(n) \\ \vdots & \ddots & \vdots \\ \sum_{n=2m}^{N_t-1} u(n)u(n-2m) & \cdots & \sum_{n=2m}^{N_t-1} u(n-2m)u(n-2m) \end{bmatrix}. \tag{121}$$

By solving the eigenvalue problem of Eq. (120) we obtain  $2m + 1$  eigenvalues  $\beta^{(r)}$  ( $r = 1, 2, \dots, 2m + 1$ ) and corresponding eigenvectors  $\mathbf{a}^{(r)}$  ( $r = 1, 2, \dots, 2m + 1$ ). Note that the residual  $R$  in Eq. (116) is given as

$$R = \mathbf{a}^T \mathbf{P}\mathbf{a} = \mathbf{a}^T \beta\mathbf{a} = \beta. \tag{122}$$

Therefore, each eigenvector gives a local minimum of the residual  $R$ . We therefore adopt the eigenvector corresponding to the minimum eigenvalue, which minimizes the residual  $R$ . Hereafter,  $\mathbf{a}$  denotes the eigenvector corresponding to the minimum eigenvalue.

The next step is to determine the complex frequencies from the eigenvector  $\mathbf{a}$ . The characteristic equation of the  $2m$ -order homogeneous AR equation is given as

$$\sum_{j=0}^{2m} a(j)z^{-j} = 0. \tag{123}$$

This equation is the  $2m$ -degree algebraic equation for  $z^{-1}$ , and has  $2m$  complex solutions, which can be solved by a numerical procedure. As mentioned before, a complex solution is always accompanied by its complex conjugate, and therefore there are  $m$  independent solutions  $z_l$  ( $l = 1, 2, \dots, m$ ) in Eq. (123). These solutions are related to the complex frequencies ( $f_l - ig_l$ ) as

$$z_l = e^{2\pi i(f_l - ig_l)} \tag{124}$$

or

$$f_l - ig_l = \ln z_l / (2\pi i). \tag{125}$$

The  $Q$  factor is defined through the complex frequencies as

$$Q_l = f_l / (-2g_l). \tag{126}$$

The final step is to determine the amplitudes of the solutions  $z_l$ . If we set a conjugate pair as  $z_l = z_{m+l}^*$ , the general solution of Eq. (123) is given as

$$\begin{aligned} x(k) &= \sum_{l=1}^m (\alpha_l z_l^k + \alpha_{m+l} z_{m+l}^k) \\ &= \sum_{l=1}^m (\alpha_l z_l^k + (\alpha_l z_l^k)^*). \end{aligned} \tag{127}$$

Here  $\alpha_l$  is the complex amplitude given as

$$\alpha_l = r_l e^{i\phi_l}, \tag{128}$$

where  $r_l$  and  $\phi_l$  are the real amplitude and phase, respectively. This is equivalently written as

$$\alpha_l = \alpha_l^r + i\alpha_l^i, \tag{129}$$

where  $\alpha_l^r$  and  $\alpha_l^i$  are the real and imaginary parts of  $\alpha_l$ , respectively. Then, the general solution of Eq. (127) can be rewritten

$$x(k) = \sum_{l=1}^m (2\alpha_l^r C_l(k) - 2\alpha_l^i S_l(k)), \tag{130}$$

where  $C_l(k) = e^{2\pi g_l k} \cos(2\pi f_l k)$  and  $S_l(k) = e^{2\pi g_l k} \sin(2\pi f_l k)$ . The observed time series thus satisfies the following equation

$$u(k) - \sum_{l=1}^m (2\alpha_l^r C_l(k) - 2\alpha_l^i S_l(k)) = e(k) \tag{131}$$

$(k = 0, 1, \dots, N_t - 1)$

or, in matrix form

$$\mathbf{u} - \mathbf{A}\boldsymbol{\alpha} = \mathbf{e}, \tag{132}$$

where

$$\mathbf{u} = (u(0), u(1), \dots, u(N_t - 1))^T, \tag{133}$$

$$\boldsymbol{\alpha} = (2\alpha_1^r, \dots, 2\alpha_m^r, 2\alpha_1^i, \dots, 2\alpha_m^i)^T, \tag{134}$$

$$\mathbf{e} = (e(0), e(1), \dots, e(N_t - 1))^T, \tag{135}$$

and

$$\mathbf{A} = \begin{bmatrix} C_1(0) & \dots & C_m(0) \\ \vdots & & \vdots \\ C_1(N_t - 1) & \dots & C_m(N_t - 1) \\ & -S_1(0) & \dots & -S_m(0) \\ & \vdots & & \vdots \\ & -S_1(N_t - 1) & \dots & -S_m(N_t - 1) \end{bmatrix}. \tag{136}$$

We estimate  $\boldsymbol{\alpha}$  by minimizing the residual  $R_a$  as follows:

$$R_a = (\mathbf{u} - \mathbf{A}\boldsymbol{\alpha})^T(\mathbf{u} - \mathbf{A}\boldsymbol{\alpha}). \tag{137}$$

This is an ordinary least-squares problem, and the solution is given as

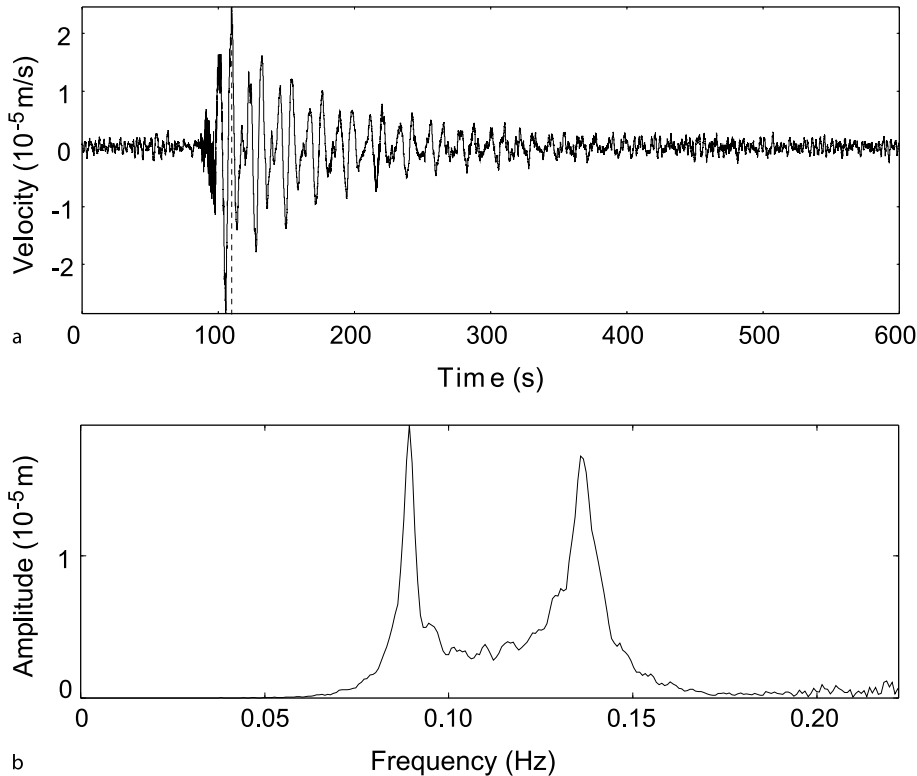
$$\boldsymbol{\alpha} = [\mathbf{A}^T\mathbf{A}]^{-1}\mathbf{A}^T\mathbf{u}. \tag{138}$$

The real amplitudes  $r_l$  and phases  $\phi_l$  at  $k = 0$  are determined from estimated  $\boldsymbol{\alpha}$  as

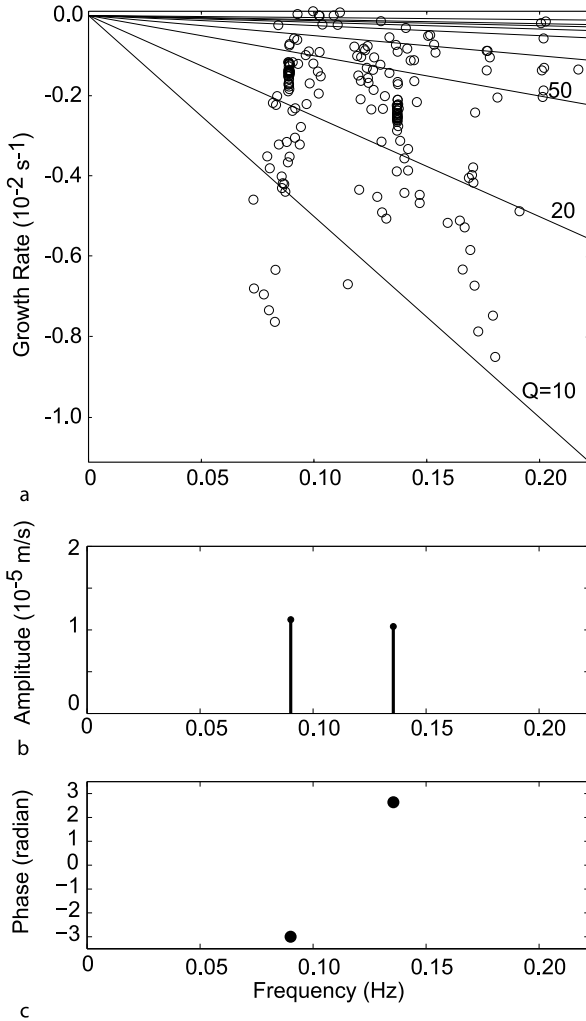
$$r_l = \sqrt{(\alpha_l^r)^2 + (\alpha_l^i)^2} \quad (l = 1, \dots, m), \tag{139}$$

$$\phi_l = \tan^{-1}(\alpha_l^i/\alpha_l^r) \quad (l = 1, \dots, m). \tag{140}$$

An example of the Sompi analysis of a VLP waveform (Fig. 13a) is shown in Fig. 14. The VLP waveform in Fig. 13a shows decaying harmonic oscillations, which are interpreted as the resonances of a magmatic dike triggered by a time-localized excitation [65,72]. The amplitude spectrum of the VLP waveform based on DFT is shown in Fig. 13b, which indicates two dominant peaks. As such,



Volcano Seismic Signals, Source Quantification of, Figure 13  
**a** VLP waveform observed at Hachijo Island, Japan, associated with a VT earthquake swarm [65,72]. **b** Amplitude spectrum of the waveform displayed in Fig. 13a



Volcano Seismic Signals, Source Quantification of, Figure 14  
 An example of the Sompi spectral analysis of the VLP waveform of Fig. 13a, in which Sompi was applied to the waveform after the time of the maximum amplitude indicated by the dashed line in Fig. 13a. **a** A frequency-growth rate diagram for estimated complex frequencies using trial AR orders between 4 and 60. Solid lines represent lines along which the Q factor is constant. We can identify two densely populated regions, which represent the signal, while other scattered points represent incoherent noise. **b** Amplitudes and **c** phases at the time of the maximum amplitude using the two complex frequencies determined at an AR order of 40

DFT provides frequency and amplitude information simultaneously. Q values may be determined by the bandwidths of these spectral peaks at halves their peak amplitudes. The estimated Q values, however, can be biased, because the whole waveform including the tail of decaying oscillations as well as the onset of growing oscillations is

used in this approach. A box-car window can be applied to extract the decaying oscillations, but this distorts the spectrum and may lead to biased estimations of Q values.

In Sompi, on the other hand, we delete the onset portion of the signal and use the decaying harmonic oscillations only, which satisfy the homogeneous assumption of the AR equation. Since the number of signals (oscillations) is not known a priori, we use trial AR orders and solve the eigenvalue problems for individual trial AR orders to determine the complex frequencies, which are plotted in a frequency-growth rate (f-g) diagram. Figure 14a is the f-g diagram of the complex frequencies determined from the VLP waveform by using trial AR orders between 4 and 60. Densely populated regions on the f-g diagram represent the signal for which the complex frequencies are stably determined for different AR orders, while scattered points represent incoherent noise. Sompi then determines the complex amplitudes  $re^{i\phi}$  by using the estimated complex frequencies. The optimum AR order and the number of signals may be determined by the two-parameter Akaike information criterion (AIC) [79] or the extended information criterion (EIC) [128]. Figure 14b and c plot the amplitudes  $r$  and phases  $\phi$ , respectively, at the beginning of the decaying harmonic oscillations and using the two complex frequencies determined at an AR order of 40.

Some remarks on the theory of the Sompi method may be required. Although Hori et al. [50] and Kumazawa et al. [75] stated that Sompi is a new spectral method. This may be incorrect, as suggested by Ulrych and Sacchi [128]. Let us consider again the AR Eq. (115), which is equivalently rewritten as

$$x(k) = \sum_{j=1}^{2m} a'(j)x(k-j), \tag{141}$$

where  $a'(j) = -a(j)/a(0)$ . From Eq. (114), we have

$$u(k) = x(k) + e(k) = \sum_{j=1}^{2m} a'(j)x(k-j) + e(k). \tag{142}$$

Substituting  $u(k-j) = x(k-j) + e(k-j)$  into Eq. (142) we obtain

$$u(k) = \sum_{j=1}^{2m} a'(j)u(k-j) + e(k) - \sum_{j=1}^{2m} a'(j)e(k-j). \tag{143}$$

This is the special autoregressive-moving average (ARMA) model discussed by Ulrych and Clayton [127], who showed that the AR coefficients in Eq. (143) can be determined by an eigenvalue problem that is essentially the

same as that given in Eq. (120). In this sense, Sompi is not a new method. However, the Sompi theory, which is based not only on mathematics or information theory but also on physical concepts, yielded a systematic approach to estimate the complex frequencies and complex amplitudes, which can be directly related to the characteristic properties of a linear dynamic system and the system excitation. This approach has been favorably used in analysis of LP and VLP events to estimate the complex frequencies and their temporal variations e. g., [38,65,66,70,72,78,85, 91,93].

**Fluid-Solid Interactions**

Understanding the physics of the fluid-solid interactions is crucial to interpreting source mechanisms and spectral characters of volcano-seismic signals. The pioneering study on the fluid-solid interaction was by Biot [14], who studied seismic waves in a fluid-filled borehole and showed the existence of a slow wave propagating along the borehole boundary. This is called the tube wave. Aki et al. [2] first studied a fluid-filled crack, although the motion of the fluid in the crack was not adequately treated in their study. Chouet [22,23,24] fully studied the fluid-filled crack and demonstrated the existence of a slow wave in the fluid-filled crack, which was called the crack wave. We first consider a fluid layer sandwiched between two homogeneous elastic half-spaces following Ferrazzini and Aki [39] to examine the physical properties of the crack wave. Then, we examine the resonance of a crack of finite size containing a fluid by using the approach proposed by Chouet [22]. For simplicity, we assume an inviscid fluid.

**Crack Wave**

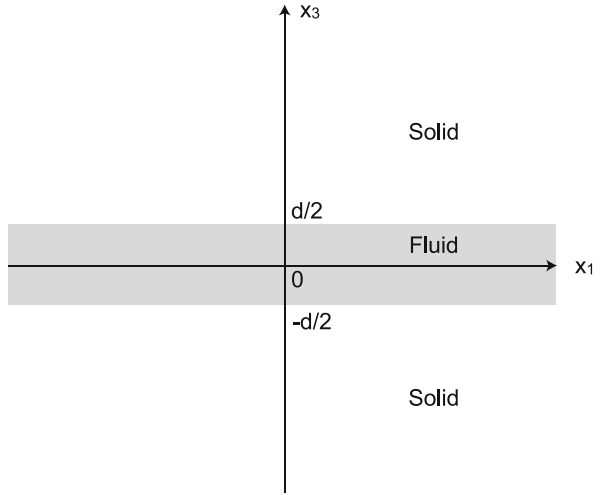
Let us consider a compressible fluid layer of thickness  $d$ , which is sandwiched between two elastic half-spaces bounded at  $x_3 = -d/2$  and  $x_3 = d/2$  (Fig. 15). The equation of motion for the fluid is given by

$$\rho_f \frac{\partial^2 u_1^f}{\partial t^2} = -\frac{\partial p}{\partial x_1}, \tag{144}$$

$$\rho_f \frac{\partial^2 u_3^f}{\partial t^2} = -\frac{\partial p}{\partial x_3}. \tag{145}$$

Here,  $\rho_f$  is the fluid density,  $u_1^f$  and  $u_3^f$  are the  $x_1$  and  $x_3$  components of the displacement in the fluid, respectively, and  $p$  is the fluid pressure satisfying the following continuity equation:

$$p = -b \left( \frac{\partial u_1^f}{\partial x_1} + \frac{\partial u_3^f}{\partial x_3} \right), \tag{146}$$



Volcano Seismic Signals, Source Quantification of, Figure 15  
**A fluid layer of thickness  $d$  sandwiched between two elastic half-spaces bounded at  $x_3 = -d/2$  and  $x_3 = d/2$**

where  $b$  is the bulk modulus of the fluid. The equation of motion for the solid is

$$\rho_s \frac{\partial^2 u_1}{\partial t^2} = \frac{\partial \sigma_{11}}{\partial x_1} + \frac{\partial \sigma_{13}}{\partial x_3}, \tag{147}$$

$$\rho_s \frac{\partial^2 u_3}{\partial t^2} = \frac{\partial \sigma_{13}}{\partial x_1} + \frac{\partial \sigma_{33}}{\partial x_3}. \tag{148}$$

Here,  $\rho_s$  is the solid density,  $u_1$  and  $u_3$  are the  $x_1$  and  $x_3$  components of the displacement in the solid, respectively, and  $\sigma_{11}$ ,  $\sigma_{33}$ , and  $\sigma_{13}$  are the components of stress:

$$\sigma_{11} = \lambda \left( \frac{\partial u_1}{\partial x_1} + \frac{\partial u_3}{\partial x_3} \right) + 2\mu \left( \frac{\partial u_1}{\partial x_1} \right), \tag{149}$$

$$\sigma_{33} = \lambda \left( \frac{\partial u_1}{\partial x_1} + \frac{\partial u_3}{\partial x_3} \right) + 2\mu \left( \frac{\partial u_3}{\partial x_3} \right), \tag{150}$$

$$\sigma_{13} = \mu \left( \frac{\partial u_1}{\partial x_3} + \frac{\partial u_3}{\partial x_1} \right). \tag{151}$$

The boundary conditions at the fluid-solid interfaces ( $x_3 = \pm d/2$ ) are

$$u_3 = u_3^f, \tag{152}$$

$$\sigma_{33} = -p, \tag{153}$$

$$\sigma_{13} = 0. \tag{154}$$

The first and second conditions mean that the normal displacement and stress across the interface are both continuous. The third condition states that the shear stress vanishes at the interface.

For waves propagating in the  $x_1$  direction, the displacements in the fluid may be written in the forms

$$u_1^f = r_1(x_3; k, \omega)e^{i(kx_1 - \omega t)}, \tag{155}$$

$$u_3^f = ir_2(x_3; k, \omega)e^{i(kx_1 - \omega t)}, \tag{156}$$

Substituting these equations into Eq. (146) we obtain

$$p = -ir_3e^{i(kx_1 - \omega t)}, \tag{157}$$

where  $k$  and  $\omega$  are the wavenumber and angular frequency, respectively

$$r_3 = b \left( kr_1 + \frac{dr_2}{dx_3} \right). \tag{158}$$

Substituting Eqs. (155) and (156) into Eqs. (144) and (145), respectively, and using Eq. (158) we have

$$\rho_f \omega^2 r_1 = kr_3, \tag{159}$$

$$-\rho_f \omega^2 r_2 = \frac{dr_3}{dx_3}. \tag{160}$$

Equations (158), (159), and (160) can be arranged in the following matrix form:

$$\begin{bmatrix} \frac{dr_2}{dx_3} \\ \frac{dr_3}{dx_3} \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{b} - \frac{k^2}{\rho_f \omega^2} \\ -\rho_f \omega^2 & 0 \end{bmatrix} \begin{bmatrix} r_2 \\ r_3 \end{bmatrix}. \tag{161}$$

For the solid, the displacement components of waves propagating in the  $x_1$  direction may be written as

$$u_1 = y_1(x_3; k, \omega)e^{i(kx_1 - \omega t)}, \tag{162}$$

$$u_3 = iy_2(x_3; k, \omega)e^{i(kx_1 - \omega t)}. \tag{163}$$

Substituting these equations into Eqs. (147), (148), (149), (150), and (151) we obtain the relationship:

$$\begin{bmatrix} \frac{dy_1}{dx_3} \\ \frac{dy_2}{dx_3} \\ \frac{dy_3}{dx_3} \\ \frac{dy_4}{dx_3} \end{bmatrix} = \begin{bmatrix} 0 & k & \frac{1}{\mu} & 0 \\ \frac{-k\lambda}{\lambda+2\mu} & 0 & 0 & \frac{1}{\lambda+2\mu} \\ k^2\zeta - \rho_s\omega^2 & 0 & 0 & \frac{k\lambda}{\lambda+2\mu} \\ 0 & -\rho_s\omega^2 - k & 0 & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}, \tag{164}$$

where  $\zeta = \frac{4\mu(\lambda+\mu)}{\lambda+2\mu}$  and

$$y_3 = \mu \left( \frac{dy_1}{dx_3} - ky_2 \right), \tag{165}$$

$$y_4 = (\lambda + 2\mu) \frac{dy_2}{dx_3} + k\lambda y_1. \tag{166}$$

The boundary condition Eqs.(152), (153), and (154) can then be written as

$$y_2 = r_2, \tag{167}$$

$$y_4 = r_3, \tag{168}$$

$$y_3 = 0, \tag{169}$$

respectively, at  $x_3 = \pm d/2$ . The systems of Eqs. (161) and (164) are to be solved under the boundary conditions of Eqs. (167), (168), and (169), in which we seek solutions satisfying the following two additional conditions

$$r_2(-x_3) = -r_2(x_3) \quad -d/2 \leq x_3 \leq d/2, \tag{170}$$

$$y_1, y_2 \rightarrow 0 \quad x_3 \rightarrow \pm\infty. \tag{171}$$

The first condition means that motion in the fluid layer is symmetric with respect to  $x_3$ . There is the antisymmetric motion in the fluid layer with respect to  $x_3$ . This motion is the same as that observed in a fluid layer overlying a solid half-space, such as ocean, and has been investigated by Biot [15] and Tolstoy [124]. This motion does not produce the slow wave, and I do not discuss it here.

The solution for the systems of Eqs. (161) and (164) under the conditions mentioned above can be determined by the following numerical procedure, which was proposed by Takeuchi and Saito [121]. Since we assume symmetric motion in the fluid layer, it is enough to consider the region  $x_3 \geq 0$ . We choose  $x_3 = h$ , where  $h$  is appropriately far from the fluid-solid interface to satisfy  $y_1 = y_2 = 0$ . At  $x_3 = h$ , we assume two independent initial values  $y_1 = y_2 = y_4 = 0, y_3 = 1$  and  $y_1 = y_2 = y_3 = 0, y_4 = 1$  and integrate the equation for the solid (164) with values of  $k$  and  $\omega$  by the Runge-Kutta method. Then, we have two independent solutions  $y_i^A(x_3; k, \omega)$  and  $y_i^B(x_3; k, \omega)$  ( $i = 1, 2, 3, 4$ ). Any solution of Eq. (164) may be expressed as a linear combination of the two solutions:

$$y_i(x_3; k, \omega) = Ay_i^A(x_3; k, \omega) + By_i^B(x_3; k, \omega) \quad (i = 1, 2, 3, 4), \tag{172}$$

where  $A$  and  $B$  are arbitrary constants of integration for the two solutions. At  $x_3 = 0$ , we assume initial values  $r_2 = 0$  and  $r_3 = 1$ , and integrate the equation for the fluid (161) with values of  $k$  and  $\omega$ . Then, we have a solution  $r_i^C(x_3; k, \omega)$  ( $i = 1, 2$ ), and any solution of Eq. (161) may be given as

$$r_i(x_3; k, \omega) = Cr_i^C(x_3; k, \omega) \quad (i = 2, 3), \tag{173}$$

where  $C$  is an arbitrary constant of integration. From the boundary conditions of Eqs. (167), (168), and (169), we have the following relationships at  $x_3 = d/2$ :

$$\begin{aligned} Ay_2^A(d/2; k, \omega) + By_2^B(d/2; k, \omega) - Cr_2^C(d/2; k, \omega) &= 0, \\ Ay_4^A(d/2; k, \omega) + By_4^B(d/2; k, \omega) - Cr_3^C(d/2; k, \omega) &= 0, \\ Ay_3^A(d/2; k, \omega) + By_3^B(d/2; k, \omega) &= 0. \end{aligned} \quad (174)$$

These equations have a nontrivial solution if, and only if,

$$\det \begin{vmatrix} y_2^A(d/2; k, \omega) & y_2^B(d/2; k, \omega) & -r_2^C(d/2; k, \omega) \\ y_4^A(d/2; k, \omega) & y_4^B(d/2; k, \omega) & -r_3^C(d/2; k, \omega) \\ y_3^A(d/2; k, \omega) & y_3^B(d/2; k, \omega) & 0 \end{vmatrix} = 0. \quad (175)$$

The dispersion relation of the fluid-solid coupled waves can be determined by searching for  $k$  and  $\omega$  that satisfy Eq. (175). From Eq. (174) we have

$$\frac{B}{A} = -\frac{y_3^A(d/2; k, \omega)}{y_3^B(d/2; k, \omega)}, \quad (176)$$

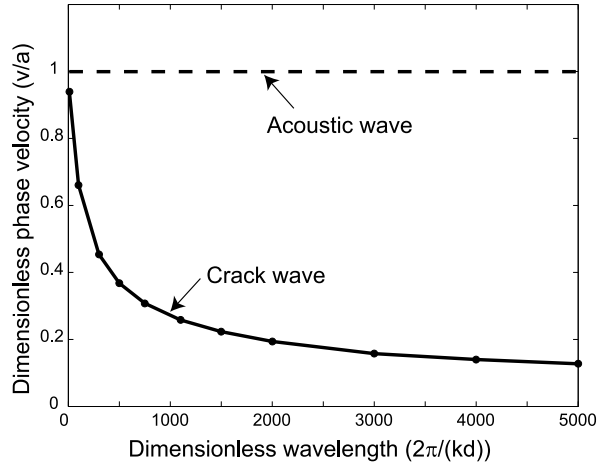
$$\frac{C}{A} = \frac{y_2^A(d/2; k, \omega)}{r_2^C(d/2; k, \omega)} - \frac{y_3^A(d/2; k, \omega)}{y_3^B(d/2; k, \omega)} \frac{y_2^B(d/2; k, \omega)}{r_2^C(d/2; k, \omega)}, \quad (177)$$

and the eigenfunctions are given as

$$y_i(x_3; k, \omega)/A = y_i^A(x_3; k, \omega) + (B/A)y_i^B(x_3; k, \omega) \quad (i = 1, 2, 3, 4), \quad (178)$$

$$r_i(x_3; k, \omega)/A = (C/A)r_i^C(x_3; k, \omega) \quad (i = 2, 3). \quad (179)$$

The dispersion relationship and eigenfunctions of the crack wave calculated by the above procedure are shown in Figs. 16 and 17, respectively. In this example, I used the following parameter values:  $\rho_s = 1.10\rho_f$ ,  $b = 0.06\mu$ , and  $\lambda = 2\mu$ , which mimic a dike containing basaltic magma with a gas-volume fraction of 15%. The dispersion relationship in Fig. 16 clearly indicates that the phase velocity of the fluid-solid coupled wave depends on the wavelength, in which the velocity decreases as the wavelength increases. This wave is called the crack wave. Figure 17 displays dimensionless normal displacements ( $y_2$  and  $r_2$ ) and normal stress and pressure ( $y_4$  and  $r_3$ ) as a function of dimensionless  $x_3$ . Figure 17 indicates that the normal displacement is symmetric with respect to  $x_3$ . This motion causes the deformation of the fluid-solid boundary. The deformation has the effect of lowering the phase velocity, which is examined below.



Volcano Seismic Signals, Source Quantification of, Figure 16  
The phase velocities of the acoustic wave (dashed line) and crack wave (solid line). These are normalized by the fluid acoustic velocity, and plotted as a function of the wavelength normalized by the thickness of the fluid layer

### Fluid-Filled Crack Model

Chouet [22,23,24] studied the motion of a finite crack containing a fluid in an infinite elastic medium. Following Chouet [22], we consider a crack set in the plane  $x_3 = 0$ , which extends from  $-W/2$  to  $W/2$  along the  $x_1$  axis and from 0 to  $L$  in the  $x_2$  direction (Fig. 18). The thickness of the crack is  $d$ , which is much smaller than  $L$ . Chouet [22] used the equations of motion for the solid given as

$$\rho_s \frac{\partial v_i}{\partial t} = \frac{\partial \sigma_{ij}}{\partial x_j}, \quad (180)$$

with the stress-strain relationship

$$\frac{\partial \sigma_{ij}}{\partial t} = \lambda \frac{\partial v_k}{\partial x_k} \delta_{ij} + \mu \left( \frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i} \right), \quad (181)$$

where  $v_j$  is particle velocity of the solid and  $i = 1, 2, 3$ . The equations of motion for the fluid are given as

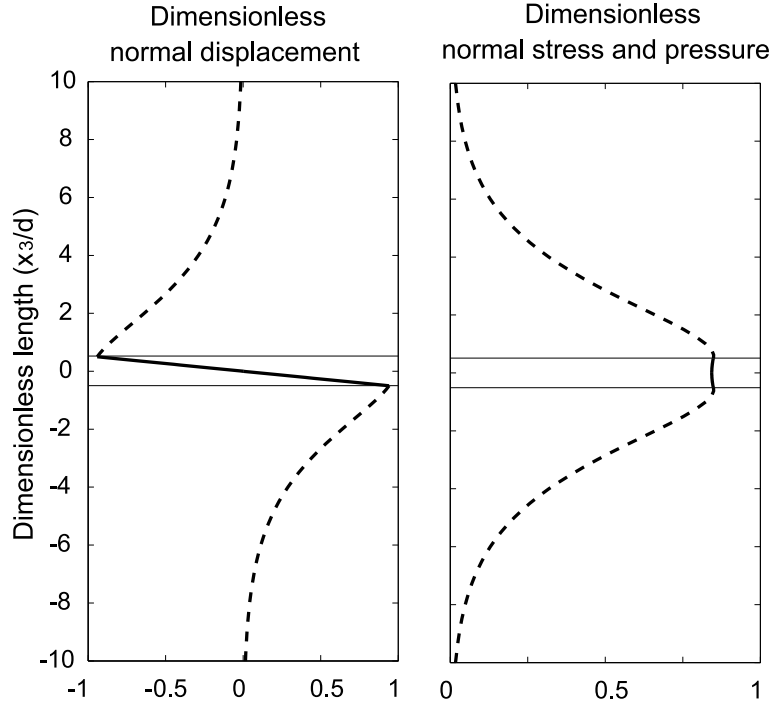
$$\rho_f \frac{\partial^2 u_i^f}{\partial t^2} = -\frac{\partial p}{\partial x_i}, \quad (182)$$

with the equation of continuity

$$p = -b \frac{\partial u_k^f}{\partial x_k}. \quad (183)$$

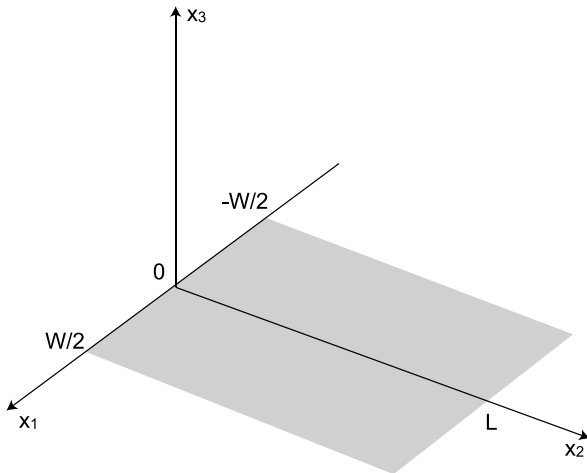
The boundary conditions on the crack surface  $S_1$  are

$$u_3^f = u_3, \quad \sigma_{33} = -p, \quad \sigma_{13} = \sigma_{23} = 0, \quad (184)$$



Volcano Seismic Signals, Source Quantification of, Figure 17

Dimensionless normal displacements, normal stress, and pressure as a function of  $x_3$  normalized by the thickness of the fluid layer. *Thin solid lines* represent the boundaries between the fluid and solid, and *thick dashed and solid lines* represent the functions in the fluid and solid regions, respectively



Volcano Seismic Signals, Source Quantification of, Figure 18

Coordinates and geometry of a crack set in the plane  $x_3 = 0$ , which extends from  $-W/2$  to  $+W/2$  on the  $x_1$  axis and from  $0$  to  $L$  on the  $x_2$  axis

which are the continuity of normal displacement and stress and the vanishment of shear stress on  $S_1$ . Let us introduce  $u_d$ , which represents the normal displacement of the crack wall and is identical to  $u_3^f = u_3$  on  $S_1$ .

We integrate Eq. (183) with respect to  $x_3$  within the crack and divide it by  $d$ :

$$\begin{aligned} & \frac{1}{d} \int_{-d/2}^{d/2} p dx_3 \\ &= -b \frac{1}{d} \int_{-d/2}^{d/2} \left[ \frac{\partial u_1^f}{\partial x_1} + \frac{\partial u_2^f}{\partial x_2} + \frac{\partial u_3^f}{\partial x_3} \right] dx_3. \end{aligned} \quad (185)$$

If we define

$$P = \frac{1}{d} \int_{-d/2}^{d/2} p dx_3 \quad (186)$$

and

$$U_i^f = \frac{1}{d} \int_{-d/2}^{d/2} u_i^f dx_3 \quad (i = 1, 2), \quad (187)$$

Eq. (185) is written as

$$P = -b \left( \frac{\partial U_1^f}{\partial x_1} + \frac{\partial U_2^f}{\partial x_2} + \frac{1}{d} \int_{-d/2}^{d/2} \frac{\partial u_3^f}{\partial x_3} dx_3 \right). \quad (188)$$

Since  $u_3^f$  is symmetric with respect to  $x_3(u_3^f(-x_3) = -u_3^f(x_3))$ , we find that

$$\int_{-d/2}^{d/2} \frac{\partial u_3^f}{\partial x_3} dx_3 = \int_{-d/2}^{d/2} du_3^f = [u_3^f]_{-d/2}^{d/2} = 2u_3^f(d/2). \tag{189}$$

Note that  $u_3^f(d/2) = u_3(d/2) = u_d$  from the boundary condition of Eq. (184) and, therefore, we have

$$P = -b \left( \frac{\partial U_1^f}{\partial x_1} + \frac{\partial U_2^f}{\partial x_2} + \frac{2}{d} u_d \right). \tag{190}$$

From Eq. (182) we obtain

$$\rho_f \frac{\partial^2 U_1^f}{\partial t^2} = -\frac{\partial P}{\partial x_1}, \tag{191}$$

$$\rho_f \frac{\partial^2 U_2^f}{\partial t^2} = -\frac{\partial P}{\partial x_2}. \tag{192}$$

Using the fluid velocity, Eqs. (191), (192), and (190) can be equivalently written as

$$\rho_f \frac{\partial V_1^f}{\partial t} = -\frac{\partial P}{\partial x_1}, \tag{193}$$

$$\rho_f \frac{\partial V_2^f}{\partial t} = -\frac{\partial P}{\partial x_2}, \tag{194}$$

$$\frac{\partial P}{\partial t} = -b \left( \frac{\partial V_1^f}{\partial x_1} + \frac{\partial V_2^f}{\partial x_2} + \frac{2}{d} v_d \right), \tag{195}$$

where  $V_i^f = \partial U_i^f / \partial t$  ( $i = 1, 2$ ) and  $v_d$  is the normal velocity of the crack wall.

Thus, the three-dimensional equations of motion and continuity are reduced to two-dimensional equations with respect to  $V_1^f, V_2^f, P$ , and  $v_d$ . In this problem, the boundary conditions are

$$\sigma_{33} = -p, \quad \sigma_{13} = \sigma_{23} = 0, \quad \text{on } S_1, \tag{196}$$

$$u_3 = 0, \quad \sigma_{13} = \sigma_{23} = 0, \quad \text{on } S_2, \tag{197}$$

where  $S_1$  denotes the crack surface and  $S_2$  denotes the crack plane outside the crack. The latter condition expresses that the displacement in the solid is symmetric with respect to  $x_3$ . In addition to these boundary conditions, Chouet [22] assumed zero mass transfer in and out of the crack, which are as follows:

$$V_1^f = v_1 \quad \text{along } |x_1| = W/2, 0 \leq x_2 \leq L, \tag{198}$$

$$V_2^f = v_2 \quad \text{along } -W/2 \leq x_1 \leq W/2, x_2 = 0, x_2 = L. \tag{199}$$

The equations of motion and stress-strain relationship for the solid in the form of Eqs. (180) and (181) and the equations of motion and continuity for the fluid in the forms of Eqs. (193), (194) and (195) are solved under the boundary conditions of Eqs. (196), (197), (198), and (199). Chouet [22] solved this problem by using a finite-difference method, in which the following scaling was used:

$$\bar{x}_i = x_i/L, \tag{200}$$

$$\bar{t} = \alpha t/L, \tag{201}$$

$$\bar{\sigma}_{ij} = \sigma_{ij}/\sigma_0, \tag{202}$$

$$\bar{u}_i = u_i \mu / (L \sigma_0), \tag{203}$$

$$\bar{v}_i = v_i \mu / (\alpha \sigma_0), \tag{204}$$

where nondimensional variables are indicated by a bar,  $\sigma_0$  is an effective stress, and  $\alpha$  is the compressional wave velocity  $\alpha = \sqrt{(\lambda + 2\mu)/\rho_s}$ . Let us define the first-order difference operators in time and space as:

$$\Delta_t f(l; i, j, k) = f(l + 1; i, j, k) - f(l; i, j, k), \tag{205}$$

$$\Delta_{x1} f(l; i, j, k) = f(l; i + 1, j, k) - f(l; i, j, k), \tag{206}$$

$$\Delta_{x2} f(l; i, j, k) = f(l; i, j + 1, k) - f(l; i, j, k), \tag{207}$$

$$\Delta_{x3} f(l; i, j, k) = f(l; i, j, k + 1) - f(l; i, j, k), \tag{208}$$

where  $f(l; i, j, k)$  represents any function at time  $l\Delta t$  and position  $[i\Delta, j\Delta, k\Delta]$  or  $[i\Delta_a, j\Delta_a, k\Delta_a]$ . Here,  $\Delta t$  is the time increment and  $\Delta$  and  $\Delta_a$  are the grid spacings for the solid and fluid, respectively. A difference form of Eq. (195) is then obtained as

$$\frac{\Delta_t P}{\Delta t} = -b \left( \frac{\Delta_{x1} V_1^f}{\Delta_a} + \frac{\Delta_{x2} V_2^f}{\Delta_a} + \frac{2}{d} v_d \right) \tag{209}$$

or

$$\Delta_t P = -b H_a \left( \Delta_{x1} V_1^f + \Delta_{x2} V_2^f \right) - 2 \frac{b}{d} \Delta t v_d, \tag{210}$$

where  $H_a = \Delta t / \Delta_a$ . Using the scaling given by Eqs. (200) to (204), Eq. (210) can be further modified to a nondimensional form:

$$\Delta_t \bar{P} = - \left( \frac{b}{\mu} \right) \alpha H_a \left( \Delta_{x1} \bar{V}_1^f + \Delta_{x2} \bar{V}_2^f \right) - 2C \Delta \bar{t} \bar{v}_d, \tag{211}$$

where  $C = (b/\mu)(L/d)$  is called the crack stiffness [2]. In a similar manner, we obtain nondimensional difference forms of equations of motion for the fluid (193) and (194) as

$$\Delta_t \bar{V}_1^f = - \left( \frac{a}{\alpha} \right)^2 \left( \frac{\mu}{b} \right) \alpha H_a \Delta_{x1} \bar{P}, \tag{212}$$



$$\Delta_t \bar{v}_2^f = -\left(\frac{a}{\alpha}\right)^2 \left(\frac{\mu}{b}\right) \alpha H_a \Delta_{x2} \bar{P}, \quad (213)$$

where  $a = \sqrt{b/\rho_f}$  is the acoustic velocity of the fluid and the ratios  $\alpha/a$  and  $b/\mu$  are related to  $\rho_f$  and  $\rho_s$  through the relationship

$$\frac{\rho_f}{\rho_s} = \left(\frac{\mu}{\lambda + 2\mu}\right) \left(\frac{\alpha}{a}\right)^2 \left(\frac{b}{\mu}\right). \quad (214)$$

Nondimensional difference forms of the equations of motion and the stress-strain relationship for the solid (Eqs. (180) and (181)) are

$$\Delta_t \bar{v}_i = \frac{1}{2 + (\lambda/\mu)} \alpha H \Delta_{xi} \bar{\sigma}_{ij}, \quad (215)$$

$$\Delta_t \bar{\sigma}_{ij} = \alpha H \left[ \frac{\lambda}{\mu} \Delta_{xk} \bar{v}_k \delta_{ij} + (\Delta_{xi} \bar{v}_i + \Delta_{xj} \bar{v}_j) \right], \quad (216)$$

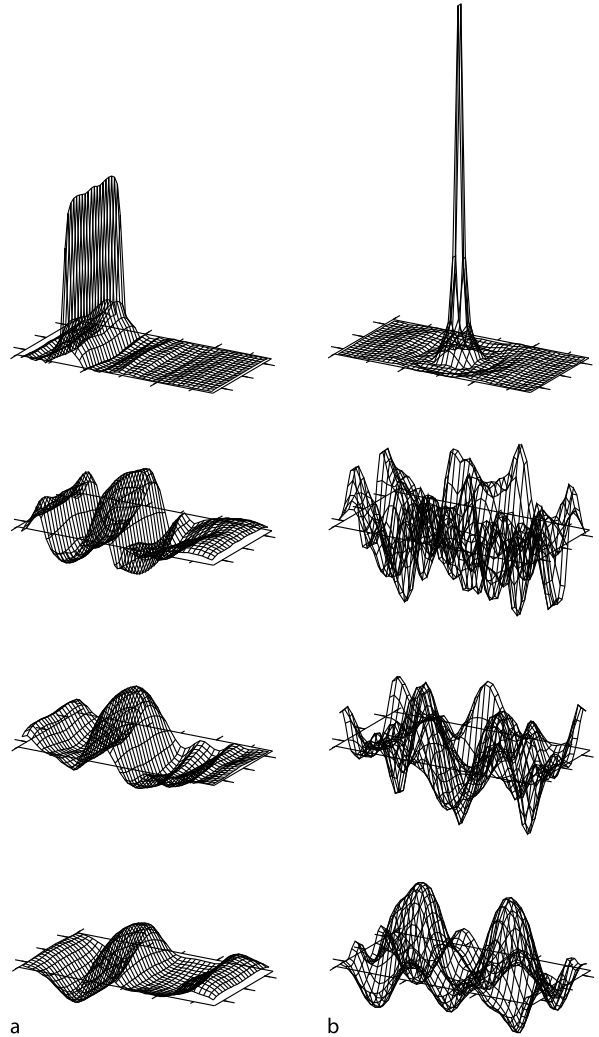
where  $H = \Delta t/\Delta$ . To satisfy the conditions for numerical stability in the solid and fluid, Chouet [22] used the following relationships:

$$H = \frac{\Delta t}{\Delta} = \frac{1}{4\alpha}, \quad (217)$$

$$H_a = \frac{\Delta t}{\Delta_a} = \frac{1}{2a}. \quad (218)$$

Applying a step increase in the fluid pressure over an area of the crack, the nondimensional difference equations for the solid and fluid are solved by a finite-difference method with a staggered grid scheme [129] to determine the normal velocity of the crack wall  $v_d$ . This source space-time function is used to synthesize the ground response to the fluid-filled crack. Note that Chouet [22] added an artificial dissipation term to the equations for particle velocities in the solid (Eq. (215)) and for fluid velocities (Eqs. (212) and (213)) to suppress spurious numerical oscillations associated with discretization in time and space in the finite difference calculations.

Snapshots of the normal velocity of the crack wall  $v_d$  for a fluid-filled crack excited by a step increase in the pressure applied over a strip extending over the entire width of the crack are shown in Fig. 19a. The results show that the crack oscillation is dominated by the longitudinal modes with wavelengths  $L$  and  $2L/5$ . Figure 19b shows snapshots of  $v_d$  for the crack with the excitation applied over a small area at the center of the crack. The crack oscillation shown in Fig. 19b displays more complex spatial patterns than those of Fig. 19a, in which both the longitudinal and transverse modes with wavelengths  $2L/3$  and  $2W/3$ , respectively, are dominantly excited. Chouet [22] showed that the dimensionless phase velocity  $v/a$  of the crack wave



Volcano Seismic Signals, Source Quantification of, Figure 19 **a** Snapshots of the normal velocity of the crack wall  $v_d$  for a fluid-filled crack excited by a step increase in the pressure applied over a strip extending over the entire width of the crack. **b** Snapshots of  $v_d$  for the crack with the excitation applied over a small area at the center of the crack (modified from Nakano et al. [95])

sustaining the crack resonance depends on the crack stiffness  $C$ . The  $Q$  factor of the crack resonance almost monotonically increases with the increasing  $\alpha/a$  [67,68].

The slow wave and its dependence on the crack stiffness may be understood by the following simple theoretical consideration. Using Eqs. (190), (191), and (192), we obtain the following relationship:

$$\frac{\partial^2}{\partial t^2} \left( P + \frac{2b}{d} u_d \right) = \left( \frac{b}{\rho_f} \right) \left( \frac{\partial^2 P}{\partial x_1^2} + \frac{\partial^2 P}{\partial x_2^2} \right). \quad (219)$$

Note that  $a = \sqrt{b/\rho_f}$  is the acoustic velocity of the fluid.

Let us consider a simple case for which  $u_d$  is proportional to  $P$ :

$$u_d = \epsilon P, \quad (220)$$

where  $\epsilon$  is a proportionality constant. Using the scaling given in Eqs. (202) and (203), the constant  $\epsilon$  may be written in a nondimensional form as

$$\epsilon = \bar{\epsilon} L / \mu, \quad (221)$$

where  $\bar{\epsilon}$  is a nondimensional proportionality constant. Using Eqs. (220) and (221), Eq. (219) can be written as

$$\left(1 + 2\bar{\epsilon} \frac{bL}{\mu d}\right) \frac{\partial^2 P}{\partial t^2} = \left(\frac{b}{\rho_f}\right) \left(\frac{\partial^2 P}{\partial x_1^2} + \frac{\partial^2 P}{\partial x_2^2}\right) \quad (222)$$

or, equivalently,

$$\frac{\partial^2 P}{\partial t^2} = \left(\frac{b_e}{\rho_f}\right) \left(\frac{\partial^2 P}{\partial x_1^2} + \frac{\partial^2 P}{\partial x_2^2}\right), \quad (223)$$

where  $b_e$  is an effective bulk modulus defined as

$$b_e = b / (1 + 2\bar{\epsilon} C). \quad (224)$$

Here,  $C = (b/\mu)(L/d)$  is the crack stiffness. Equation (223) is the two-dimensional wave equation with phase velocity  $a_e = \sqrt{b_e/\rho_f}$ . If no deformation of the crack wall occurs ( $\bar{\epsilon} = 0$ ), the phase velocity  $a_e$  is equal to the fluid acoustic velocity  $a$ . If the deformation of the crack wall occurs such that the wall moves inward to increase fluid pressure ( $\bar{\epsilon} < 0$ ),  $a_e$  becomes smaller than  $a$  through a reduction of the effective bulk modulus, which depends on the crack stiffness as given in Eq. (224).

The slow crack wave leads to more realistic estimates of the size and volume of a fluid-filled resonator as compared to a resonator with spherical geometry e. g. [35,40,41,115]. The slow wave has been used to interpret spectral characteristics of LP and VLP events observed at various active volcanoes [29,44,52,65,66,67,70,72,73,85,88,91,98,117,131].

### Future Directions

Recent volcano seismological studies have contributed greatly to achieving a better understanding of magmatic and hydrothermal systems. It is clear that the theoretical development of source models and analysis techniques has played an essential role in recent advances. The current studies in volcano seismology may be summarized as follows: (a) physical understanding and modeling of source dynamics of LP and VLP events [34,99,100,125], (b) development of a waveform inversion method of volcano-seis-

mic signals for an extended source to achieve a better understanding of the resonance characteristics and triggering mechanisms [95], (c) extension of the crack model to incorporate more realistic fluid acoustic properties [65,131], and (d) dense broadband seismic observation and the use of waveform inversion techniques to monitor active volcanoes e. g., [11,37,74]. Future studies building on these will contribute further to our ability to quantify the sources of volcano-seismic signals, which will help us in our efforts to predict eruptive behavior and thus mitigate volcanic hazards through improved seismic monitoring of magmatic and hydrothermal activity.

### Acknowledgment

I am deeply grateful to Masaru Nakano for numerous discussions on all the subjects presented in this manuscript. I thank Yasuko Takei for constructive comments on the phenomenological source representation. Comments from Pablo Palacios, Luca D'Auria, Takeshi Nishimura, and an anonymous reviewer helped improve the manuscript. I used the Generic Mapping Tools (GMT) [130] in the preparation of figures.

### Appendix A: Green's Functions

Here, I briefly explain the relationship between the displacement and Green's functions. To simplify the explanation, I use two-dimensional equations in an infinite medium. The extension of the equations into three-dimension is straightforward. Green's functions defined by Eq. (5) are explicitly written as

$$\rho \frac{\partial^2 G_{11}}{\partial t^2} = (\lambda + 2\mu) \frac{\partial^2 G_{11}}{\partial x_1^2} + (\lambda + \mu) \frac{\partial^2 G_{21}}{\partial x_2 \partial x_1} + \mu \frac{\partial^2 G_{11}}{\partial x_2^2} + \delta(\mathbf{x} - \boldsymbol{\eta}) \delta(t - \tau), \quad (A1)$$

$$\rho \frac{\partial^2 G_{21}}{\partial t^2} = (\lambda + 2\mu) \frac{\partial^2 G_{21}}{\partial x_2^2} + (\lambda + \mu) \frac{\partial^2 G_{11}}{\partial x_2 \partial x_1} + \mu \frac{\partial^2 G_{21}}{\partial x_1^2}, \quad (A2)$$

$$\rho \frac{\partial^2 G_{12}}{\partial t^2} = (\lambda + 2\mu) \frac{\partial^2 G_{12}}{\partial x_1^2} + (\lambda + \mu) \frac{\partial^2 G_{22}}{\partial x_2 \partial x_1} + \mu \frac{\partial^2 G_{12}}{\partial x_2^2}, \quad (A3)$$

$$\rho \frac{\partial^2 G_{22}}{\partial t^2} = (\lambda + 2\mu) \frac{\partial^2 G_{22}}{\partial x_2^2} + (\lambda + \mu) \frac{\partial^2 G_{12}}{\partial x_2 \partial x_1} + \mu \frac{\partial^2 G_{22}}{\partial x_1^2} + \delta(\mathbf{x} - \boldsymbol{\eta}) \delta(t - \tau). \quad (A4)$$

Note that  $(G_{11}, G_{21})$  represents the  $x_1$  and  $x_2$  components of the wavefield at  $(\mathbf{x}, t)$ , which is excited by the impulse in the  $x_1$  direction applied at  $\mathbf{x} = \boldsymbol{\eta}$  and  $t = \tau$ . Similarly,  $(G_{12}, G_{22})$  represents the  $x_1$  and  $x_2$  components of the wavefield excited by the impulse in the  $x_2$  direction at  $\mathbf{x} = \boldsymbol{\eta}$  and at  $t = \tau$ . Therefore,  $i$  and  $j$  in  $G_{ij}$  represent the component and direction of the impulse, respectively. To specify the relationship between the receiver and source, we use the notation  $G_{ij}(\mathbf{x}, t; \boldsymbol{\eta}, \tau)$ . We obtain  $G_{ij}(\mathbf{x}, t; \boldsymbol{\eta}, \tau) = G_{ij}(\mathbf{x}, t - \tau; \boldsymbol{\eta}, 0)$ , since Green's functions are independent of the time of origin.

The displacement  $u_i(\mathbf{x}, t)$  satisfies Eq. (3), which is explicitly written as

$$\rho \frac{\partial^2 u_1}{\partial t^2} = (\lambda + 2\mu) \frac{\partial^2 u_1}{\partial x_1^2} + (\lambda + \mu) \frac{\partial^2 u_2}{\partial x_1 \partial x_2} + \mu \frac{\partial^2 u_1}{\partial x_2^2} + f_1^S(\mathbf{x}, t), \quad (\text{A5})$$

$$\rho \frac{\partial^2 u_2}{\partial t^2} = (\lambda + 2\mu) \frac{\partial^2 u_2}{\partial x_2^2} + (\lambda + \mu) \frac{\partial^2 u_1}{\partial x_1 \partial x_2} + \mu \frac{\partial^2 u_2}{\partial x_1^2} + f_2^S(\mathbf{x}, t). \quad (\text{A6})$$

Equation (6) indicates that  $u_i$  is described by the following relationships:

$$u_1(\mathbf{x}, t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [f_1^S(\boldsymbol{\eta}, \tau) G_{11}(\mathbf{x}, t - \tau; \boldsymbol{\eta}, 0) + f_2^S(\boldsymbol{\eta}, \tau) G_{12}(\mathbf{x}, t - \tau; \boldsymbol{\eta}, 0)] d\eta_1 d\eta_2 d\tau, \quad (\text{A7})$$

$$u_2(\mathbf{x}, t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [f_1^S(\boldsymbol{\eta}, \tau) G_{21}(\mathbf{x}, t - \tau; \boldsymbol{\eta}, 0) + f_2^S(\boldsymbol{\eta}, \tau) G_{22}(\mathbf{x}, t - \tau; \boldsymbol{\eta}, 0)] d\eta_1 d\eta_2 d\tau. \quad (\text{A8})$$

We can verify that the displacement given by Eqs. (A7) and (A8) satisfies Eqs. (A5) and (A6) in the following way. We can rewrite Eqs. (A5) and (A6) as

$$f_1^S(\mathbf{x}, t) = \rho \frac{\partial^2 u_1}{\partial t^2} - (\lambda + 2\mu) \frac{\partial^2 u_1}{\partial x_1^2} - (\lambda + \mu) \frac{\partial^2 u_2}{\partial x_1 \partial x_2} - \mu \frac{\partial^2 u_1}{\partial x_2^2}, \quad (\text{A9})$$

$$f_2^S(\mathbf{x}, t) = \rho \frac{\partial^2 u_2}{\partial t^2} - (\lambda + 2\mu) \frac{\partial^2 u_2}{\partial x_2^2} - (\lambda + \mu) \frac{\partial^2 u_1}{\partial x_1 \partial x_2} - \mu \frac{\partial^2 u_2}{\partial x_1^2}. \quad (\text{A10})$$

We denote the right-hand side of Eq. (A9) as  $L_1$ , which is derived from Eqs. (A7) and (A8) as

$$L_1 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left[ \rho \left( f_1^S \frac{\partial^2 G_{11}}{\partial t^2} + f_2^S \frac{\partial^2 G_{12}}{\partial t^2} \right) - (\lambda + 2\mu) \left( f_1^S \frac{\partial^2 G_{11}}{\partial x_1^2} + f_2^S \frac{\partial^2 G_{12}}{\partial x_1^2} \right) - (\lambda + \mu) \left( f_1^S \frac{\partial^2 G_{21}}{\partial x_1 \partial x_2} + f_2^S \frac{\partial^2 G_{22}}{\partial x_1 \partial x_2} \right) - \mu \left( f_1^S \frac{\partial^2 G_{11}}{\partial x_2^2} + f_2^S \frac{\partial^2 G_{12}}{\partial x_2^2} \right) \right] d\eta_1 d\eta_2 d\tau. \quad (\text{A11})$$

This equation can be modified as follows:

$$L_1 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left[ f_1^S \left\{ \rho \frac{\partial^2 G_{11}}{\partial t^2} - (\lambda + 2\mu) \frac{\partial^2 G_{11}}{\partial x_1^2} - (\lambda + \mu) \frac{\partial^2 G_{21}}{\partial x_1 \partial x_2} - \mu \frac{\partial^2 G_{11}}{\partial x_2^2} \right\} + f_2^S \left\{ \rho \frac{\partial^2 G_{12}}{\partial t^2} - (\lambda + 2\mu) \frac{\partial^2 G_{12}}{\partial x_1^2} - (\lambda + \mu) \frac{\partial^2 G_{22}}{\partial x_1 \partial x_2} - \mu \frac{\partial^2 G_{12}}{\partial x_2^2} \right\} \right] d\eta_1 d\eta_2 d\tau. \quad (\text{A12})$$

From Eq. (A3), we find that the second integral of Eq. (A12) is zero. Therefore, Eq. (A12) from Eq. (A1) becomes

$$L_1 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_1^S(\boldsymbol{\eta}, \tau) \delta(\mathbf{x} - \boldsymbol{\eta}) \delta(t - \tau) d\eta_1 d\eta_2 d\tau = f_1^S(\mathbf{x}, t). \quad (\text{A13})$$

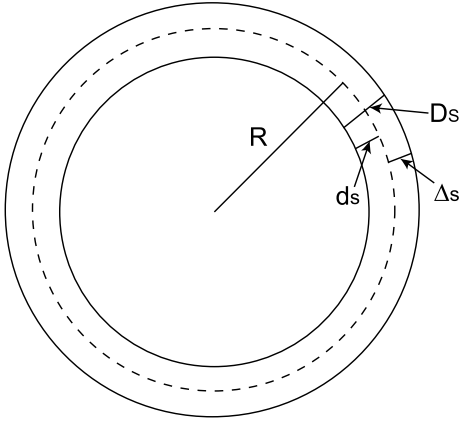
Similarly, we can show that the right-hand side of Eq. (A10) is equivalent to  $f_2^S(\mathbf{x}, t)$ , and thus the displacement in the forms of Eqs. (A7) and (A8) satisfies the equation of motion (A5) and (A6).

## Appendix B: Moment Tensor for a Spherical Source

Let us consider the moment density tensors for three tensile cracks in the planes  $\xi_1 = 0$ ,  $\xi_2 = 0$ , and  $\xi_3 = 0$ . If we sum and average these three tensors and assume that  $[u_1] = [u_2] = [u_3] = D_s$ , we obtain the following moment density tensor:

$$\mathbf{m} = D_s \begin{pmatrix} \lambda + 2\mu/3 & 0 & 0 \\ 0 & \lambda + 2\mu/3 & 0 \\ 0 & 0 & \lambda + 2\mu/3 \end{pmatrix}. \quad (\text{B1})$$

This represents the moment density tensor for the isotropic expansion of a cubic element. Following Müller [89], we consider a spherical crack surface of radius  $R$



Volcano Seismic Signals, Source Quantification of, Figure A1  
**A spherical crack surface of radius  $R$  where a constant radial expansion  $D_s = (d_s + \Delta_s)$  occurs. Here, the inner wall of the crack moves inward by  $d_s$  and the outer wall moves outward by  $\Delta_s$  [89]**

where a constant radial expansion  $D_s = (d_s + \Delta_s)$  occurs (Fig. A1). Here, the inner wall of the crack moves inward by  $d_s$  and the outer wall moves outward by  $\Delta_s$ . The moment tensor of the spherical expansion may be obtained by integration of the moment density tensor (B1) over the surface  $R$ :

$$\mathbf{M} = \Delta V \begin{pmatrix} \lambda + 2\mu/3 & 0 & 0 \\ 0 & \lambda + 2\mu/3 & 0 \\ 0 & 0 & \lambda + 2\mu/3 \end{pmatrix}, \quad (\text{B2})$$

where  $\Delta V$  is the volume given as

$$\Delta V = 4\pi R^2 D_s = 4\pi R^2 (d_s + \Delta_s). \quad (\text{B3})$$

The volume  $4\pi R^2 d_s$  is caused by the inward motion, which compresses the sphere. The volume  $\Delta V_s = 4\pi R^2 \Delta_s$ , on the other hand, is caused by the outward motion, which excites seismic waves in the region outside the sphere.  $\Delta V_s$  can be determined by solving an elastostatic boundary-value problem in the following way. We assume an isotropic medium, and denote the regions inside and outside the sphere as regions 1 and 2, respectively. Since the motion is radial only, the equation of motion is given as e. g., [121]

$$\rho \frac{\partial^2 u}{\partial t^2} = \frac{1}{r^2} \frac{\partial}{\partial r} (r^2 \sigma_{rr}) - \frac{1}{r} (\sigma_{\theta\theta} + \sigma_{\phi\phi}) \quad (\text{B4})$$

and

$$\sigma_{rr} = (\lambda + 2\mu) \frac{\partial u}{\partial r} + \lambda \frac{2}{r} u, \quad (\text{B5})$$

$$\sigma_{\theta\theta} = \sigma_{\phi\phi} = \lambda \frac{\partial u}{\partial r} + (\lambda + \mu) \frac{2}{r} u, \quad (\text{B6})$$

where  $u$  is the radial displacement and  $\sigma_{rr}$ ,  $\sigma_{\theta\theta}$ , and  $\sigma_{\phi\phi}$  are the stress components in the spherical coordinate. Equations (B4) and (B5) apply to both regions 1 and 2. Substituting Eq. (B5) into Eq. (B4) and setting  $\rho(\partial^2 u_r / \partial t^2) = 0$ , we obtain the following static equilibrium equation:

$$\frac{\partial^2 u}{\partial r^2} + \frac{2}{r} \frac{\partial u}{\partial r} - \frac{2}{r^2} u = 0. \quad (\text{B7})$$

This equation has two solutions:  $u = ar$  and  $u = b/r^2$ , where  $a$  and  $b$  are constants. The former is the interior solution for region 1 ( $r \leq R$ ), and the latter is the exterior solution for region 2 ( $r \geq R$ ). I denote the interior and exterior solutions as  $u_i$  and  $u_e$ , respectively. The constants  $a$  and  $b$  are determined by the boundary conditions for the radial displacement and the continuity of the radial stress at  $r = R$ :

$$u_e(R) - u_i(R) = \frac{b}{R^2} - aR = D_s, \quad (\text{B8})$$

$$\sigma_{rr}^i(R) - \sigma_{rr}^e(R) = -\frac{4\mu}{R^3} b - (3\lambda + 2\mu)a = 0, \quad (\text{B9})$$

where  $\sigma_{rr}^i$  and  $\sigma_{rr}^e$  are the radial stresses in regions 1 and 2, respectively. Accordingly, we obtain

$$u_i(r) = -\frac{4\mu D_s}{3(\lambda + 2\mu)} \frac{r}{R} \quad (r \leq R), \quad (\text{B10})$$

$$u_e(r) = \frac{(\lambda + 2\mu/3) D_s}{\lambda + 2\mu} \frac{R^2}{r^2} \quad (r \geq R). \quad (\text{B11})$$

Then, we obtain

$$d_s = -u_i(R) = \frac{4\mu D_s}{3(\lambda + 2\mu)}, \quad (\text{B12})$$

$$\Delta_s = u_e(R) = \frac{(\lambda + 2\mu/3) D_s}{\lambda + 2\mu}. \quad (\text{B13})$$

Equation (B3) can be modified as

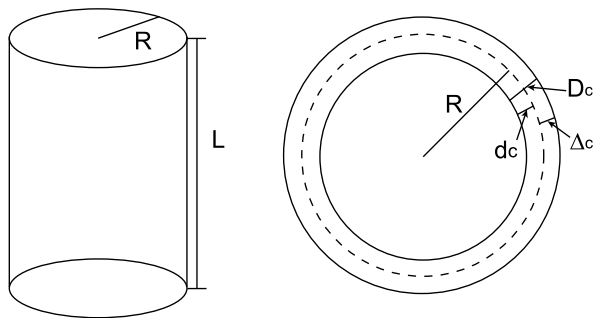
$$\Delta V = 4\pi R^2 \Delta_s (D_s / \Delta_s) = \frac{\lambda + 2\mu}{\lambda + 2\mu/3} \Delta V_s. \quad (\text{B14})$$

Finally, we obtain the moment tensor for the spherical expansion as

$$\mathbf{M} = (\lambda + 2\mu) \Delta V_s \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \quad (\text{B15})$$

### Appendix C: Moment Tensor for a Cylindrical Source

We consider the moment density tensors for two tensile cracks in the planes  $\xi_1 = 0$  and  $\xi_2 = 0$ , where  $\xi_1$  and  $\xi_2$



**Volcano Seismic Signals, Source Quantification of, Figure A2**  
**A vertical cylinder of length  $L$  and radius  $R$ . The cylinder surface can be regarded as a cylindrical crack, where the radial expansion  $D_c = (d_c + \Delta_c)$  occurs. Here, the inner wall of the crack moves inward by  $d_c$  and the outer wall moves outward by  $\Delta_c$  [89]**

are two horizontal axes. If we sum and average these two tensors and assume that  $[u_1] = [u_2] = D_c$ , we obtain

$$m = D_c \begin{pmatrix} \lambda + \mu & 0 & 0 \\ 0 & \lambda + \mu & 0 \\ 0 & 0 & \lambda \end{pmatrix}. \tag{C1}$$

This represents the moment density tensor for the expansion of a cubic element in the two horizontal directions. Let us consider a vertical cylinder of length  $L$  and radius  $R$ . The cylinder surface at radius  $R$  can be regarded as a cylindrical crack, where the radial expansion  $D_c = (d_c + \Delta_c)$  occurs (Fig. A2). The moment tensor for the radial expansion of the cylinder may be obtained by integration of the moment density tensor (C1) over the surface  $R$ :

$$M = \Delta V \begin{pmatrix} \lambda + \mu & 0 & 0 \\ 0 & \lambda + \mu & 0 \\ 0 & 0 & \lambda \end{pmatrix}, \tag{C2}$$

where  $\Delta V$  is the volume given as

$$\Delta V = 2\pi R L D_c = 2\pi R L (d_c + \Delta_c). \tag{C3}$$

We obtain the static equilibrium equation for the radial expansion of the cylinder  $u$  as

$$\frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} - \frac{1}{r^2} u = 0. \tag{C4}$$

This equation has internal and external solutions, which are given as  $u = ar$  and  $u = b/r$ , respectively. The constants  $a$  and  $b$  are determined by the boundary conditions

$$u_e(R) - u_i(R) = \frac{b}{R} - aR = D_c, \tag{C5}$$

$$\sigma_{rr}^i(R) - \sigma_{rr}^e(R) = -\frac{2\mu}{R^2} b - 2(\lambda + \mu)a = 0, \tag{C6}$$

where superscripts  $i$  and  $e$  denote the interior and exterior solutions. We then obtain

$$d_c = -u_i(R) = \frac{\mu D_s}{(\lambda + 2\mu)}, \tag{C7}$$

$$\Delta_c = u_e(R) = \frac{(\lambda + \mu) D_c}{\lambda + 2\mu}, \tag{C8}$$

and

$$\Delta V = \frac{\lambda + 2\mu}{\lambda + \mu} \Delta V_c. \tag{C9}$$

The moment tensor for the vertical cylinder is therefore given as

$$M = \frac{\lambda + 2\mu}{\lambda + \mu} \Delta V_c \begin{pmatrix} \lambda + \mu & 0 & 0 \\ 0 & \lambda + \mu & 0 \\ 0 & 0 & \lambda \end{pmatrix}. \tag{C10}$$

Let us rotate the vertical cylinder with axis orientation angles  $\phi$  and  $\theta$  (Fig. 8b). This can be done in the following steps: (1) fix the cylinder and (2) rotate the  $\xi_1$  and  $\xi_2$  axes around the  $\xi_3$  axis through an angle  $-\phi$ , and (3) further rotate the  $\xi_1$  and  $\xi_3$  axes around the  $\xi_2$  axis through an angle  $-\theta$ . The rotation matrix  $R$  is given as

$$R = \begin{pmatrix} \cos \theta & 0 & -\sin \theta \\ 0 & 1 & 0 \\ \sin \theta & 0 & \cos \theta \end{pmatrix} \begin{pmatrix} \cos \phi & -\sin \phi & 0 \\ \sin \phi & \cos \phi & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{C11}$$

$$= \begin{pmatrix} \cos \theta \cos \phi & -\cos \theta \sin \phi & -\sin \theta \\ \sin \phi & \cos \phi & 0 \\ \sin \theta & -\sin \theta \sin \phi & \cos \theta \end{pmatrix}. \tag{C12}$$

Using the matrix  $R$ , we obtain the moment tensor for a cylinder with the axis orientation angles  $(\phi, \theta)$  as

$$M' = R^T M R, \tag{C13}$$

which leads to Eq. (34).

## Bibliography

### Primary Literature

1. Aki K, Richards PG (2002) Quantitative seismology, 2nd ed. University Science Books, Sausalito
2. Aki K, Fehler M, Das S (1977) Source mechanism of volcanic tremor: Fluid-driven crack models and their application to the 1963 Kilauea eruption. *J Volcanol Geotherm Res* 2:259–287

3. Almendros J, Chouet BA, Dawson PB (2001) Spatial extent of a hydrothermal system at Kilauea Volcano, Hawaii, determined from array analyses of shallow long-period seismicity 1. Method. *J Geophys Res* 106:13565–13580
4. Almendros J, Chouet BA, Dawson PB (2001) Spatial extent of a hydrothermal system at Kilauea Volcano, Hawaii, determined from array analyses of shallow long-period seismicity 2. Results. *J Geophys Res* 106:13581–13597
5. Almendros J, Chouet BA, Dawson PB, Huber C (2002) Mapping the sources of the seismic wave field at Kilauea Volcano, Hawaii, using data recorded on multiple seismic antennas. *Bull Seism Soc Am* 92:2333–2351
6. Almendros J, Chouet BA, Dawson PB, Bond T (2002) Identifying elements of the plumbing system beneath Kilauea Volcano, Hawaii, from the source locations of very-long-period signals *Geophys J Int* 148:303–312
7. Aoyama H, Takeo M (2001) Wave properties and focal mechanisms of N-type earthquakes at Asama volcano. *J Volcanol Geotherm Res* 105:163–182
8. Arciniega-Ceballos A, Chouet BA, Dawson PB (1999) Very long-period signals associated with vulcanian explosions at Popocatepetl Volcano, Mexico. *Geophys Res Lett* 26:3013–3016
9. Arciniega-Ceballos A, Chouet BA, Dawson PB (2003) Long-period events and tremor at Popocatepetl volcano (1994–2000) and their broadband characteristics. *Bull Volcanol* 65:124–135
10. Aster R, Mah S, Kyle P, McIntosh W, Dunbar N, Johnson J, Ruiz M, McNamara S (2003) Very long period oscillations of Mount Erebus Volcano. *J Geophys Res* 108:2522. doi:10.1029/2002JB002101
11. Auger E, D'Auria L, Martini M, Chouet BA, Dawson PB (2006) Real-time monitoring and massive inversion of source parameters of Very-Long-Period (VLP) seismic signals: An application to Stromboli Volcano, Italy. *Geophys Res Lett* 33:L04301. doi:10.1029/2005GL024703
12. Backus G, Mulcahy M (1976) Moment tensors and other phenomenological descriptions of seismic sources—I. Continuous displacements. *Geophys J R Astr Soc* 46:341–361
13. Battaglia J, Aki K (2003) Location of seismic events and eruptive fissures on the Piton de la Fournaise volcano using seismic amplitudes. *J Geophys Res* 108:2364. doi:10.1029/2002JB002193
14. Biot MA (1952) Propagation of elastic waves in a cylindrical bore containing a fluid. *J Appl Phys* 23:997–1005
15. Biot MA (1952) The interaction of Rayleigh and Stoneley waves in the ocean bottom. *Bull Seism Soc Am* 42:82–92
16. Bouchon M (1979) Discrete wave number representation of elastic wave fields in three-space dimensions. *J Geophys Res* 84:3609–3614
17. Bouchon M (1981) A simple method to calculate Green's functions for elastic layered media. *Bull Seism Soc Am* 71:959–971
18. Bouchon M, Schultz CA, Toksöz MN (1996) Effect of 3D topography on seismic motion. *J Geophys Res* 101:5835–5846
19. Brandsdottir B, Einarsson P (1979) Seismic activity associated with the September 1977 deflation of the Krafla central volcano in northeastern Iceland. *J Volcanol Geotherm Res* 6:197–212
20. Chouet BA (1982) Free surface displacements in the near field of a tensile crack expanding in three dimensions. *J Geophys Res* 87:3868–3872
21. Chouet BA (1985) Excitation of a buried magmatic pipe: A seismic source model for volcanic tremor. *J Geophys Res* 90:1881–1893
22. Chouet BA (1986) Dynamics of a fluid-driven crack in three dimensions by the finite difference method. *J Geophys Res* 91:13967–13992
23. Chouet BA (1988) Resonance of a fluid-driven crack: Radiation properties and implications for the source of long-period events and harmonic tremor. *J Geophys Res* 93:4375–4400
24. Chouet BA (1992) A seismic model for the source of long-period events and harmonic tremor. In: Gasparini P, Scarpa R, Aki K (eds) *Volcanic Seismology*. Springer, Berlin, pp 133–156
25. Chouet BA (1996) New methods and future trends in seismological volcano monitoring. In: Scarpa R, Tilling RI (eds) *Monitoring and Mitigation of Volcano Hazards*. Springer, New York, pp 23–97
26. Chouet BA (1996) Long-period volcano seismicity: its source and use in eruption forecasting. *Nature* 380:309–316
27. Chouet BA (2003) *Volcano Seismology*. *Pure Appl Geophys* 160:739–788
28. Chouet BA, Julian B (1985) Dynamics of an expanding fluid-filled crack. *J Geophys Res* 90:11187–11198
29. Chouet BA, Page RA, Stephens CD, Lahr JC, Power JA (1994) Precursory swarms of long-period events at Redoubt Volcano (1989–1990), Alaska: Their origin and use as a forecasting tool. *J Volcanol Geotherm Res* 62:95–135
30. Chouet BA, Saccorotti G, Martini M, Dawson PB, De Luca G, Milana G, Scarpa R (1997) Source and path effects in the wavefields of tremor and explosions at Stromboli volcano, Italy. *J Geophys Res* 102:15129–15150
31. Chouet BA, Saccorotti G, Dawson PB, Martini M, Scarpa R, De Luca G, Milana G, Cattaneo M (1999) Broadband measurements of the sources of explosions at Stromboli Volcano, Italy. *Geophys Res Lett* 26:1937–1940
32. Chouet BA, Dawson PB, Ohminato T, Martini M, Saccorotti G, Guidicepiero F, De Luca G, Milana G, Scarpa R (2003) Source mechanisms of explosions at Stromboli Volcano, Italy, determined from moment-tensor inversions of very-long-period data. *J Geophys Res* 108:2331. doi:10.1029/2003JB002535
33. Chouet BA, Dawson PB, Arciniega-Ceballos A (2005) Source mechanism of Vulcanian degassing at Popocatepetl Volcano, Mexico, determined from waveform inversions of very long period signals. *J Geophys Res* 110:B07301. doi:10.1029/2004JB003524
34. Chouet BA, Dawson PB, Nakano M (2006) Dynamics of diffusive bubble growth and pressure recovery in a bubbly rhyolitic melt embedded in an elastic solid. *J Geophys Res* 111:B07310. doi:10.1029/2005JB004174
35. Crosson RS, Bame DA (1985) A spherical source model for low frequency volcanic earthquakes. *J Geophys Res* 90:10237–10247
36. Dawson PB, Dietel C, Chouet BA, Honma K, Ohminato T, Okubo P (1998) A digitally telemetered broadband seismic network at Kilauea Volcano, Hawaii. *US Geol Surv Open File Report* 98–108:1–121
37. Dawson PB, Whilldin D, Chouet BA (2004) Application of near real-time semblance to locate the shallow magmatic conduit at Kilauea Volcano, Hawaii. *Geophys Res Lett* 31:L21606. doi:10.1029/2004GL021163
38. De Angelis S, McNutt SR (2005) Degassing and hydrothermal activity at Mt. Spurr, Alaska during the summer of 2004 in-

- ferred from the complex frequencies of long-period events. *Geophys Res Lett* 32:L12312. doi:10.1029/2005GL022618
39. Ferrazzini V, Aki K (1987) Slow waves trapped in a fluid-filled infinite crack: Implication for volcanic tremor. *J Geophys Res* 92:9215–9223
  40. Fujita E, Ida Y (2003) Geometrical effects and low-attenuation resonance of volcanic fluid inclusions for the source mechanism of long-period earthquakes. *J Geophys Res* 108:2118. doi:10.1029/2002JB001806
  41. Fujita E, Ida Y, Oikawa J (1995) Eigen oscillation of a fluid sphere and source mechanism of harmonic volcanic tremor. *J Volcanol Geotherm Res* 69:365–378
  42. Fujita E, Ukawa M, Yamamoto E (2004) Subsurface cyclic magma sill expansions in the 2000 Miyakejima volcano eruption: Possibility of two-phase flow oscillation. *J Geophys Res* 109:B04205. doi:10.1029/2003JB002556
  43. Fukuyama E, Kubo A, Kawai H, Nonomura K (2001) Seismic remote monitoring of stress field. *Earth Planets Space* 53: 1021–1026
  44. Gil Cruz F, Chouet BA (1997) Long-period events, the most characteristic seismicity accompanying the emplacement and extrusion of a lava dome in Galeras Volcano, Colombia, in 1991. *J Volcanol Geotherm Res* 77:121–158
  45. Goldstein P, Chouet BA Array measurements and modeling of sources of shallow volcanic tremor at Kilauea volcano, Hawaii. *J Geophys Res* 99:2637–2652
  46. Hayashi Y, Morita Y (2003) An image of a magma intrusion process inferred from precise hypocenter migrations of the earthquake swarm east of the Izu Peninsula. *Geophys J Int* 153:159–174
  47. Hidayat D, Voight B, Langston C, Ratdomopurbo A, Ebeling C (2000) Broadband seismic experiment at Merapi volcano, Java, Indonesia: very-long-period pulses embedded in multiphase earthquakes. *J Volcanol Geotherm Res* 100: 215–231
  48. Hidayat D, Chouet BA, Voght B, Dawson P, Ratdomopurbo A (2002) Source mechanism of very-long-period signals accompanying dome growth activity at Merapi volcano, Indonesia. *Geophys Res Lett* 23:2118. doi:10.1029/2002GL015013
  49. Hill DP, Dawson PB, Johnston MJS, Pitt AM (2002) Very-long-period volcanic earthquakes beneath Mammoth Mountain, California. *Geophys Res Lett* 29:1370 doi:10.1029/2002GL014833
  50. Hori S, Fukao Y, Kumazawa M, Furumoto M, Yamamoto A (1989) A new method of spectral analysis and its application to the earth's free oscillations: the "Sompi" method. *J Geophys Res* 94:7535–7553
  51. Iguchi M (1994) A vertical expansion source model for the mechanisms of earthquakes originated in the magma conduit of an andesitic volcano: Sakurajima, Japan. *Bull Volcanol Soc Jpn* 39:49–67
  52. Jousset P, Neuberg JW, Sturton S (2003) Modelling the time-dependent frequency content of low-frequency volcanic earthquakes. *J Volcanol Geotherm Res* 128:201–223
  53. Julian BR (1994) Volcanic tremor: Nonlinear excitation by fluid flow. *J Geophys Res* 99:11859–11877
  54. Julian BR, Miller AD, Foulger GR (1998) Non-double-couple earthquakes 1. Theory. *Rev Geophys* 36:525–549
  55. Kanamori H, Given J (1982) Analysis of long-period seismic waves excited by the May 18, 1980, eruption of Mount St. Helens – A terrestrial monopole. *J Geophys Res* 87: 5422–5432
  56. Kanamori H, Given J, Lay T (1984) Analysis of seismic body waves excited by the Mount St. Helens eruption of May 18, 1980. *J Geophys Res* 89:1856–1866
  57. Kaneshima S et al (1996) Mechanism of phreatic eruptions at Aso volcano inferred from near-field broadband observations. *Science* 273:642–645
  58. Kawakatsu H, Ohminato T, Ito H, Kuwahara Y, Kato T, Tsuruga K, Honda S, Yomogida K (1992) Broadband seismic observation at Sakurajima Volcano, Japan. *Geophys Res Lett* 19: 1959–1962
  59. Kawakatsu H, Ohminato T, Ito H (1994) 10s-period volcanic tremors observed over a wide area in southwestern Japan. *Geophys Res Lett* 21:1963–1966
  60. Kawakatsu H et al (2000) Aso seismic observation with broadband instruments. *J Volcanol Geotherm Res* 101:129–154
  61. Kawakatsu H, Yamamoto M (2007) Volcano Seismology. In: Kamanori H (ed) *Treatise on Geophysics. Earthquake Seismology*, vol 4. Elsevier, Amsterdam, pp 389–420
  62. Kawase H (1988) Time domain response of a semi-circular canyon for incident SV, P and Rayleigh waves calculated by the discrete wavenumber boundary element method. *Bull Seism Soc Am* 78:1415–1437
  63. Kennet LN, Kerry NJ (1979) Seismic waves in a stratified half-space. *Geophys J R Astr Soc* 57:557–583
  64. Kobayashi T, Ohminato T, Ida Y (2003) Earthquakes series preceding very long period seismic signals observed during the 2000 Miyakejima volcanic activity. *Geophys Res Lett* 30:1423. doi:10.1029/2002GL016631
  65. Kumagai H (2006) Temporal evolution of a magmatic dike system inferred from the complex frequencies of very long period seismic events. *J Geophys Res* 111:B06201. doi:10.1029/2005JB003881
  66. Kumagai H, Chouet BA (1999) The complex frequencies of long-period seismic events as probes of fluid composition beneath volcanoes. *Geophys J Int* 138:F7–F12
  67. Kumagai H, Chouet BA (2000) Acoustic properties of a crack containing magmatic or hydrothermal fluids. *J Geophys Res* 105:25493–25512
  68. Kumagai H, Chouet BA (2001) The dependence of acoustic properties of a crack on the mode and geometry. *Geophys Res Lett* 28:3325–3328
  69. Kumagai H, Ohminato T, Nakano M, Ooi M, Kubo A, Inoue H, Oikawa J (2001) Very-long-period seismic signals and caldera formation at Miyake Island, Japan. *Science* 293:687–690
  70. Kumagai H, Chouet BA, Nakano M (2002) Temporal evolution of a hydrothermal system in Kusatsu-Shirane Volcano, Japan, inferred from the complex frequencies of long-period events. *J Geophys Res* 107:2236. doi:10.1029/2001JB000653
  71. Kumagai H, Chouet BA, Nakano M (2002) Waveform inversion of oscillatory signatures in long-period events beneath volcanoes. *J Geophys Res* 107:2301. doi:10.1029/2001JB001704
  72. Kumagai H, Miyakawa K, Negishi H, Inoue H, Obara K, Suet-sugu D (2003) Magmatic dike resonances inferred from very-long-period seismic signals. *Science* 299:2058–2061 (10.1126/science.1081195)
  73. Kumagai H, Chouet BA, Dawson PB (2005) Source process of a long-period event at Kilauea Volcano, Hawaii. *Geophys J Int* 161:243–254

74. Kumagai H et al (2007) Enhancing volcano-monitoring capabilities in Ecuador. *Eos trans AGU* 88:245–246
75. Kumazawa M, Imanishi Y, Fukao Y, Furumoto M, Yamamoto A (1990) A theory of spectral analysis based on the characteristic property of a linear dynamic system. *Geophys J Int* 101:613–630
76. Legrand D, Kaneshima S, Kawakatsu H (2000) Moment tensor analysis of near field broadband waveforms observed at Aso volcano, Japan. *J Volcanol Geotherm Res* 101:155–169
77. Lesage P, Glangeaud F, Mars J (2002) Applications of autoregressive models and time-frequency analysis to the study of volcanic tremor and long-period events. *J Volcanol Geotherm Res* 114:391–417
78. Lin CH, Konstantinou KI, Liang WT, Pu HC, Lin YM, You SH, Huang YP (2005) Preliminary analysis of volcanoseismic signals recorded at the Tatun Volcano Group, northern Taiwan. *Geophys Res Lett* 32:L10313. doi:10.1029/2005GL022861
79. Matsuura T, Imanishi Y, Imanari M, Kumazawa M (1990) Application of a new method of high-resolution spectral analysis, “Sompfi”, for free induction decay of nuclear magnetic resonance. *Appl Spectrosc* 44:618–626
80. McNutt SR (1996) Seismic monitoring and eruption forecasting of volcanoes: A review of the state-of-the art and cased histories. In: Scarpa R, Tilling RI (eds) *Monitoring and Mitigation of Volcano Hazards*. Springer, New York, pp 99–146
81. McNutt SR (2005) Volcanic Seismology. *Annu Rev Earth Planet Sci* 33:461–491
82. Métaxian JP, Lesage P, Dorel J (1997) Permanent tremor of Masaya volcano, Nicaragua: wavefield analysis and source location. *J Geophys Res* 102:22529–22545
83. Menke W (1989) *Geophysical data analysis: Discrete inverse theory*, Revised edition. Academic Press, San Diego
84. Minakami T (1974) Seismology of volcanoes in Japan. In: Civetta L, Gasparini P, Luongo G, Rapolla A (eds) *Physical Volcanology*. Elsevier, Amsterdam, pp 1–27
85. Molina I, Kumagai H, Yepes H (2004) Resonances of a volcanic conduit triggered by repetitive injections of an ash-laden gas. *Geophys Res Lett* 31:L03603. doi:10.1029/2003GL018934
86. Morita Y, Nakao S, Hayashi Y (2006) A quantitative approach to the dike intrusion process inferred from a joint analysis of geodetic and seismological data for the 1998 earthquake swarm off the east coast of Izu Peninsula, central Japan. *J Geophys Res* 111:B06208. doi:10.1029/2005JB003860
87. Morrissey M, Chouet BA (1997) A numerical investigation of choked flow dynamics and its application to the triggering mechanism of long-period events at Redoubt Volcano, Alaska. *J Geophys Res* 102:7965–7983
88. Morrissey M, Chouet BA (2001) Trends in long-period seismicity related to magmatic fluid compositions. *J Volcanol Geotherm Res* 108:265–281
89. Müller G (2001) Volume change of seismic sources from moment tensors. *Bull Seism Soc Am* 91:880–884
90. Nakamichi H, Hamaguchi H, Tanaka S, Ueki S, Nishimura T, Hasegawa A (2003) Source mechanisms of deep and intermediate-depth low-frequency earthquakes beneath Iwate volcano, northeastern Japan. *Geophys J Int* 154:811–828
91. Nakano M, Kumagai H (2005) Response of a hydrothermal system to magmatic heat inferred from temporal variations in the complex frequencies of long-period events at Kusatsu-Shirane Volcano, Japan. *J Volcanol Geotherm Res* 147:233–244
92. Nakano M, Kumagai H (2005) Waveform inversion of volcano-seismic signals assuming possible source geometries. *Geophys Res Lett* 32:L12302. doi:10.1029/2005GL022666
93. Nakano M, Kumagai H, Kumazawa M, Yamaoka K, Chouet BA (1998) The excitation and characteristic frequency of the long-period volcanic event: An approach based on an inhomogeneous autoregressive model of a linear dynamic system. *J Geophys Res* 103:10031–10046
94. Nakano M, Kumagai H, Chouet BA (2003) Source mechanism of long-period events at Kusatsu-Shirane Volcano, Japan, inferred from waveform inversion of the effective excitation functions. *J Volcanol Geotherm Res* 122:149–164
95. Nakano M, Kumagai H, Chouet BA, Dawson PB (2007) Waveform inversion of volcano-seismic signals for an extended source. *J Geophys Res* 112:B02306. doi:10.1029/2006JB004490
96. Neuberg JW (2000) Characteristics and causes of shallow seismicity in andesite volcanoes. *Philos Trans R Soc Lond A* 358:1533–1546
97. Neuberg JW, Luckett R, Ripepe M, Braun T (1994) Highlights from a seismic broadband array on Stromboli Volcano. *Geophys Res Lett* 21:749–752
98. Neuberg JW, Luckett R, Baptie B, Olsen K (2000) Models of tremor and low-frequency earthquake swarms on Montserrat. *J Volcanol Geotherm Res* 101:83–104
99. Neuberg JW, Tuffen H, Collier L, Green D, Powell T, Dingwell D (2006) The trigger mechanism of low frequency earthquakes on Montserrat. *J Volcanol Geotherm Res* 153:37–50
100. Nishimura T (2004) Pressure recovery in magma due to bubble growth. *Geophys Res Lett* 31:L12613. doi:10.1029/2004GL019810
101. Nishimura T, Chouet BA (2003) A numerical simulation of magma motion, crustal deformation, and seismic radiation associated with volcanic eruptions. *Geophys J Int* 153:699–718
102. Nishimura T, Nakamichi H, Tanaka S, Sato M, Kobayashi T, Ueki S, Hamaguchi H, Ohtake M, Sato H (2000) Source process of very long period seismic events associated with the 1998 activity of Iwate Volcano, northeastern Japan. *J Geophys Res* 105:19135–19417
103. Nishimura T, Ueki S, Yamawaki T, Tanaka S, Hashino H, Sato M, Nakamichi H, Hamaguchi H (2002) Broadband seismic signals associated with the 2000 volcanic unrest of Mount Bandai, northeastern Japan. *J Volcanol Geotherm Res* 119:51–59
104. O’Brien GS, Bean CJ (2004) A 3D discrete numerical elastic lattice method for seismic wave propagation in heterogeneous media with topography. *Geophys Res Lett* 31:L14608. doi:10.1029/2004GL020069
105. Ohminato T (2006) Characteristics and source modeling of broadband seismic signals associated with the hydrothermal system at Satsuma-Iwojima volcano, Japan. *J Volcanol Geotherm Res* 158:467–490
106. Ohminato T, Chouet BA (1997) A free-surface boundary condition for including 3D topography in the finite difference method. *Bull Seism Soc Am* 87:494–515
107. Ohminato T, Chouet BA, Dawson PB, Kedar S (1998) Waveform inversion of very long period impulsive signals associated with magmatic injection beneath Kilauea Volcano, Hawaii. *J Geophys Res* 103:23839–23862
108. Ohminato T, Takeo M, Kumagai H, Yamashina T, Oikawa J, Koyama E, Tsuji H, Urabe T (2006) Vulcanian eruptions



- with dominant single force components observed during the Asama 2004 volcanic activity in Japan. *Earth Planets Space* 58:583–593
109. Ripperger J, Igel H, Wasserman J (2003) Seismic wave simulation in the presence of real volcano topography. *J Volcanol Geotherm Res* 128:31–44
  110. Rowe CA, Aster RC, Kyle PR, Schlue JW, Dibble RR (1998) Broadband recording of Strombolian explosions and associated very-long-period seismic signals on Mount Erebus volcano, Ross Island, Antarctica. *Geophys Res Lett* 25: 2297–2300
  111. Rubin AM, Gillard D (1998) Dike-induced seismicity: theoretical considerations. *J Geophys Res* 103:10017–10030
  112. Rubin AM, Gillard D, Got JL (1998) A reinterpretation of seismicity associated with the January 1983 dike intrusion at Kilauea Volcano, Hawaii. *J Geophys Res* 103:10003–10015
  113. Saccorotti G, Del Pezzo E (2000) A probabilistic approach to the inversion of data from a seismic array and its application to volcanic signals. *Geophys J Int* 143:249–261
  114. Saccorotti G, Chouet BA, Dawson PB (2001) Wavefield properties of a shallow long-period event and tremor at Kilauea Volcano, Hawaii. *J Volcanol Geotherm Res* 109:163–189
  115. Sakuraba A, Oikawa J, Imanishi Y (2002) Free oscillations of a fluid sphere in an infinite elastic medium and long-period volcanic earthquakes. *Earth Planets Space* 54:91–106
  116. Stump BW, Johnson LR (1977) The determination of source properties by the linear inversion of seismograms. *Bull Seism Soc Am* 67:1489–1502
  117. Sturton S, Neuberg JW (2006) The effects of conduit length and acoustic velocity on conduit resonance: Implications for low-frequency events. *J Volcanol Geotherm Res* 151:319–339
  118. Takei Y, Kumazawa M (1994) Why have the single force and torque been excluded from seismic source models? *Geophys J Int* 118:20–30
  119. Takei Y, Kumazawa M (1995) Phenomenological representation and kinematics of general seismic sources including the seismic vector modes. *Geophys J Int* 121:641–662
  120. Takeo M, Yamasato H, Furuya I, Seino M (1990) Analysis of long-period seismic waves excited by the November 1987 eruption of Izu-Oshima Volcano. *J Geophys Res* 95:19377–19393
  121. Takeuchi H, Saito M (1972) Seismic surface waves. In: Bolt BA (ed) *Methods in Computational Physics*, vol 11. Academic Press, New York, pp 217–295
  122. Tameguri T, Iguchi M, Ishihara K (2002) Mechanism of explosive eruptions from moment tensor analyses of explosion earthquakes at Sakurajima volcano, Japan. *Bull Volcanol Soc Jpn* 47:197–216
  123. Toda S, Stein R, Sagiya T (2002) Evidence from the AD 2000 Izu islands earthquake swarm that stressing rate governs seismicity. *Nature* 419:58–61
  124. Tolstoy I (1954) Dispersive properties of a fluid layer overlying a semi-infinite elastic solid. *Bull Seism Soc Am* 44:493–512
  125. Tuffen H, Dingwell D (2005) Fault textures in volcanic conduits: evidence for seismic trigger mechanisms during silicic eruptions. *Bull Volcanol* 67:370–387
  126. Uehira K, Takeo M (1994) The source of explosive eruptions of Sakurajima volcano, Japan. *J Geophys Res* 99:17775–17789
  127. Ulrych TJ, Clayton RW (1976) Time series modeling and maximum entropy. *Phys Earth Planet Inter* 12:188–200
  128. Ulrych TJ, Sacchi MD (1995) Sompi, Pisarenko and the extended information criterion. *Geophys J Int* 122:719–724
  129. Virieux J (1986) P-SV wave propagation in heterogeneous media: Velocity-stress finite-difference method. *Geophysics* 51:889–901
  130. Wessel P, Smith WHF (1998) New improved version of Generic Mapping Tools released. *Eos trans AGU* 122:149–164
  131. Yamamoto M (2005) Volcanic fluid system inferred from broadband seismic signals. Ph D Thesis, University of Tokyo
  132. Yamamoto M, Kawakatsu H, Kaneshima S, Mori T, Tsutsui T, Sudo Y, Morita Y (1999) Detection of a crack-like conduit beneath active crater at Aso volcano, Japan. *Geophys Res Lett* 26:3677–3680
  133. Yamamoto M, Kawakatsu H, Yomogida K, Koyama J (2002) Long-period (12 sec) volcanic tremor observed at Usu 2000 eruption: Seismological detection of a deep magma plumbing system. *Geophys Res Lett* 29:1329. doi:10.1029/2001GL013996
  134. Yamamura K, Kawakatsu H (1998) Normal-mode solutions for radiation boundary conditions with an impedance contrast. *Geophys J Int* 134:849–855

### Books and Reviews

- Dahlen FA, Tromp J (1998) *Theoretical Global Seismology*. Princeton University Press, Princeton
- Kay SM, Marple SL (1981) Spectrum analysis – A modern perspective. *Proc IEEE* 69:1380–1419

## Volcanoes, Non-linear Processes in

BERNARD CHOUET

US Geological Survey, Menlo Park, USA

### Article Outline

Glossary

Definition of the Subject

Introduction

Description of Seismic Sources in Volcanoes

Sources of Long-Period Seismicity

Source Processes of Very-Long-Period Signals

Future Directions

Acknowledgment

Bibliography

### Glossary

**Bubbly liquid** Liquid-gas mixture in which the gas phase is distributed as discrete bubbles dispersed in a continuous liquid phase.

**Choked flow** Flow condition where the transonic flow of a compressible fluid through a constriction becomes choked. As the fluid flows through a narrowing channel, it accelerates at a rate that depends on the ratio of channel inlet pressure to channel outlet pressure. At a critical pressure ratio, which is a function of the fractional change of cross-sectional area, the speed of the fluid accelerating through the nozzle-like constriction reaches a maximum value equal to its sound speed. This flow condition is termed choked flow.

**Crack wave** A dispersive wave generated by fluid-solid interaction in a fluid-filled crack embedded in an elastic solid. The crack-wave speed is always smaller than the acoustic speed of the fluid.

**Diffuse interface** Thin mixing layer constituting the interface between two immiscible fluids. Interface dynamics is governed by molecular forces and is described by the mixing energy of the two fluid components.

**Gas slug** A gas bubble whose diameter is approximately the diameter of the pipe in which the slug is flowing. In a slug ascending a vertical pipe, the nose of the bubble has a characteristic spherical cap and the gas in the rising bubble is separated from the pipe wall by a falling film of liquid.

**Long-period (LP) event** A seismic event originating under volcanoes with emergent onset of *P* waves, no distinct *S* waves, and a harmonic signature with periods in the range 0.2–2 s as compared to a tectonic earthquake

of the same magnitude. LP events are attributed to the involvement of fluid such as magma and/or water in the source process (see also VLP event).

**Magma** Molten rock containing variable amounts of dissolved gases (principally water vapor, carbon dioxide, and sulfur dioxide), crystals (principally silicates and oxides), and, occasionally, preexisting solid rock fragments.

**Moment tensor** A symmetric second-order tensor that completely characterizes an internal seismic point source. For an extended source, it represents a point-source approximation and can be quantified from an analysis of seismic waves whose wavelengths are much longer than the source dimensions.

**Nonlinear process** Process involving physical variables governed by nonlinear equations that reflect the fundamental micro-scale dynamics of the system. Nonlinear phenomena can arise from a number of different sources, including convective acceleration in fluid dynamics, constitutive relations, boundary conditions representing gas-liquid interfaces, nonlinear body forces, and geometric nonlinearities arising from large-scale deformation.

**Phase-field method** Method treating the interface between two immiscible fluids as a diffuse thin layer; also known as the diffuse interface method. In this approach, the physical and chemical properties of the interfacial layer are defined by a phase-field variable  $\phi$ , whose dynamics are expressed by the Cahn–Hilliard convection-diffusion equation. Use of the phase-field variable avoids the necessity of tracking the interface and yields the correct interfacial tension from the free energy stored in the mixing layer.

**Very-long-period (VLP) event** A seismic event originating under volcanoes with typical periods in the range 2–100 s. VLP events are attributed to the involvement of fluid such as magma and/or water in the source process. Together with LP events, VLP events represent a continuum of fluid oscillations, including acoustic and inertial effects resulting from perturbations in the flow of fluid through conduits.

**Volatile** A chemical compound that is dissolved in magma at high pressure and exsolves from the melt and appears as a low-density gas at low pressure. The most common volatiles in magma are water, carbon dioxide, and sulfur dioxide.

**Waveform inversion** Given an assumed model of the wave-propagation medium, a procedure for determining the source mechanism and source location of a seismic event based on matching observed waveforms with synthetics calculated with the model.

## Definition of the Subject

Magma transport is fundamentally episodic in character as a result of the inherent instability of magmatic systems at all time scales. This episodicity is reflected in seismic activity, which originates in dynamic interactions between gas, liquid and solid along magma transport paths involving complex geometries. The geometrical complexity plays a central role in controlling flow disturbances and also providing specific sites where pressure and momentum changes in the fluid are effectively coupled to the Earth.

Recent technological developments and improvements in the seismological instrumentation of volcanoes now allow the surface effects of subterranean volcanic processes to be imaged in unprecedented detail. Through detailed analyzes of the seismic wavefields radiated by volcanic activity, it has become possible for volcano seismologists to make direct measurements of the volcanic conduit response to flow processes, thus opening the way for detailed modeling of such processes.

However, unlike the description of seismic waves, which is based on the linear equations of elastodynamics, the description of the flow processes underlying the observed seismic source mechanisms is governed by the nonlinear equations of fluid dynamics. In the classic Navier–Stokes equations of fluid mechanics, the nonlinearity resides in the convective acceleration terms in the equations of conservation of momentum. In volcanic fluids, further complexity arises from the strong nonlinear dependence of magma rheology on temperature, pressure, and water and crystal content, and nonlinear characteristics of associated processes underlying the physico-chemical evolution of liquid-gas mixtures constituting magma. An example of nonlinearity in two-phase fluid mixtures are the changing boundary conditions associated with the internal surfaces separating the phases, a situation commonly encountered during the spinodal decomposition of binary fluids [61]. Spinodal decomposition occurs when a mixture of two species *A* and *B* of fluid forming a homogeneous mixture at high temperature undergoes spontaneous demixing following a temperature drop (quench) below some critical temperature. As the temperature drops below the critical temperature, the initially homogeneous mixture becomes locally unstable, and a period of interdiffusion is initiated, which leads to the formation of patches of *A*-rich and *B*-rich fluids separated by sharp interfaces. In the late stages of demixing following a deep quench, the interfacial thickness reaches a molecular scale and the interfacial tension approaches an equilibrium surface tension. As it approaches local equilibrium, the mixture continues to evolve to-

ward a minimization of energy by reducing its interfacial area, and stresses arising from changes in the local interfacial curvature drive fluid motion. A smooth evolution of the interfacial structure ensues, occasionally interrupted by local coalescence, breakup and reconnection, until final equilibrium is reached between the two bulk domains. Similar processes underlie the physics that governs the vesiculation, fragmentation, and collapse of bubble-rich suspensions to form separate melt and vapor in response to decreasing pressure along the ascent path of magma.

All of the above factors contribute to a diversity of oscillatory processes that intensifies as magma rheology becomes more complex; increasingly diverse behavior also characterizes hydrothermal fluids. Heat transfer to the hydrothermal system from magmatic gases streaming through networks of fractures pervading the rock matrix may induce boiling in ground water, resulting in the formation of steam bubbles and attendant bubble oscillations, thereby generating sustained tremor of hydrothermal origin.

Refined understanding of magma and hydrothermal transport dynamics therefore requires multidisciplinary research involving detailed field measurements, laboratory experiments, and numerical modeling. The comprehensive breadth of seismology must be used to understand and interpret the wide variety of seismic signals encountered in various volcanic processes. This effort must then be complemented by laboratory experiments and numerical modeling to better understand the complex flow dynamics at the origin of the source mechanisms imaged from seismic data. Such research is fundamental to monitoring and interpreting the subsurface migration of magma that often leads to eruptions, and thus enhances our ability to forecast hazardous volcanic activity.

## Introduction

Degassing is the main driving force behind most volcanic phenomena. The separation of vapor and melt phases leads to the formation of bubbles, the presence of which decreases magma density, enhances magma buoyancy and propels magma ascent [140]. Bubbles in magma also naturally increase magma compressibility, resulting in a low acoustic speed of the bubbly mixture compared to that of the liquid phase [28,36,37,56]. At shallow depths, the high compressibility of the bubbles may further contribute to the generation of pressure oscillations in bubble clouds [28,145]. Bubble coalescence and fragmentation can also contribute to pressure disturbances [85]. In hydrothermal fluids, the collapse of small vapor bub-

bles can be the source of pressure pulses and sustained tremor [70,72].

Bubble dynamics play a key role in the transport of magmatic and hydrothermal fluids, not only as sources of acoustic energy, but also in providing a sharp contrast in velocity between the fluid and encasing solid, which facilitates the entrapment of acoustic energy at the source. This aspect of the source often manifests itself in the form of long-lasting, long-period harmonic oscillations produced by sustained resonance at the source.

Other types of gaseous fluid mixtures also favor source resonance. For example, gases laden with solid particles, or gases mixed with liquid droplets, may produce velocity contrasts that are similar to, or stronger, than those associated with bubbly liquids. Indeed, dusty gases made of micron-sized particles, or misty gases made of micron-sized droplets, can sustain resonance at the source over durations that far exceed those achieved with bubbly fluids [77].

Large gas slugs bursting at the liquid surface can act as active sources of acoustic and seismic radiation [59, 118,136]. Gas slugs traversing discontinuities in conduit geometry can also cause transient liquid motions inducing pressure and momentum changes [60], which radiate elastic waves via their coupling to the conduit walls [32,35,109]. In magmas where a high fluid viscosity impedes flow, the diffusion of volatiles from a super-saturated melt may result in gradual pressurization of the melt and attendant deformation of the encasing rock matrix [33,34,104,124]. The fluid motions resulting from unsteady slug dynamics, large degassing bursts, and diffusion-dominated processes typically produce signals with characteristic periods longer than those commonly associated with acoustic resonance.

Oscillatory processes are thus ubiquitous during magma flow and are a natural expression of the release of thermo-chemical and gravitational energy from volcanic fluids. The radiated elastic energy reflects both active and passive processes at the source, and features a large variety of signals over a wide range of periods.

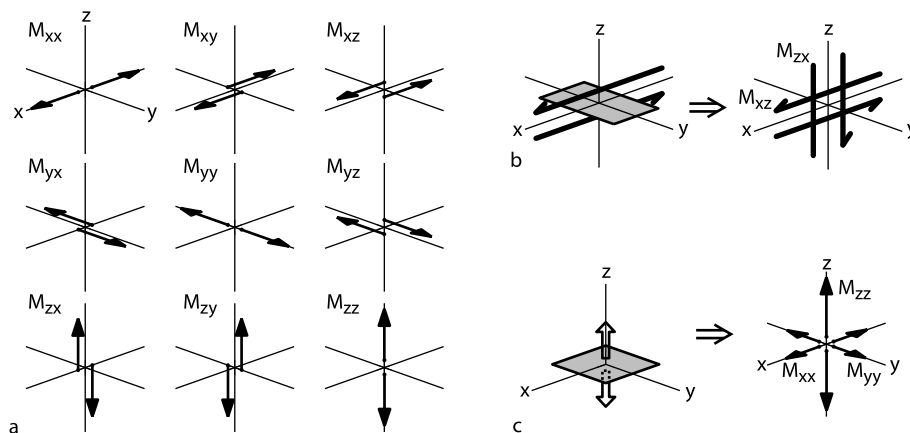
Seismic signals originating in the dynamics of magmatic and hydrothermal fluids typically include Long-Period (LP) events and tremor [27,28]. LP events resemble small tectonic earthquakes in duration but differ in their characteristic frequency range and harmonic signature. Tremor is characterized by a harmonic signal of sustained amplitude lasting from minutes to days, and sometimes for months or even longer. In many instances, LP events and tremor are found to have essentially the same temporal and spectral components, suggesting that a common source process, differing only in duration, underlies these

two types of events. Accordingly, LP events and tremor are often grouped under the common appellation LP seismicity. The periods in which LP seismicity is observed typically range from 0.2 to 2 seconds, and the characteristic oscillations of LP signals are commonly viewed as a result of acoustic resonance in a fluid-filled cavity. Slower processes involving inertial forces associated with unsteady mass transport may radiate seismic signals extending to periods much longer than 2 seconds. These types of signals commonly fall under the appellation of Very-Long-Period (VLP) seismicity. Collectively, LP and VLP seismicities provide a comprehensive view of mass transport dynamics under a volcano.

The present chapter offers a brief review of the state of the art in volcano seismology and addresses basic issues in the quantitative interpretation of processes operative in active volcanic systems. The chapter starts with an introduction of the seismic methodology used to quantify the source of volcano seismicity, with emphasis on sources originating in the dynamics of volcanic fluids. A review of some of the representative source mechanisms of LP and VLP signals and of their implications for volcanic processes follows in Sects. “Sources of Long-Period Seismicity” and “Source Processes of Very-Long-Period Signals”. A final section describes a mesoscale computational approach for simulating two-phase flows of complex magmatic fluids. This method, called the phase-field method, grew out of the original work of [19], and relies on an energy-based variational formalism to treat interfacial dynamics and complex rheology in a unified framework. The phase-field method has been successfully applied to the modeling of complex micro-structured two-phase fluids in the field of engineering materials and appears well adapted to the numerical simulation of nonlinear volcanic processes.

## Description of Seismic Sources in Volcanoes

A general kinematic description of seismic sources in volcanoes is commonly based on a moment-tensor and single-force representation of the source [3] (Fig. 1a). The seismic-moment tensor consists of nine force couples, with each corresponding to one set of opposing forces (dipoles or shear couples). This symmetric second-order tensor allows a description of any generally oriented discontinuity in the Earth in terms of equivalent body forces. For example, slip on a fault can be represented by an equivalent force system involving a superposition of two force couples of equal magnitudes – a double couple (Fig. 1b). Similarly, a tensile crack has an equivalent force system made of three vector dipoles, with dipole magnitudes with



Volcanoes, Non-linear Processes in, Figure 1

**a** The nine possible couples corresponding to the moment-tensor components describing the equivalent force system for a seismic source in the Earth. **b** Slip on a fault can be described by a superposition of two force couples, in which each force couple is represented by a pair of forces offset in the direction normal to the force. Sources involving slip on a fault thus have an equivalent force system in the form of a double couple composed of four forces. **c** A tensile crack has a representation in the form of three vector dipoles, in which each dipole consists of a pair of forces offset in the direction of the force

ratios  $1:1:(\lambda + 2\mu)/\lambda$ , in which  $\lambda$  and  $\mu$  are the Lamé coefficients of the rock matrix [28], and where the dominant dipole is oriented normal to the crack plane (Fig. 1c). Injection of fluid into the crack will cause the crack to expand and act as a seismic source. In general, magma movement between adjacent segments of conduit can be represented through a combination of volumetric sources of this type. Because mass-advection processes can also generate forces on the Earth, a complete description of volcanic sources commonly requires the consideration of single forces in addition to the volumetric source components expressed in the moment tensor. For example, a volcanic eruption can induce a force system that includes a contraction of the conduit/reservoir system in response to the ejection of fluid, and a reaction force from the eruption jet [64] (Fig. 2). Some volcanic processes can be described by a single-force mechanism only. An example is the single-force mechanism attributed to the massive landslide observed at the start of the 1980 eruption of Mount St. Helens [63,66]. A single-force source model was also proposed by Uhira et al. [134] to explain the mechanism of dome collapses at Unzen Volcano, Japan. In general terms, a single force can be generated by an exchange of linear momentum between the source and the rest of the Earth [129]. Thus, a single force on the Earth may result from an acceleration of the center of mass of the source in the direction opposite to the force, or deceleration of the center of mass of the source in the same direction as the force [129].

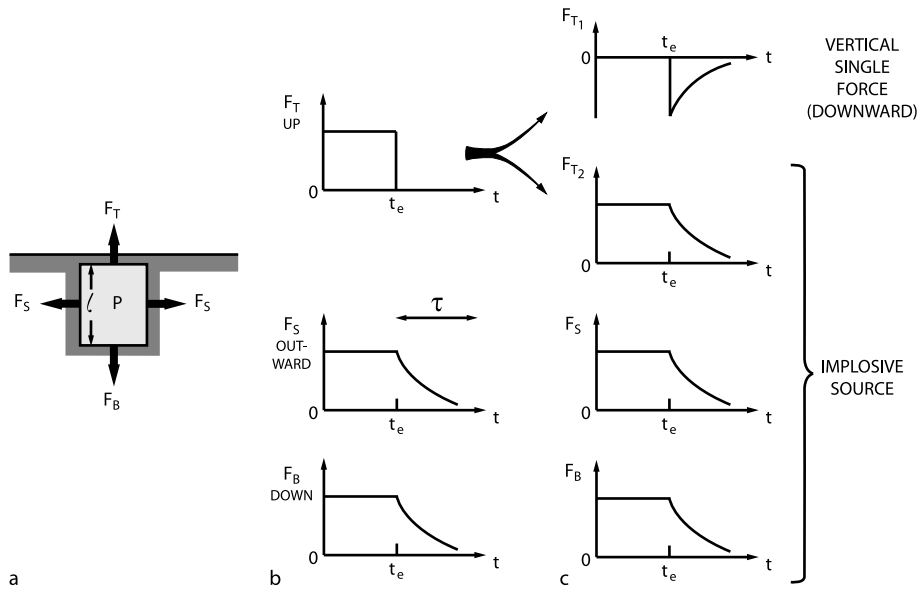
## Waveform Inversion

Waveform inversions solving for the amplitudes and time histories of the moment-tensor and single-force components of the source have become the primary tool to identify and understand the mechanisms of generation of LP and VLP signals recorded by broadband seismometers. When the wavelengths of observed seismic waves are much longer than the spatial extent of the source, the source may be approximated by a point source, and the force system represented by the moment-tensor and single-force components is localized at this point. The waveform inversion procedure requires a calculation of the impulse responses (Green's functions) of the medium to this force system for each receiver component. The displacement field generated by a seismic source is described by the representation theorem which, for a point source, may be written as e. g., Chouet, [28]

$$u_n(t) = F_p(t) * G_{np}(t) + M_{pq}(t) * G_{np,q}(t),$$

$$p, q = x, y, z, \quad (1)$$

where  $u_n(t)$  is the  $n$ -component of seismic displacement at a receiver at time  $t$ ,  $F_p(t)$  is the time history of the force applied in the  $p$ -direction,  $M_{pq}(t)$  is the time history of the  $pq$ -component of the moment tensor, and  $G_{np}(t)$  is the Green tensor which relates the  $n$ -component of displacement at the receiver position with the  $p$ -component of impulsive force at the source position. The notation,  $q$ , indicates spatial differentiation with respect to the



Volcanoes, Non-linear Processes in, Figure 2

Force system equivalent to a volcanic eruption (after [64]). **a** Pressurized cavity model for a volcanic eruption. A shallow vertically oriented cylindrical cavity initially sealed at the top by a lid contains a pressurized inviscid fluid which exerts an upward vertical force  $F_T$  on the lid, a horizontal outward force  $F_S$  on the sidewall, and a vertical downward force  $F_B$  on the bottom of the cylinder. **b** Time histories of forces acting on the top, side, and bottom cavity walls of the cavity. The eruption is simulated by the sudden removal of the lid at time  $t = t_e$ , at which point the force  $F_T$  vanishes instantaneously and the fluid pressure in the cylinder starts to decrease with a characteristic time constant  $\tau$  fixed by the mass flux of the eruption, i.e.,  $\tau \sim \ell/v$ , where  $\ell$  is the length of the cylinder and  $v$  is the mean fluid velocity inside the cylinder. Since the forces  $F_S$  and  $F_B$  are both proportional to pressure, they decrease with the same time constant  $\tau$ . **c** Decomposition of the force system to a downward single force and implosive source. The force  $F_T$  in **b** is decomposed into a vertical downward component  $F_{T1}$  and vertical upward component  $F_{T2}$  in such a way that  $F_{T2}$  has the same time history as  $F_S$  and  $F_B$ . As a result, the three forces  $F_{T2}$ ,  $F_S$ , and  $F_B$  form an implosive source so that the eruption mechanism is represented by the superposition of a downward vertical force (the reaction force of the volcanic jet) with this volumetric implosion

$q$ -coordinate and the symbol  $*$  denotes convolution. Summation over repeated indices is implied.

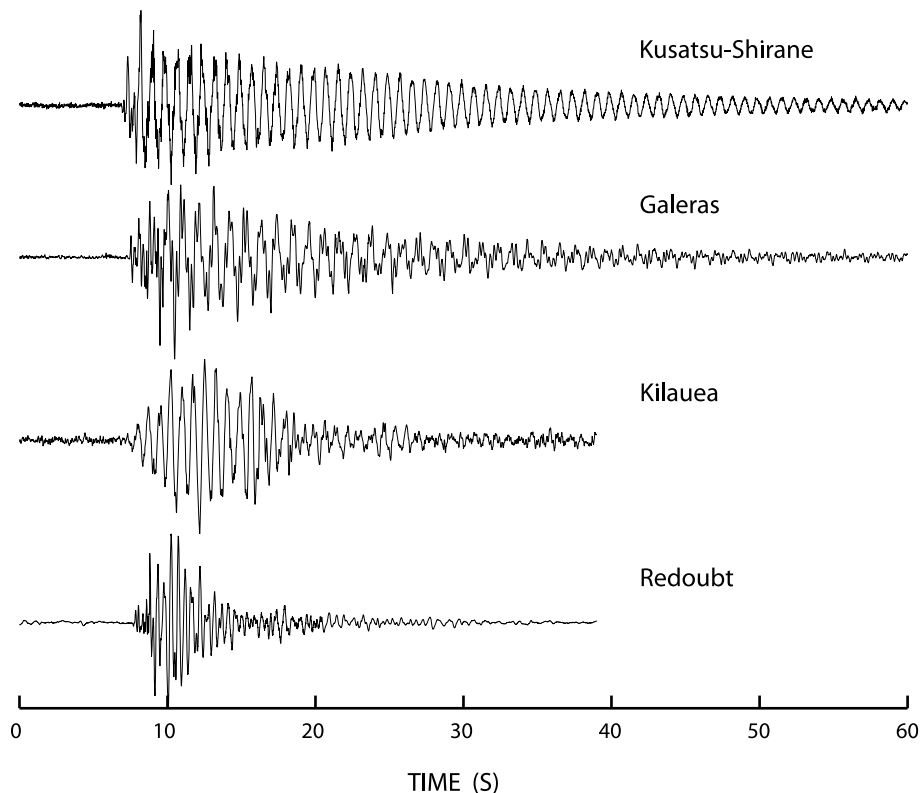
A common approach to the quantification of the seismic source mechanism relies on a discretized representation of the volcanic edifice based on a digital elevation model. Using this model, Green's functions are then calculated by the finite difference method [107]. Once the Green's functions are known, the nine independent mechanisms in Eq. (1) are retrieved through least-squares inversion expressing the standard linear problem  $\mathbf{d} = \mathbf{Gm}$  in this equation [32].

### Sources of Long-Period Seismicity

Long-Period (LP) seismicity, including individual LP events and tremor, is commonly observed in relation to magmatic and hydrothermal activities in volcanic areas and is recognized as a precursory phenomenon for eruptive activity (e. g., Chouet, [27]). The waveform of the LP event is characterized by simple decaying harmonic oscil-

lations except for a brief interval at the event onset (Fig. 3). This signature is commonly interpreted as oscillations of a fluid-filled resonator in response to a time-localized excitation. By the same token, tremor may be viewed as oscillations of the same resonator in response to a sustained excitation. LP events are particularly useful in the quantification of magmatic and hydrothermal processes because the properties of the resonator system at the source of this event can be inferred from the properties of the decaying oscillations in the tail of the seismogram. The damped oscillations in the LP coda are quantified by two parameters,  $T$ , and  $Q$ , where  $T$  is the period of the dominant mode of oscillation, and  $Q$  is the quality factor of the oscillatory system representing the combined effects of intrinsic and radiation losses.

Interpretations of the oscillating characteristics of LP sources have mostly relied on a model of fluid-driven crack [23,24,25]. This model, which has the most natural geometry that satisfies mass-transport conditions at depth beneath a volcano, is supported by results from

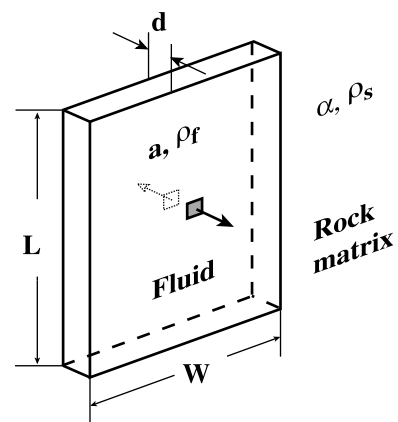


Volcanoes, Non-linear Processes in, Figure 3

Typical signatures of long-period events observed at Kusatsu-Shirane, Galeras, Kilauea, and Redoubt Volcanoes. The signatures are all characterized by a harmonic coda following a signal onset enriched in higher-frequencies

inversions of LP waveforms recorded at several volcanoes [80,82,97,137]. A simplified two-dimensional model of fluid-driven crack was originally proposed by Aki et al. [2]. Although this model considered both the driving excitation and geometry appropriate for transport, the fluid inside the crack was only treated as a passive cushion that did not support the acoustic propagation of the pressure disturbance caused by the motion of the crack wall. An extension of this model including active fluid participation was proposed by Chouet and Julian [22], who considered a simultaneous solution of the elastodynamics and fluid dynamics for a two-dimensional crack. This model was further extended to three dimensions by Chouet [23] and was extensively studied by Chouet [24,25].

Chouet's three-dimensional model consists of a single isolated rectangular crack embedded in an infinite elastic body and assumes zero mass transfer in and out of the crack (Fig. 4). Crack resonance is excited by a pressure transient applied symmetrically on both walls over a small patch of crack wall. In this model, the crack aperture is assumed to be much smaller than the seismic

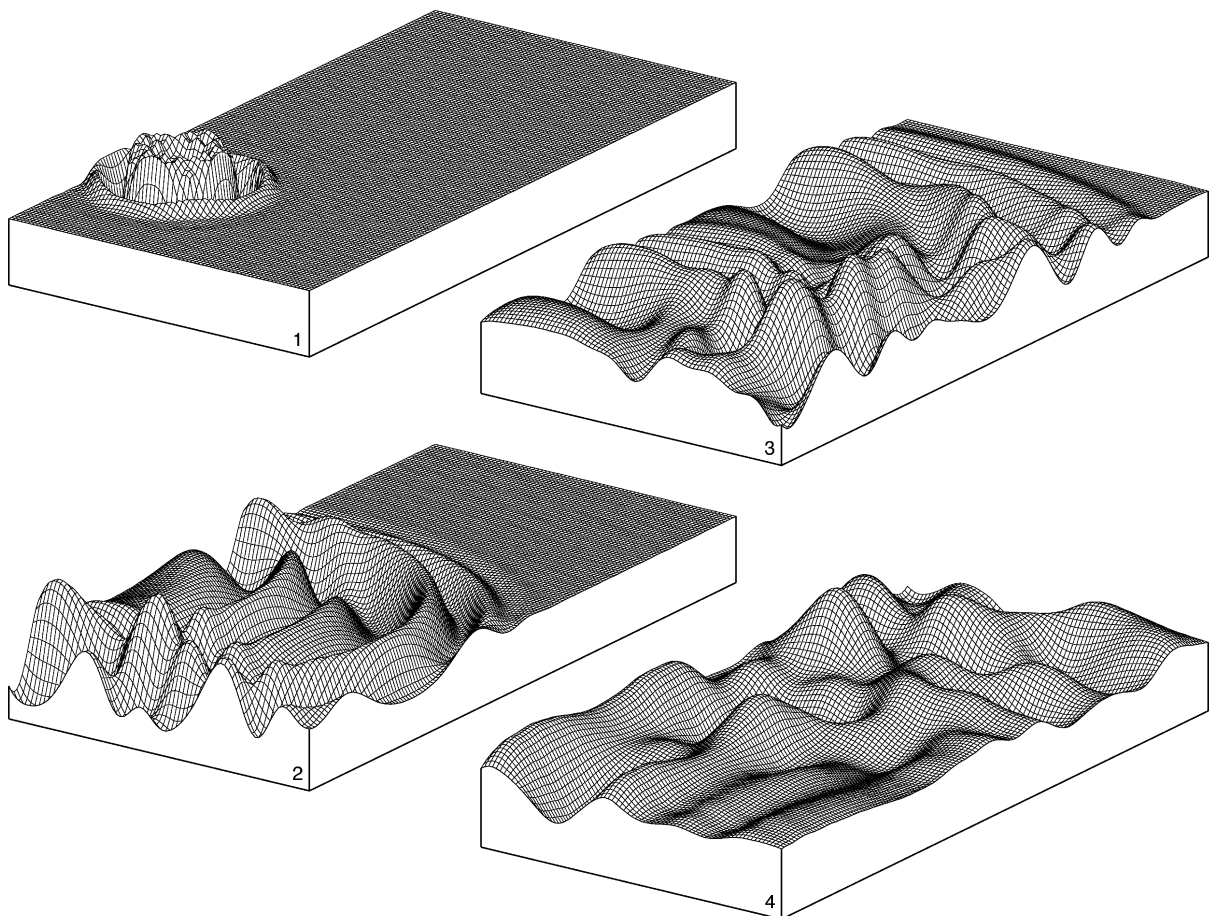


Volcanoes, Non-linear Processes in, Figure 4

Geometry of the fluid-filled crack model of Chouet [23]. The crack has length  $L$ , width  $W$ , and aperture  $d$ , and contains a fluid with sound speed  $a$  and density  $\rho_f$ . The crack is embedded in an elastic solid with compressional wave velocity  $\alpha$  and density  $\rho_s$ . Excitation of the crack is provided by a pressure transient applied symmetrically on both walls over the small areas such as that indicated by the gray patch and dotted small square

wavelengths of interest and the motion of the fluid inside the crack is treated as two-dimensional in-plane motion. The fluid dynamics are represented by the conservation of mass and momentum for the two in-plane components of fluid velocity averaged over the crack aperture, and convective terms in the momentum equations are neglected assuming their contributions are negligibly small compared to the time derivatives of pressure and flow velocity. A solution of the coupled equations of fluid dynamics and elastodynamics is then obtained by using a finite-difference approach. As formulated, the model accounts for the radiation loss only; intrinsic losses due to dissipation mechanisms within the fluid are treated separately [77].

Figure 5 shows snapshots of the normal component of particle velocity on the crack wall obtained at four different times following the onset of crack resonance. In this particular example, the crack excitation is triggered by a step in pressure applied at a point located a quarter of the crack width from the left edge of the crack and a quarter of the crack length from the front edge of the crack. The first snapshot at the upper left shows the propagation of the elastic disturbance on the crack surface shortly after the onset of the pressure transient. The second snapshot at the lower left shows the motion of the crack wall following the first reflections from the front, left and right edges of the crack. The third snapshot at the upper right represents the wall motion immediately preceding the reflection



Volcanoes, Non-linear Processes in, Figure 5

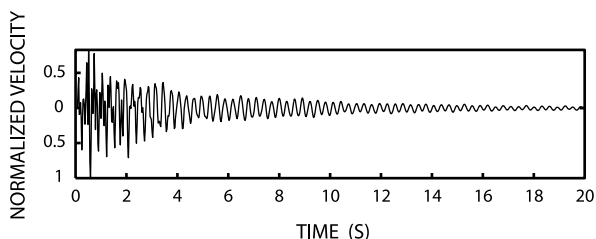
Snapshots of the normal component of particle velocity on the surface of a fluid-filled crack excited by a step in pressure applied over a small patch of the crack wall. The snapshots (1–4) show the propagation of the disturbance on the crack wall at four different times in the history of crack oscillation specified by the number of time steps,  $N$ , in the finite-difference solution obtained by Chouet [23]. The time step has size  $\Delta t = 0.003125L/\alpha$  (see Fig. 4 for explanation of symbols). Depicted are the shape of the wall motion at  $N = 72$  (snapshot 1),  $N = 312$  (snapshot 2),  $N = 560$  (snapshot 3), and  $N = 1200$  (snapshot 4). (See text for discussion)



from the back edge of the crack, and the fourth snapshot at the lower right depicts the shape of the lateral and longitudinal modes of crack resonance, which by then are well established.

The excitation of modes in the crack model depends on the position of the pressure transient, the extent of crack surface affected by the transient, the temporal characteristics of the transient, and crack stiffness  $C = (b/\mu)(L/d)$ , where  $b$  is the bulk modulus of the fluid,  $\mu$  is the rigidity of the rock,  $L$  is the crack length, and  $d$  is the crack aperture [2,23]. An example of synthetic seismogram representing the vertical ground velocity response in the near field of a fluid-filled crack excited into resonance by a step transient in pressure is shown in Fig. 6. The synthetics display sustained oscillations that bear strong resemblance to observed LP waveforms (compare with Fig. 3), convincingly demonstrating the critical involvement of active fluid participation in the source process of volcanic signals.

Chouet's studies demonstrated the existence of a slow wave propagating along the crack wall, which he named the "crack wave". The asymptotic behavior of this wave was investigated by Ferrazini and Aki [42] in an analytical study of normal modes trapped in a liquid layer sandwiched between two elastic half spaces. The crack wave speed is slower than the sound velocity of the liquid and decreases with increasing crack stiffness and increasing wavelength; this result reflects the elasticity of the crack wall, whose effect is to decrease the effective bulk modulus of the fluid. These slow characteristics of the crack wave lead to more realistic estimates of crack dimensions compared to estimates based on the sound speed of a fluid embedded in a resonator with perfectly rigid walls. Source dimensions estimated from LP data based on this model



Volcanoes, Non-linear Processes in, Figure 6

Synthetic velocity waveform observed at an epicentral distance of 500 m and azimuth of  $45^\circ$  from the crack trace for a vertical crack (vertical extent  $L = 150$  m and horizontal extent  $W = 75$  m) buried at a depth of 500 m in a homogeneous half space. The aspect ratio is  $L/d = 10^4$ , and the parameters of the fluid and solid are  $a = 300$  m/s,  $\rho_f = 2120$  kg/m<sup>3</sup>, and  $\alpha = 4500$  m/s,  $\rho_s = 2650$  kg/m<sup>3</sup>. (Reproduced from [77])

typically range from tens to several hundred meters [79, 82,120]. Detailed analyzes of the dependence of crack resonance on fluid composition carried out by Kumagai and Chouet [77], and systematic investigations of LP signatures based on their results, suggest that dusty gases or bubbly basalt are common fluids in LP events of magmatic origin [27,45,76], and that wet gases, steam and bubbly water are typically representative of the source of LP events of hydrothermal origin [79,120].

The model of Chouet [23] does not address the excitation mechanism of LP events or tremor. Rather, the spatio-temporal properties of the pressure transient triggering the crack resonance are preset as kinematic conditions in the model. The usefulness of this model is thus restricted to a quantification of the crack resonance and properties of the fluids at the source of LP events. For a better perspective of the excitation mechanism of LP seismicity, observations of the source over a wider band of frequencies are necessary, along with an adequate sampling of the evolutionary characteristics of volcanic seismicity. Observations carried out in different volcanic settings point to a wide variety of LP excitation mechanisms originating in magmatic-hydrothermal interactions, as well as magmatic instabilities.

### Magmatic-Hydrothermal Interactions

In hydrothermal systems, LP seismicity can be induced by surges in the heat transfer from an underlying magma body. Kusatsu-Shirane Volcano in central Japan provides a prime example of the character of such interactions. Three crater lakes (Yugama, Mizugama, and Karagama) occupy the summit of the volcano along with numerous hot springs, pointing to an active magmatic system and enhanced circulation of hydrothermal fluids. Geochemical studies provide further support for hydrothermal activity resulting from the interaction of hot volcanic gases with groundwater [55,106]. LP events marked by nearly monochromatic oscillatory signatures have been frequently observed at this volcano (see Fig. 3) [43,46], and temporal variations in the decay characteristics of these events have also been documented [96]. Kumagai et al. [79] performed a systematic study of the temporal characteristics of LP events observed at Kusatsu-Shirane over a period from August 1992 through January 1993, and Kumagai et al. [80] carried out waveform inversions of a typical event during this time interval. Results from these studies point to the repeated excitation of a fixed sub-horizontal crack at depths near 200 m under the Yugama crater lake, and consistently explain temporal variations in LP signatures by the dynamic response of this crack

to a magmatic heat pulse. In this interpretation, the crack contains an initially wet gas, which became gradually drier with time, suggesting a “drying out” of the hydrothermal system in response to the heat pulse. Further analyzes by Nakano et al. [97] based on waveform inversions of the effective excitation functions (the signal components remaining after removal of the resonance characteristics) of individual LP events confirm the crack mechanism imaged by Kumagai et al. [80] and shed further light on the source process. This process involves a collapse and recovery of the sub-horizontal crack accompanied by an upward-directed force, and is consistent with a gradual buildup of pressure in the crack causing repeated discharges of steam from the crack. The vertical force imaged with the volumetric component of the source reflects the release of gravitational energy that occurs as a slug of steam ejected from the crack ascends toward the surface and is replaced by a downward flow of cooler water within a conduit linking the crack to the base of the Yugama crater lake. In this scenario, crack resonance is triggered by the sudden collapse of the crack induced by the venting of steam.

Similar mechanisms of interaction between magmatic and hydrothermal systems have been inferred for Kilauea, Mount St. Helens, and Redoubt Volcanoes. At Kilauea Volcano, Hawaii, Almendros et al. [4] imaged a shallow hydrothermal system from frequency-slowness analyses of LP seismicity recorded on three small-aperture seismic arrays deployed in the summit caldera. Located within the top 500 m below the caldera floor, this reservoir is perched immediately above the shallowest segment of magma conduit identified by [109] from inversions of VLP signals produced during a transient in the magma flow feeding the long-lived and ongoing east rift eruption of this volcano [48]. Episodic deflations of the Kilauea summit during brief pauses in the flow of magma are commonly followed by rapid reinflation of the summit during the resumption of flow, and subsequent slower deflation back to an original state preceding the pause [21]. Larger summit reinflation episodes associated with more energetic surges of magma are accompanied by the emission of VLP signals, production of shallow brittle failure earthquakes, and enhanced LP seismicity in the hydrothermal reservoir [39,109]. Detailed analyses of seismic array data by Almendros et al. [5] further document the persistence of low-amplitude tremor generated by shallow hydrothermal activity in response to the flux of heat from the underlying magma conduit. Additional support for this scenario also comes from analyses of the source process of a hydrothermal LP event [82], which point to the resonance of a horizontal crack at depth of  $\sim 150$  m immediately above the conduit imaged by Ohminato et al. [109]. The observed

frequencies and attenuation characteristics of crack resonance are consistent with a crack filled with either bubbly water or steam [82].

The inferred scenario of magmatic-hydrothermal interactions at Kilauea, which is in harmony with all of the above observations, involves a chain of causes and effects that starts with a volumetric deformation of the underlying shallow segment of magma conduit induced by the surging magma. This conduit response causes brittle fractures in the surrounding rock matrix, and thus enhances the upward transport of magmatic gases through the fractured medium to shallow depths, where the hot gases heat and activate the hydrothermal system. The boiling of groundwater and attendant production of steam then raise the overall pressure of hydrothermal fluids in fractures permeating the medium, thereby leading to repeated impulsive discharges of fluid and resonant excitation of individual fractures.

The current dome-building eruption of Mount St. Helens Volcano, Washington, is marked by the extrusion of a stiff plug of dacitic magma [40,91] accompanied by shallow, repetitive LP events that have been dubbed “drumbeats” by scientists monitoring Mount St. Helens [94]. VLP signals commonly accompany the LP events, but are generally detected only in the immediate vicinity of the crater. Using high-quality data from a temporary broadband seismic network deployed in 2004–2005, Waite et al. [137] carried out systematic waveform inversions for both LP and VLP signals. Results from these inversions point to the perturbation of a common crack system linking the magma conduit and shallow water-saturated region of the volcano. A scenario consistent with the imaged source mechanisms involves the repeated pressurization of a sub-horizontal crack that lies beneath the growing lava dome in the southern part of the crater. Steam produced by the heating of ground water, as well as steam exsolved from the magma, feed into the crack and pressurize it. Upon reaching a critical pressure threshold, steam is vented out of the crack through fractures leading to the surface. The drop in pressure triggers a partial collapse of the crack, which resonates for 5 to 10s of seconds. The collapse of the crack in turn induces sagging of the overlying dome and triggers a passive response in the magma conduit that is observed in the VLP band. The observed force and volumetric components of this VLP response are both consistently explained as the result of a perturbation in an otherwise smooth steady flow of magma upward through a dike and into a sill underlying the old dome. In this model, the dike represents the top of the old conduit that fed the 1980–1986 dome-building eruptions, and the sill represents a bypass below the old dome. Accord-

ing to this interpretation, at the start of renewed activity in September 2004, magma moving up the conduit found an easier pathway to the surface by breaking out of the conduit near the surface and pushing sub-horizontally to the south rather than intruding the older dome complex. The imaged sill represents this segment of conduit, and the hydrothermal crack represents an extension of this fracture into the shallow water-soaked region beneath the crater floor. This picture of Mount St. Helens as a steam engine is quite distinct from an earlier model in which the seismic drumbeats were assumed to represent stick-slip motion between the extruding lava and conduit walls (see below). A sustained supply of heat and fluid from the magmatic system is necessary to keep the crack pressurized and keep the drumbeats beating in the model elaborated by Waite et al. [137]. As for Kilauea, a proper interpretation of this activity has important implications for the assessment of the potential for phreatic eruptions at Mount St. Helens.

Another perspective on a different type of instability presumed to be associated with dome growth activity at Mount St. Helens was proposed by Iverson et al. [57]. Their model, originally developed before results from the detailed seismic analyses by Waite et al. [137] became available, hypothesizes that repetitive stick-slip motion along the perimeter of the extruding solid plug may cause observable earthquakes. Although not supported by the character of LP seismicity observed at Mount St. Helens, this model nevertheless offers an intriguing view into a possible oscillatory behavior of plug flow. Oscillations in extrusion velocity originate in the interaction of plug inertia, a variable upward force due to magma pressure, and downward force due to plug mass. Assuming that a steady ascent of compressible magma drives the upward extrusion of a solidified plug exhibiting nonlinear rate-weakening friction along its margins, Iverson et al. [57] infer that stick-slip oscillations might be an inevitable component of such an extrusion process; however, the predicted amplitude of individual slip events, estimated at a few millimeters, suggests this mechanism may be difficult to detect. The seismic source associated with a stick-slip extrusion event may manifest itself as a combination of a near-vertical force (assuming vertical spine growth) representing the reaction force on the Earth due to the upward acceleration of mass, and double couple resulting from shear near the plug margins where viscosity is very high and magma is in the glass transition. This particular mechanism does not involve LP source resonance and is apparently not part of the dominant expression of seismicity at Mount St. Helens, but could possibly be operating at a level below the threshold of seismic detection. Such mechanism may be observable in dome-building eruptions where the extrusion rate

is sufficiently large and unsteady to produce oscillations of this type.

The reawakening of Redoubt Volcano, Alaska on 13 December, 1989, after 23 years of quiescence, was heralded by a rapidly accelerating swarm of repetitive LP events that merged into sustained tremor a few hours before the eruption onset [26,126]. These LP events were interpreted by Chouet et al. [26] as the resonant excitation of a crack linking a shallow, low-pressure hydrothermal system to a deeper supercharged magma-dominated reservoir. The initiation of the swarm was attributed by these authors to the failure of a barrier separating the two reservoirs. The actual mechanism triggering the LP excitation in this model is the unsteady choking of a supersonic flow of magmatic steam driven by the pressure gradient existing between the two reservoirs. Unsteady choking itself is the result of fluctuations in crack outlet pressure associated with the reaction of the hydrothermal system to the injection of mass and heat from below. Similarly, the emergence of sustained tremor late in the swarm is interpreted as a change in choked-flow regime related to a gradual weakening of the pressure gradient driving the flow [92]. Numerical simulations of the transonic flow through a crack featuring a nozzle-like constriction were performed by Morrissey and Chouet [95] using the Navier–Stokes equations representing a mixture of gas and suspended solid particles. In these simulations, shock waves develop immediately downstream from the nozzle-like constriction, and the flow acts as an energy transducer that converts smooth low-amplitude fluctuations of outlet flow pressure into strongly-amplified and repetitive step-like pressure transients applied to the crack wall immediately downstream from the nozzle. The magnitude of the pressure transient at the walls is fixed by the pressure of the supplying reservoir and the geometry of the nozzle aperture, presumably remaining constant as long as these conditions prevail. The temporal pattern of outlet pressure fluctuations controls the areal extent of the crack wall impacted by shock waves and, given a roughly constant shock magnitude, fixes the amplitude of the force applied to the wall by the fluid. This variable force applied at a fixed location along the crack wall is viewed as the force responsible for the scaling of amplitudes and the similarity of waveforms observed in the LP swarm at Redoubt [26].

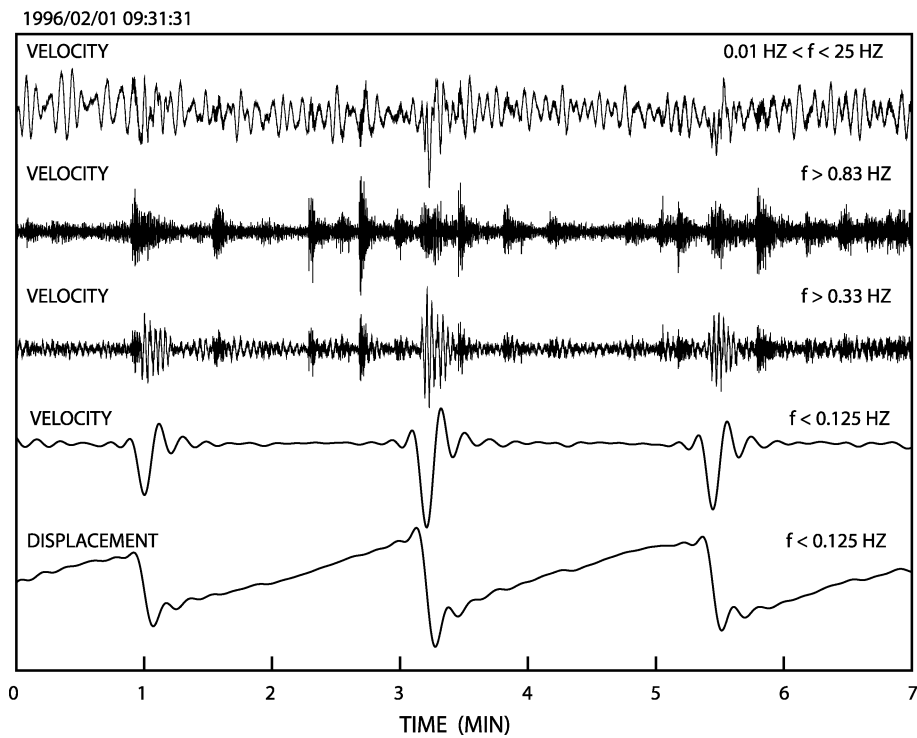
### Magmatic LP Events

In low-viscosity basaltic systems, LP seismicity may be part of a broadband source process associated with the unsteady transport of magma and gases through conduits. This is well illustrated in a broadband seismic record ob-

tained at Kilauea (Fig. 7); most interesting is the displacement record shown at bottom of figure. This record, obtained after filtering the raw broadband signal (top trace in Fig. 7) with a 0.125 Hz low-pass filter, features a repetitive sawtooth signal with rise time of 2–3 min and drop time of 5–10 s. The observed repetitiveness of the VLP signal is a clear indication of the repeated action of a non-destructive source. Noteworthy also are the characteristic LP signatures obtained by filtering the broadband signal with a 0.33-Hz high-pass filter (third trace from the top). These LP events display a fixed dominant period of 2.5 s, whose onsets coincide with the start of the dropdown in the VLP sawtooth displacement signals. Such coincidence is strongly suggestive of a causative relationship between the VLP and LP signals.

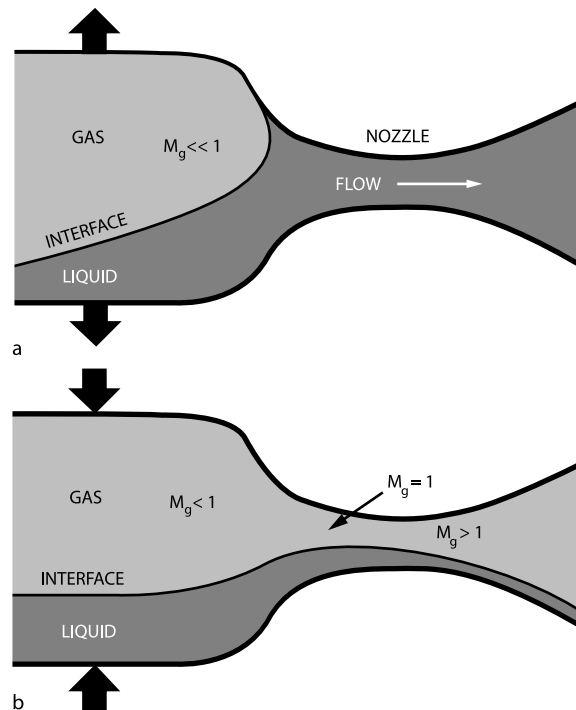
The data in Fig. 7 are part of a seismic sequence observed during a surge in the magma flow feeding the ongoing east rift eruption of Kilauea, and the source mechanism of the VLP signals obtained by Ohmiato et al. [109] from waveform inversions of these data points to a repeated cycle of deformation of a sub-horizontal crack located roughly 1 km below the floor of the summit caldera. Each sawtooth in the recorded displacement signal represents a slow dilation of the crack, followed by a rapid collapse of the crack. The volume change associated with individual sawtooths ranges from 1000 to 4000 m<sup>3</sup> [109].

A conceptual model consistent with the seismic source mechanism imaged by Ohmiato et al. [109] is shown in Fig. 8. This model involves the injection of a large slug of gas into a sub-horizontal crack with a narrow constricted



Volcanoes, Non-linear Processes in, Figure 7

Broadband record and associated filtered signals obtained at Kilauea Volcano during a volcanic event on February 1, 1996. The broadband signal represents ground velocity and is filtered in various frequency bands to produce five records for the same 7-min time interval. The top trace shows the broadband signal ( $0.01 < f < 25$  Hz), which is dominated by the oceanic wave-action microseism with dominant periods in the range 3 to 7 s. The second trace shows the signal obtained after application of a high-pass filter ( $f > 0.83$  Hz). The result is equivalent to a typical short-period record and shows a series of events superimposed on a background of tremor. The third trace also has a high-pass filter applied but with a lower corner frequency ( $f > 0.33$  Hz); LP signals with a dominant period of about 2.5 s are enhanced in this record. The fourth trace shows the signal when a low-pass filter is applied ( $f < 0.125$  Hz); a repetitive VLP signal consisting of pulses with period of about 20 s is observed. The fifth trace is the corresponding displacement record obtained with the same low-pass filter ( $f < 0.125$  Hz), showing a repetitive sawtooth pattern with rise time of 2–3 min and drop time of 5–10 s. Notice that the onset of the LP signal seen above coincides with the onset of the dropdown in the VLP displacement. (Reproduced from [38])



Volcanoes, Non-linear Processes in, Figure 8

Conceptual model of separated gas-liquid flow through a converging-diverging nozzle under choked conditions (see pp. 71–74 in [138]). **a** Inflation phase showing gas accumulation upstream of the nozzle, leading to pressure buildup and deformation of the crack. The gas slug is essentially stationary and the Mach number of the gas is  $M_g \ll 1$ . This stage coincides with the upgoing ramp in the sawtooth displacement signal at the bottom of Fig. 7. **b** Separated gas-liquid flow through the nozzle under compound choked conditions. The gas flow is choked ( $M_g = 1$ ) in the nozzle defined by the gas-liquid interface and upper solid wall, and is supersonic ( $M_g > 1$ ) immediately downstream of the nozzle. There exists a maximum possible gas flow rate that is fixed by the rate of liquid discharge for given upstream conditions and fixed throat geometry. In the limit, liquid fills the duct entirely and there is no gas flow. (Reproduced from [109])

outlet; both liquid magma and gas flow through this outlet constriction. Initially, liquid magma occupies the entire outlet cross section and gas accumulates in the crack, forming a layer of gas on top of the liquid and deforming the crack walls as the back pressure of gas increases in the crack. Eventually, when the back pressure buildup in the crack reaches a critical threshold, the gas slug deforms and rapidly squeezes past the constriction along with the liquid, triggering a rapid deflation (collapse) of the crack in the process. This sequence is then repeated with the next gas slug. In this scenario, the restraining force of the liquid on the gas acts like a self-activated viscosity-controlled valve. In the stratified flow through the nozzle, the liquid moves at a steady slow pace on the order of m/s but the gas flow itself is choked in the narrow opening between the liquid/gas interface and the upper wall [138]. The formation of a shock associated with compound choking of the flow is viewed here as a trigger of acoustic oscillations of the liquid/gas mixture, which is at the origin of the LP signature

with characteristic period of 2.5 s observed in conjunction with the rapidly downgoing part of the sawtooth displacement signal shown in Fig. 7.

At Stromboli Volcano, Italy, broadband seismic signals including strong VLP and LP components are seen to accompany explosions. Waveform inversions of the VLP signals performed by Chouet et al. [32] show that the processes associated with eruptions involve volumetric changes marking a cycle of pressurization-depressurization-repressurization of the conduit, coupled with a single force component, both of which may be viewed as the result of a piston-like action of the magma associated with the disruption of a gas slug transiting through a sudden enlargement in conduit aperture in the steeply-inclined upper segment of conduit [32,60] (see Sect. “**Slug Distribution at Stromboli**” for details). The LP signal in that case is attributed to the oscillatory response of the shallowest segment of the fluid-filled conduit associated with the rapid expansion and ejection of the slug [29].

LP seismicity is also intimately linked with Vulcanian activity. The complex magma rheology, marked by a strong nonlinear dependence of viscosity on water content and temperature [49,147], finite yield strength, strain-rate dependent viscosity, and transition to brittle solid behavior at high strain rates [133] contributes to widely varying oscillatory behaviors seen in Vulcanian systems. At Galeras Volcano, Colombia, LP events displaying extended quasi-monochromatic coda were found to be linked to the release of ash-laden gases from a shallow magma body through a preexisting system of cracks bisecting the dome capping the vent [45]. Similar to Galeras, eruptive activity at Popocatepetl Volcano, Mexico, is dominated by emissions of steam, gas and ash, and by repeated formation and destruction of lava domes, all of which generate a wide variety of signals [10,11]. LP events accompanying degassing bursts are commonly associated with VLP signals and display waveform features that remain stationary from event to event. The source of these events coincides with, or is very close to, the source of VLP signals, 1.5 km below the crater floor [11,33]. This LP activity represents an integral part of a degassing process similar to that inferred for Galeras, where the sudden expulsion of a pressurized pocket of gases induces elastic deformation of the conduit coupled with acoustic resonance of the conduit (see further discussion in Sect. “**Coupled Diffusive-Elastic Pressurization at Popocatepetl Volcano**”).

LP events at Soufrière Hills, Montserrat, have been linked to the action of a repetitive, non-destructive, stationary source 1.5 km below the crater floor [100]. The temporal relationship between LP events and tilt patterns at Soufrière Hills clearly shows that LP seismicity starts in synchronicity with the passage of tilt through a turning point marked by a maximum in time derivative of the tilt signal. Seismicity subsequently ceases as tilt goes through a second turning point [100]. This correlation between LP events and inflation-deflation cycles is evidence of a strong relationship linking LP seismicity to magma movement and pressure conditions at depth. Based on these observations, [100] proposed a mechanism of LP excitation originating in the brittle failure of a highly viscous magma under high strain rate. In their model, loss of gas and heat from the magma at the conduit wall leads to the generation of a sharp viscosity gradient at the wall. This in turn leads to a buildup of shear stress near the wall, where the flowing magma undergoing glass transition can fail in a brittle manner. The resulting brecciated zone along the conduit boundary provides a natural pathway for escaping gases. The LP events may then be viewed as an expression of acoustic resonance occurring in well-defined channels in this zone, or possibly as an expression of acoustic energy

trapped in an adjacent region of conduit filled with bubbly magma [36,56].

The LP sources described so far all represent shallow sources occurring in the top 2 km below the surface. Although thoroughly documented, such sources do not represent the total expression of LP seismicity. An activity of deep (> 5 km depth) LP events has also been noted in several areas of the world. In particular, episodes of deep (30–60 km) and intermediate-depth (5–15 km) LP seismicity are often observed under Kilauea [74,75]. Analyses of the intermittency of LP activity were carried out by Shaw and Chouet [122] and Shaw and Chouet [123], who viewed such tremor episodes as the relaxation oscillations of a percolation network of coupled magma-filled fractures. In their interpretation, each individual fracture in the network acts as a source of LP events, and sets of fractures give rise to multiple LP events and episodes of tremor. The aggregate nature of fracture propagation and vibration then lead to the existence of universal scaling laws that give rise to the long-recognized frequency invariance of LP events [122,123].

Other episodic bursts have been noted at depths of 10–20 km under Mammoth Mountain, California, where LP and VLP events have been seen to occur synchronously with spasmodic bursts of small brittle failure earthquakes [54]. A process involving the transport of a slug of CO<sub>2</sub>-rich hydrous magmatic fluid derived from a plexus of basaltic dikes and sills has been invoked to explain such occurrences [54]. LP events at depths near 30 km under Mount Pinatubo, Philippines, have been attributed to a basaltic intrusion thought to have triggered processes leading to the cataclysmic eruption of June 15 [139]. Small LP events attributed to magmatic activity have also been detected at depths of 25–40 km beneath active volcanoes in northeastern Japan [47], and at depths of 10–45 km below volcanoes of the Aleutian Arc [114]. In all these instances, the occurrence of deep LP events appears to be more directly related to deep magma supply dynamics than near-surface or surface activity. Because of our lack of knowledge concerning the character and dynamics of deep-seated fluid transport under volcanoes, the actual origins of these events remain enigmatic.

### Summary of Inferred Excitation Mechanisms for LP Seismicity

Shallow hydrothermal LP events at Kusatsu-Shirane, Kilauea, and Mount St. Helens all share a common excitation mechanism involving the repeated pressurization of a steam-filled fracture, causing the venting of steam, collapse of the fracture and recharge, in response to heat

transfer from an underlying magma body. At Mount St. Helens, where the LP source is located very close to the magma conduit, LP activity is also seen to trigger a passive response of the conduit itself. LP seismicity observed at Redoubt represents a more energetic form of magmatic-hydrothermal interaction where an unsteady choked flow of magmatic gases provides a natural source of pressure perturbation at the origin of LP events and tremor. Magmatic LP events produced during bursts of ash-laden gases associated with Vulcanian activity at Galeras and Popocatepetl involve a pumping mechanism in shallow fractures similar to that inferred for hydrothermal LP events at Kusatsu-Shirane, Kilauea, and Mount St. Helens. LP events accompanying endogenous dome growth at Soufrière Hills appear to be related to a more complex process involving the production of shear fractures in highly viscous magma at the glass transition, injection of dusty gases into these fractures, and resonance of the fracture network and/or possibly resonance of a bubble-rich magma excited by the energy release from the brittle failure. Magmatic LP events in the basaltic systems at Kilauea and Stromboli are the result of pressure disturbances generated during the transit of large slugs of gas through conduit discontinuities.

A nonlinear excitation mechanism of LP seismicity by fluid flow has also been proposed by Julian [62]. Using a simple lumped-parameter model, Julian investigated the elastic coupling of the fluid and solid as a means to produce self-excited oscillations in a viscous incompressible liquid flowing through a channel with compliant walls. In his model, an increase in flow velocity leads to decrease of fluid pressure via the Bernoulli effect. As a result, the channel walls move inward and constrict the flow, causing an increase in fluid pressure and forcing the channel open again. The cyclic repetition of this process is the source of sustained oscillations in Julian's model. Julian demonstrated that with increasing driving pressure the model can exhibit various oscillatory behaviors resembling tremor.

As illustrated in the above examples, magmatic-hydrothermal interactions and magmatic oscillations can take on a variety of forms. Studies of the mechanisms of tremor excitation are still very much in their infancy, owing to the paucity of accurate and detailed seismic observations essential in documenting and analyzing such processes.

### Source Processes of Very-Long-Period Signals

Unlike the LP signals, which are mainly interpreted as manifestations of acoustic resonance, VLP signals are viewed as the results of inertial forces associated with per-

turbations in the flow of magma and gases through conduits. An increasingly widespread use of broadband seismometers on volcanoes has led to an increasing number of observations, and VLP signals have by now been documented in volcanic areas all over the world, including Sakurajima [67,68], Unzen [134], Aso [65,69,86,143], Satsuma-Iwojima [108,110], Iwate [101], Miyakejima [44,78], Usu [144], Bandai [103], Hachijo [81,83], and Asama [111] in Japan; Stromboli [14,30,32,35,99] in Italy; Merapi [50,51,52] in Indonesia; Kilauea [6,7,38,39,109] in Hawaii; Long Valley [53,54] and Mount St. Helens [137] in the United States; Erebus [13,119] in Antarctica; Popocatepetl [9,10,33] in Mexico; and Cotopaxi and Tungurahua [84] in Ecuador.

The wavelengths of VLP signals are in the range of tens to hundreds of kilometers, which greatly facilitates their analysis. Inversions of VLP waveforms have imaged crack geometries at the source in the forms of dikes or sills [32,81,109,143], as well as more complicated geometrical configurations involving a composite of a dike intersecting a sill [33], or composites of intersecting dikes [7,35], or two chambers connected to each other by a narrow channel [101]. Based on the characteristic period of VLP signals recorded at Hachijo Island, Japan, Kumagai [83] estimated a source with dimensions of a few km. The decaying harmonic oscillations with periods near 10 s and duration up to 300 s of the VLP signals seen at Hachijo are similar to the features of LP events with periods near 1 s (see Fig. 3) and appear consistent with the resonance of a dike containing a bubbly basalt. In that sense, these VLP signals probably represent a unique end member of the family of sources associated with resonant conduit excitation. Interestingly, the dominant periods and associated decay characteristics of the VLP signals at Hachijo were both found to vary with time, and these temporal variations were attributed by Kumagai [83] to a change in source dimensions, assuming fixed acoustic properties of the fluid at the source. A spread of about 4 km was also estimated by Chouet et al. [101] for two chambers at Iwate Volcano. The modeling of VLP waveforms at Stromboli by Nishimura et al. [35] points to two point sources, each involving two intersecting dikes, whose relative positions and orientations represent individual conduit segments with dimensions of a few hundred meters. Systematic modeling of sustained VLP tremor accompanying episodes of deflation-inflation-deflation at Kilauea has suggested a complex magma pathway composed of a plexus of intersecting dikes extending over the depth range 0.6–1.6 km below the summit caldera [7]. A crack-like conduit extending over the depth range 0.3–2.5 km has also been inferred under Aso Volcano, Japan [143].

Contrasting the above-discussed findings are results obtained by Ohminato [111] from waveform inversions of VLP signals produced by Vulcanian explosions at Asama Volcano, Japan. At Asama, the contribution from a vertical single force with magnitude  $10^{10} - 10^{11}$  N was found to dominate the observed waveforms, and no obvious volumetric component was identified at the source. The depth of the VLP source is  $\sim 200$  m beneath the summit crater, and the source-time history features two downward force components separated by an upward force component. The initial downward component was attributed by Ohminato [111] to the sudden removal of the lid capping the pressurized conduit (this force is analogous to the force  $F_{T_1}$  in Fig. 2), and the subsequent upward force was interpreted by these authors as a drag force induced by viscous magma moving up the conduit. Ohminato [111] also attributed the final downward force component to an explosive fragmentation of the magma, whose effect is to effectively cancel viscous drag so that the downward force due to jet recoil again dominates the VLP signal.

Available studies demonstrate that detailed investigations of the VLP oscillation characteristics of the source are critically important to our understanding of volcanic fluid dynamics. Below, I review two in-depth investigations of VLP signals representative of Vulcanian degassing bursts at Popocatepetl and gas-slug transport dynamics at Stromboli.

### Coupled Diffusive-Elastic Pressurization at Popocatepetl Volcano

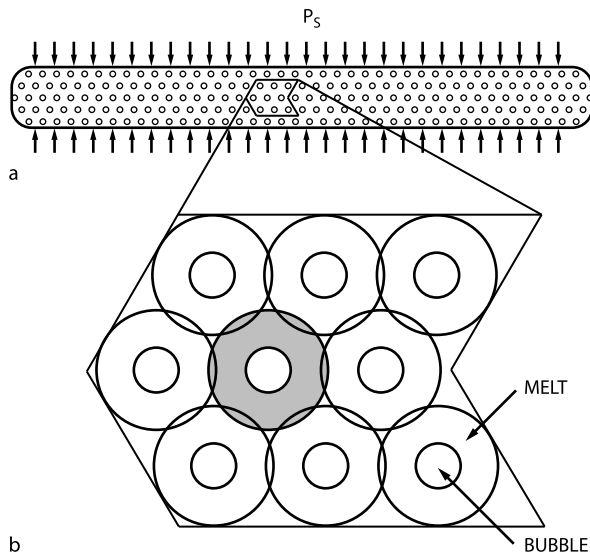
Vulcanian degassing bursts at Popocatepetl Volcano, Mexico, have been observed to be closely linked to the degassing of a sill-shaped volume of magma at a depth of 1.5 km below its summit crater [33]. The variations in volume of this sill were interpreted by Chouet et al. [33] as reflecting cyclic pressure oscillations originating in the magma filling the sill. Based on these observations, these authors proposed a model in which static magma in the sill becomes supersaturated because of groundmass crystallization. In this conceptual view, volatile exsolution and diffusion of gas from the melt into bubbles increase the internal pressure of the bubbles, because bubble expansion is impeded by the viscous resistance of the surrounding liquid and by the confining effects due to the finite yield strength of the overlying column of magma and surrounding solid rock. Elastic inflation of the sill occurs as a result of bubble pressurization, and this inflation proceeds until the critical yield strength of the magma column is exceeded and magma starts flowing out of the sill. Magma fragmentation induced by viscous shear near the conduit

wall then causes coalescence and collapse of bubbles intersected by fractures, allowing gas escape through a transient network of fractures. This, in turn, induces a pressure decrease in the sill, which results in the collapse and welding of the fracture network that shuts down the gas-escape pathway. Repeated cycles of shear-induced fracture and welding of magma provide a ratchet mechanism by which the separated gas phase in the magma can be recharged and evacuated.

The expansion dynamics of bubbles in supersaturated magma have been addressed in many previous studies under conditions of constant melt pressure [87,88,90,115,117,121,125], or constant decompression rate [116,131,132]. A canonical model of bubble growth in magma that includes the effect of finite spacing of bubbles in the melt was first developed by Proussevitch et al. [115]. This model considers a suspension of gas bubbles in an incompressible volatile-bearing liquid. The suspension is modeled as a three-dimensional lattice of closely-packed spherical cells, where each identical elementary cell is composed of a gas bubble surrounded by a shell of liquid. A step drop in pressure in the melt induces volatile exsolution and bubble expansion is driven by the diffusion of gas into the bubble. In this model, only the volatiles contained in the shell of liquid surrounding the bubble contribute to the bubble expansion so that bubble growth is limited by how closely packed the bubbles are in the liquid. Using the model of Proussevitch et al. [115], Chouet et al. [34] and Shimomura et al. [124] obtained a dynamic solution that includes the effect of melt compressibility. In their model, pressure recovery is driven by bubble growth in a supersaturated magma that is stressed by the surrounding crust as the magma volume increases with bubble expansion. Both models consider the pressure recovery following an instantaneous pressure drop in the melt. The magma is embedded in an infinite, homogeneous, elastic solid and consists of a melt containing numerous small spherical gas bubbles of identical size. No new bubbles are created and no bubbles are lost during pressure recovery, and the gas in the bubbles is assumed to be a perfect gas. Gravity and other body forces are not considered.

The conduit geometry considered by Chouet et al. [34] applies specifically to the observation made at Popocatepetl [33] and consists of a penny-shaped crack containing a melt with an isotropic distribution of bubbles (Fig. 9a). The initial oversaturation of the melt resulting from a pressure drop  $\Delta P_0$  is distributed uniformly in the melt, and each bubble grows by diffusion of volatiles from a surrounding shell of melt with finite radius (Fig. 9b). As diffusion is a slow process, with time scale much longer than that of heat transfer into the bubble [56], bubble





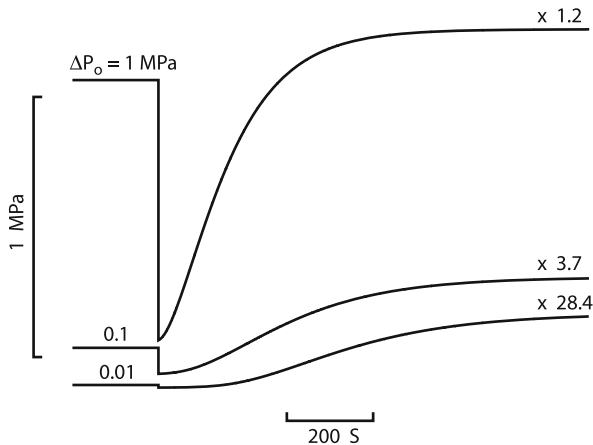
Volcanoes, Non-linear Processes in, Figure 9  
**Model used by Chouet et al. [34] to interpret the source mechanism of Vulcanian degassing bursts at Popocatepetl Volcano, Mexico. a** Schematic illustration (cross section) of penny-shaped crack embedded in an elastic solid. The crack is under confining pressure  $P_s$ , and contains a melt with an isotropic distribution of gas bubbles of identical sizes. **b** Detailed view of the distribution of elementary cells composing the bubbly liquid (after [115]). Each cell (as in gray-shaded example) consists of a gas bubble surrounded by a finite volume of melt. The elementary cells are organized in a three-dimensional lattice with slight overlap of the cells, where the volumes of intersecting melt and gaps are equal, so that the gas-volume fraction is  $(R/S)^3$ , where  $R$  is the bubble radius and  $S$  is the outer shell radius

growth is assumed to be isothermal. The mathematical model describing an elementary cell therefore consists of three equations: (1) a diffusion equation governing the transfer of volatiles in the melt shell; (2) an equation expressing the radial motion of the bubble; and (3) a relation representing the ideal gas approximation. These equations are then subjected to three boundary conditions expressing the phase equilibrium at the bubble wall, mass flux at the bubble wall, and zero mass flux through the outer shell wall. As the bubble grows by mass transfer from the melt, melt is compressed and the surrounding rock matrix is deformed as a result. The pressure in the melt is balanced by the stress applied by the surrounding elastic medium, and in both Chouet et al. [34] and Shimomura et al. [124], this is formalized by assuming a quasi-static deformation of the magma conduit. Melt shrinkage from dehydration [105] is assumed to be negligible compared to melt compression due to bubble growth and is not considered.

In their treatment of the initial condition in the bubble, Chouet et al. [34] follow Proussevitch et al. [115] and assume an instantaneous drop in gas pressure inside the bubble. This induces an instantaneous drop in volatile concentration at the bubble wall, thereby producing a step-like gradient of concentration that kick-starts the diffusion process. This condition is different from that assumed by Shimomura et al. [124], who follow the model of Navon and Lyakhovsky [98], in which the bubble initially retains the original gas pressure and immediately expands in response to the difference between internal and external pressures. The relaxation expansion of the bubble then induces a decrease in gas pressure inside the bubble, which leads to the establishment of a gradient in volatile concentration at the bubble wall that starts the diffusion. Although this leads to distinct features in the instantaneous bubble response, this has essentially no effect on the volumetric response of the bubble at longer time scales.

Figure 10 shows the pressure recovery calculated by Chouet et al. [34] for input step drops  $\Delta P_0 = 0.01, 0.1,$  and  $1$  MPa applied to a bubbly rhyolitic melt encased in a penny-shaped crack with radius  $100$  m and thickness  $5$  m under ambient pressure  $P_s = 40$  MPa appropriate for the depth of the source imaged for Popocatepetl Volcano [33]. The bubbles have a fixed initial radius  $10^{-6}$  m, and the bubble number density is  $10^{12} \text{ m}^{-3}$ . Melt density, viscosity, and diffusivity are  $2300 \text{ kg m}^{-3}$ ,  $10^6 \text{ Pa s}$ , and  $10^{-11} \text{ m}^2 \text{ s}^{-1}$ , respectively; surface tension is  $0.2 \text{ Nm}^{-1}$ , and the bulk modulus of melt and elastic rigidity of the rock are both  $10^{10} \text{ Pa}$ . The amplitude ratio of pressure recovery to initial pressure drop displays a strong nonlinear sensitivity to the magnitude of the input pressure drop, resulting in magnifications ranging from about  $1$  to nearly  $30$  over the range of input transients considered (Fig. 10). This overpressurization upon recovery was previously noted in a static solution obtained earlier by Nishimura [104] and becomes largest for tiny bubbles and a stiff elastic medium [34,124]. Figure 10 also indicates that the speed of recovery is a function of the input transient, with the crack response becoming markedly slower for small pressure transients compared to larger ones.

The variation of time and amplitude scales seen in Fig. 10 suggests that a wide range of responses may be possible for cracks containing melts with a wider population of bubble sizes than the single-size bubble populations considered by Chouet et al. [34]. The net drop in volatile concentration associated with diffusion-driven bubble growth is quite small. Indeed, the results obtained by Chouet et al. [34] point to concentration drops  $\sim 0.003 - 0.035 \text{ wt\%}$  for initial bubble radii of  $10^{-5} - 10^{-7}$  m in penny-shaped cracks with aperture to radius ratios

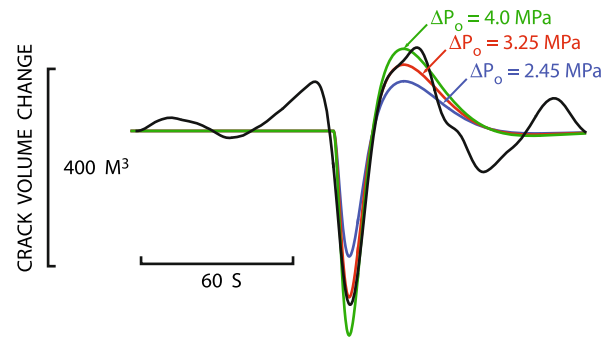


Volcanoes, Non-linear Processes in, Figure 10  
Dependence of pressure recovery on the size of the input transient,  $\Delta P_0$ , in the model of Chouet et al. [34]. The magnification factor for each response is indicated at the upper right of each pressure trace

0.01 – 0.25. The net concentration drop is small because pressure recovery in the melt acts to raise the saturation level of volatiles in the melt. This has important implications for the overall history of degassing in a Vulcanian system, as it suggests that this process may be repeated many times without significantly depleting the gases in the melt body.

The source process imaged under Popocatepetl involves a sequence of inflation, deflation, and reinflation, reflecting a cycle of pressurization, depressurization, and repressurization within a time interval of 3–5 min [33]. The volumetric component of the source processes in the sill associated with an eruption on 23 May 2000 is shown in Fig. 11. Other eruptions were found to produce similar waveform characteristics, hence the event in Fig. 11 may be viewed as an appropriate representation of overall source dynamics associated with these Vulcanian eruptions. For comparison, Fig. 11 also shows the volume change of a magma-filled sill calculated with the model of Chouet et al. [34]. The sill response has been band-pass filtered in the same band as the observed signal, and is illustrated for three distinct pressure drops. The model parameters in Fig. 11 are identical to those used in Fig. 10. A realistic fit of the peak-to-trough amplitude of the volume change imaged at Popocatepetl is obtained for an input pressure drop  $\Delta P_0 = 3.25$  MPa.

Unfortunately, the longer-period components and other specific features of the sill response in the model are lost in the band-limited version of the signal compatible with observations made at Popocatepetl, so that the responses obtained in this band for different model param-



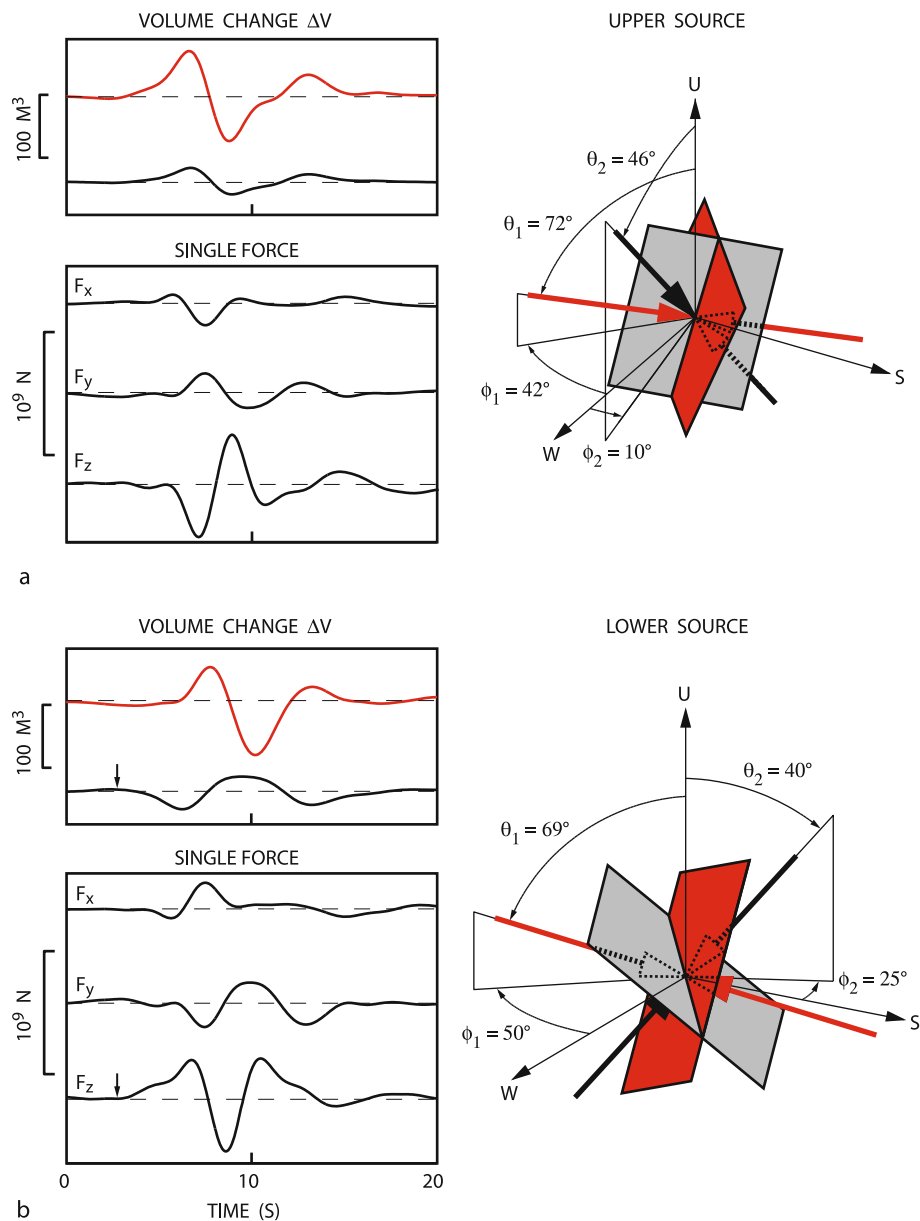
Volcanoes, Non-linear Processes in, Figure 11  
Source-time function of volume change obtained by Chouet et al. [33] for a sill under Popocatepetl Volcano during a degassing burst on 23 May 2000 (black line), and corresponding volume changes in a sill filled with a bubbly melt in response to a step drop  $\Delta P_0$  in pressure (colored lines). The data from Popocatepetl and solutions from the crack model have both been band-pass filtered in the 15–70 s band

eters are all very similar [34]. Although it is not possible to infer specific values of melt viscosity, volatile diffusivity, initial bubble radius, bubble number density, or sill aperture to radius ratio based on available seismic data, these results strongly support the idea that diffusion pumping of bubbles in the magma under Popocatepetl provides a viable mechanism for pressure recovery following a pressure drop induced by a degassing event.

### Slug Disruption at Stromboli

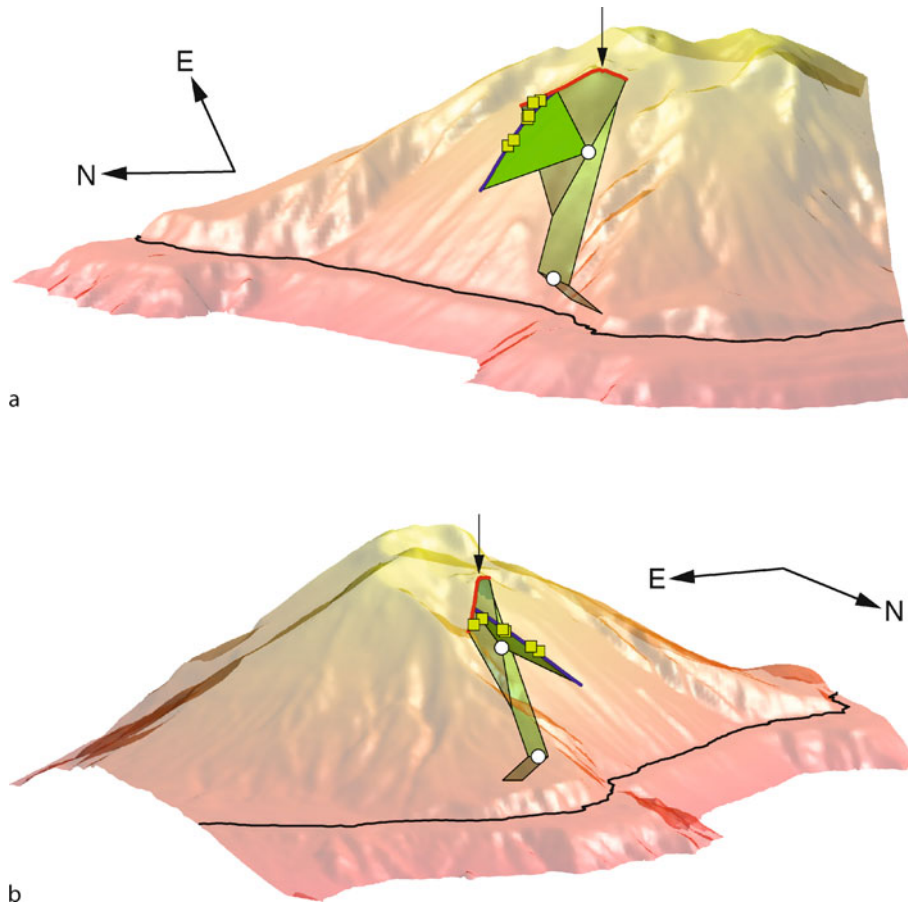
Eruptive behavior at Stromboli is characterized by mild, intermittent explosive activity, during which well-collimated jets of gases laden with molten lava fragments burst in short eruptions typically lasting 5–15 s. Modeling of VLP seismic data recorded during explosive activity in 1997 has imaged two distinct dike structures representative of explosive eruptions from two different vents located near the northern and southern perimeters of the summit crater [32,35].

Figures 12 and 13 show two representations of the conduit geometry underlying the northern vent area. Figure 12 shows the seismic source mechanisms obtained from inversion of VLP waveforms; the mechanisms include two point sources, each of which marks a flow disruption site along the upper conduit. The upper source (Fig. 12a) is located 300 m below the crater floor and represents a bifurcation in the conduit; this is the main flow disruption site for gas slugs ascending toward the northern vent. The lower source (Fig. 12b), located roughly 500 m below the upper source, involves a sharp corner in the conduit and represents a secondary flow disruption site.



Volcanoes, Non-linear Processes in, Figure 12

Seismic source mechanisms imaged for the two flow disruption sites in the shallow conduit structure underlying the northern vent area at Stromboli. The two point-source mechanisms were obtained by Chouet et al. [35] from inversions of VLP waveforms recorded during explosions at Stromboli, and are positioned at different depths in the volcanic edifice (see text for details). Each source consists of two intersecting cracks and a single force with components  $F_x$  (east),  $F_y$  (north), and  $F_z$  (up). Volume changes are color-coded with the color of the cracks they represent in each source. Crack orientations are provided by the azimuth  $\phi$  and polar angle  $\theta$  of the dominant dipole normal to the crack plane, with arrow directions marking crack deflation (see Fig. 1c). The reference coordinates are W (west), S (south), and U (up). **a** Upper source. **b** Lower source. Small arrows in left panels mark the onset of deflation of the lower dike (gray dike in right panel) and synchronous start of the upward force. (Reproduced from [35])



Volcanoes, Non-linear Processes in, Figure 13

Geometry of the upper 1 km of conduit underlying the northern vent area of Stromboli. A semi-transparent view of the northwest quadrant of the volcanic edifice provides the reference for the location and geometry of the conduit, which is derived from the seismic source mechanisms in Fig. 12. Thin black line indicates sea level. The summit of the volcano is 924 m above sea level (no vertical exaggeration). The two flow disruption sites that are sources of VLP elastic radiation are indicated by small circles. The irregular red and blue lines, respectively, represent the surface traces of the dominant and subdominant dike segments constituting the shallowest portions of the conduit system. The eruptive vent is marked by an arrow, and vents temporarily active during the flank eruption in 2002–2003 are marked by green squares. The lateral extents of individual dike segments are unknown and are shown for illustrative purpose only. **a** East-looking view. **b** South-looking view

Both upper and lower sources feature a dominant crack sustaining the largest volume change (colored red), and a subdominant crack undergoing a smaller volume change (shaded gray). Both cracks in Fig. 12a display a similar sequence of inflation-deflation-inflation. The dominant crack in the lower source displays a volumetric response similar to that seen in the dominant crack at the upper source, but delayed by about 1 s with respect to the upper source. This implies a propagation speed of roughly 500 m/s between the two sources, consistent with the slow speed expected for the crack wave.

Figure 13 shows a picture of the upper conduit geometry consistent with the seismic source mechanisms imaged

in Fig. 12. The closely matching dips of the two dominant cracks in Fig. 12 point to a conduit that extends essentially straight from 80 m below sea level to the crater floor, 760 m above sea level. At a depth of 80 m below sea level the conduit features a sharp corner leading into a dike segment dipping  $40^\circ$  to the southeast. The upper dominant dike segment, and deep segment below the abrupt corner both strike northeast-southwest along a direction parallel to the elongation of the volcanic edifice and a prominent zone of structural weakness, as expressed by lineaments, dikes, and brittle structures. The surface trace of the main dike segment trends through the northern vent area, while that of the upper subsidiary segment extends

northwest-southeast in rough alignment with several vents active in the northwest quadrant of Stromboli in 2002–2003 [1]; the subsidiary dike trace intersects the main dike trace  $\sim 170$  m north of the northern vent area.

A striking aspect of the mechanisms imaged for the two sources in Fig. 12 is the presence of dominantly vertical single-force components with common-looking time histories, except for a polarity reversal in one source compared to the other. The upper source (Fig. 12a) displays an initially downward force followed by an upward force, while an upward force followed by a downward force is manifest in the lower source (Fig. 12b). These force components compensate each other so that the total momentum in the overall source volume is conserved. In contrast to the delay of  $\sim 1$  s in the volumetric components noted above, no significant delay is noted in the onsets of the vertical forces at the two sources, suggesting that transmission of the force between the two sources occurs via the faster speed (3.5 km/s [32]) of the compressional wave in the rock matrix.

Laboratory simulations carried out by James et al. [60] provide insights into the origin of the initial pressurization and downward force observed at the upper source in Fig. 12a. These experiments investigate the ascent of a slug of gas in a vertical liquid-filled tube featuring a flare that abruptly doubles the cross sectional area. The tube is instrumented with pressure transducers mounted flush with the inner tube wall, and one accelerometer mounted on the exterior of the tube, and the whole assembly is free to move in the vertical direction. Detailed measurements of the flow transients obtained by James et al. [60] show that the transit of a gas slug through the tube flare involves complex changes in flow pattern. A characteristic pinching of the slug tail is observed to occur synchronously with strong pressure and acceleration transients at the time the slug clears the flare, a picture consistent with the downward and inward motion of a liquid piston formed by the thickening film of liquid falling past the slug expanding in the wider tube. The sudden deceleration of the liquid annulus as it impinges the narrower inlet to the lower tube segment generates a pressure pulse in the liquid below the flare and also induces a downward force on the apparatus. These observations are consistent with the pressurization phase and initial downward force imaged for the upper source at Stromboli and a similar funneling mechanism was inferred by Chouet et al. [35] to be operative there. The repeatability of recorded pressure data and dependence of the magnitude of the pressure transient on slug size seen in the experiments of James et al. [60] are also in harmony with the observed spatio-temporal properties of VLP signals at Stromboli [32,35].

At the lower source, the start of the vertical force signal is synchronous with the onset of deflation of the lower dike segment (see arrows in the left panels in Fig. 12b). As the amplitude of the upward force increases, the lower dike segment continuously deflates. During the same interval, the dominant dike (red-colored volumetric trace in Fig. 12b) remains in a slightly deflated state. The lower dike reaches maximum deflation at the time the upper dike segment goes through a transition from weak contraction to expansion, and the upward force reaches its peak amplitude  $\sim 0.5$  s later. This picture is consistent with a compression of the lower dike synchronous with a downward acceleration of the liquid mass, both of which are suggestive of increasing external pressure on the conduit wall resulting from the downward vertical force acting at the upper source. Compression of the lower dike segment proceeds unimpeded until this process is overprinted by the arrival of the much slower volumetric expansion signal from the upper source.

Although not illustrated here, a similar slug disruption mechanism was imaged in the dike system underlying the southern vent [35]. In both conduit systems, the early response of the lower source relative to the volumetric disturbance arriving from the upper source may be interpreted as the passive response of the liquid to the movement of the conduit wall induced by elastic radiation from the force acting at the upper source. The scenario emerging from these dynamics is that of an upper source representing an active fluid phase and passive solid phase, and a lower source representing an active solid phase and passive fluid phase. The overall seismic source process associated with eruptions at Stromboli may then be summarized as follows. A slug of gas formed in the deeper reaches of conduit [18] rises through the lower conduit corner (the lower seismic source). At this point, the slug is most likely a few meters long and traverses this corner aseismically; past this corner, the slug expands on its way to the upper conduit bifurcation (the upper seismic source). The transit time from the lower to the upper seismic source is probably in the range of 5–15 minutes [60], hence related changes in magmatic head are well beyond the capability of broadband seismometers to detect and are not apparent within the VLP band imaged in Fig. 12. As it traverses the upper seismic source, the slug length has expanded to tens of meters and the slug is by then seismically noisy. Gravitational slumping of the liquid occurs as the slug expands through a flare in the conduit at this location. The slumping liquid rapidly decelerates in the narrowing dike neck, increasing the liquid pressure and inducing a volume expansion in the main conduit and its subsidiary branch. The rapid deceleration and associated pressurization of

the liquid couples to the conduit wall via the flare shoulders and induces a downward vertical force on the Earth. The volumetric signal propagates along the conduit at the slow speed of the crack wave, while the force signal itself propagates in the solid at the much higher speed of the compressional wave in the rock matrix, arriving at the lower conduit corner well before the crack wave. At this corner, the downward displacement of the rock induced by the force acting at the upper source impinges the bottom dike and squeezes this segment of conduit; this segment essentially acts like a spring that absorbs the downward motion of the rock. This scenario is consistent with both the small volume change of the lower dike segment, as well as its early response. The subsequent conduit response then reflects the combined effects of volumetric and mass oscillations of this liquid/gas/solid system, which are damped and eventually terminated by changing flow conditions.

The analyses carried out by Chouet et al. [35] illuminate the subsurface processes driving eruptions at Stromboli and clearly point to the key role played by the conduit geometry in controlling fluid motion and resultant processes. Each irregularity or discontinuity in the conduit provides a site where pressure and momentum changes resulting from flow processes associated with the transit of a gas slug through the discontinuity are coupled to the Earth, or where the elastic response of the conduit can couple back into pressure and momentum changes in the fluid. The resulting processes are naturally oscillatory and involve diverse dynamics that reflect the complex physico-chemical behavior of the volcanic system from the surface downward.

### Future Directions

Seismology alone cannot directly see into the conduit and resolve details of the actual fluid dynamics at the origin of the seismic source mechanisms revealed by analyses of LP or VLP signals. To develop a better understanding of fluid behavior responsible for these signals, laboratory experiments are required to explore the links between known flow processes and the resulting pressure and momentum changes. The pressure and momentum changes generated under laboratory conditions may then be compared with the pressure and momentum changes estimated from the time-varying moment-tensor and single-force components imaged from seismic data, yielding clues about the physical flow processes linked to the seismic source mechanism. Recent laboratory studies [59,60,85] have elucidated self-excitation mechanisms inherent to the fluid nonlinearity that are providing new

insights into the mechanisms imaged from seismic data. In particular, the results obtained by James et al. [60] demonstrate that direct links between the moment-tensor and single-force seismic-source mechanism and fluid-flow processes are possible and could potentially provide a wealth of information not available from seismic data alone. Together with these laboratory advances, numerical studies of multiphase flows are required to shed light on both the micro- and macrophysics of such flows (e. g., Badalassi [15]), along with models exploring the coupled dynamics of fluid and solid [102]. The key to a better understanding of volcanic processes lies in a sustained effort aimed at cross-fertilization between increasingly realistic numerical and experimental models of the fluid dynamics and elastodynamics, spatially and temporally dense field measurements of diverse geophysical signals at all frequencies, and chemical and physical evidence recorded in the eruptive products. In the next section, I present a brief summary of a modeling approach to two-phase fluids that holds great promise for the quantification of volcanic and hydrothermal source processes.

### Phase-Field Method

Liquid-gas mixtures are an ubiquitous feature of magmatic and hydrothermal systems. A major challenge in addressing these types of fluids is the description of the moving and deforming interface between the two components of fluid. Traditional fluid dynamics treats these as sharp interfaces on which matching boundary conditions must be imposed, which generally leads to intractable problems. Another difficulty is posed by the complex rheology of each component, whose internal microstructure is coupled with the flow field. The *phase-field method*, also known as the *diffuse-interface model*, circumvents these difficulties by considering the interface between components as a thin diffuse layer within which the two components are mixed and store a mixing energy. This idea follows from the original work of van der Waals [135] and expresses the properties of the interface by molecular forces and a mixing energy. In the limit of an interface width approaching zero, the diffuse-interface model reduces to the classical sharp-interface model and results in the proper expression for interfacial tension. Lowengrub and Truskinovsky [89] provide a detailed formulation of the conservative dynamics of the diffuse-interface model based on the classical procedure of Lagrangian mechanics. A review of diffuse-interface methods and related applications can also be found in Anderson et al. [8]. Noteworthy are the more recent applications by Kendon et al. [71], Xu et al. [141], Xu et al. [142], and Yue et al. [146]. The following

discussion is restricted to a brief description of the basic principles of the method based on the work of Kendon et al. [71], who consider the evolution of a symmetric binary fluid mixture after a deep quench.

**Equilibrium State** A two-phase mixture is described by two scalar fields  $n_A$  and  $n_B$  expressing the local molar densities of the fluid components  $A$  and  $B$ . The corresponding local density,  $\rho$ , is given by:

$$\rho = m_A n_A + m_B n_B, \quad (2)$$

where  $m_A$  and  $m_B$  are the molecular weights of phases  $A$  and  $B$ , respectively. The local composition of the fluid is quantified by a dimensionless parameter,  $\phi$ ,

$$\phi = \frac{n_A - n_B}{n_A + n_B}, \quad (3)$$

called the order parameter; the values of  $\phi$  span the range  $[-1, +1]$  with the value  $+1$  denoting the pure phase  $A$  and the value  $-1$  denoting the pure phase  $B$ .

The chemical equilibrium between the two phases is described through a minimization of the free energy  $\mathcal{F}$ , which represents the integral over the body of the local free-energy density (with units of  $\text{Jm}^{-3}$  or, equivalently,  $\text{Nm}^{-2}$ ), a function of the fluid composition and its gradient [19]:

$$\mathcal{F} = \int \left\{ f_0(\phi, P) + \frac{1}{2} \kappa (\nabla \phi)^2 \right\} \text{d}\mathbf{r}. \quad (4)$$

The term  $f_0(\phi, P)$  in the above equation represents the bulk free-energy density, which is assumed to take the form of a double-well with respect to  $\phi$ . The detailed functional form of the double-well potential is not important [17,71,127,128,141,142,146] and may be set as [71,141]:

$$f_0(\phi, P) = \frac{a}{2} \phi^2 + \frac{b}{4} \phi^4 + P \ln \frac{P}{P_0}, \quad (5)$$

where the polynomial terms represent the bulk properties of the fluid. The parameter  $a < 0$  and parameter  $b$  is always positive [141]. The term in  $P$  yields a positive background pressure and does not affect the phase behavior (see discussion of pressure below). The form of this latter term is selected here for convenience, with  $P_0$  representing an arbitrary reference pressure. As defined above, the function  $f_0$  features two minima for  $\phi = \pm \sqrt{-a/b}$  that correspond to the coexisting pure bulk phases. As-

suming  $b = -a$ , one obtains the equilibrium values  $\phi = \pm 1$  [141].

The squared gradient term in Eq. (4), with  $\kappa$  a positive constant, represents weakly non-local interactions between the components. The effect of these interactions favors mixing of the components in contrast to the tendency of the bulk free energy  $f_0$ , which is toward total separation of the phases into domains of pure components [146].

According to the model represented by Eq. (5), minimization of the free energy is achieved through the creation of two bulk domains at compositions  $\phi = \pm 1$  separated by interfaces across which the composition varies smoothly from one phase to the other. The change in  $\mathcal{F}$  induced by a small local change in composition is described by the chemical potential,  $\mu$ ,

$$\mu = \frac{\delta \mathcal{F}}{\delta \phi}, \quad (6)$$

and equilibrium between the coexisting bulk phases is reached when  $\mu$  is everywhere zero [19,71]. Using Eqs. (4) and (5), one obtains the following expression for  $\mu$ :

$$\mu = a\phi + b\phi^3 - \kappa \nabla^2 \phi. \quad (7)$$

The equilibrium interfacial profile is given by  $\phi(x) = \tanh(2x/\xi)$ , where  $x$  is the coordinate normal to the interface [141]. With the choice  $b = -a$ , this yields an interfacial width,  $\xi = 2\sqrt{-2\kappa/a}$ , and surface tension,  $\sigma = (2/3)\sqrt{-2a\kappa}$  [141]. The interface thickness may then be defined as the width over which 90% of the variation of  $\phi$  occurs; this yields an equilibrium interface thickness of  $1.47222\xi$ .

As interfaces in the fluid can exert non-isotropic forces, pressure is a tensor. The thermodynamic pressure tensor  $P_{ij}^{\text{th}}$  is obtained from the free energy as [41,71,127]

$$\begin{aligned} P_{ij}^{\text{th}} &= \left( P \frac{\delta \mathcal{F}}{\delta P} + \phi \frac{\delta \mathcal{F}}{\delta \phi} - \left[ f_0(\phi, P) + \frac{1}{2} \kappa (\nabla \phi)^2 \right] \right) \delta_{ij} \\ &\quad + \kappa \frac{\partial \phi}{\partial x_i} \frac{\partial \phi}{\partial x_j} \\ &= P + P_{ij}^{\text{chem}}, \end{aligned} \quad (8)$$

where  $P$  is the bulk pressure, and  $P_{ij}^{\text{chem}}$  is the chemical pressure component given by

$$\begin{aligned} P_{ij}^{\text{chem}} &= \left[ \frac{1}{2} a \phi^2 + \frac{3}{4} b \phi^4 - \kappa \phi \nabla^2 \phi - \frac{1}{2} \kappa (\nabla \phi)^2 \right] \delta_{ij} \\ &\quad + \kappa \frac{\partial \phi}{\partial x_i} \frac{\partial \phi}{\partial x_j}, \end{aligned} \quad (9)$$

in which  $\delta_{ij}$  is the Kronecker symbol ( $\delta_{ij} = 0$  for  $i \neq j$ , and  $\delta_{ij} = 1$  for  $i = j$ ). Note that the expression be-

tween brackets contributes to the isotropic fluid pressure and that only the last term is anisotropic.

**Conservation Equations** The temporal evolution of the fluid composition  $\phi$  is described by the Cahn–Hilliard diffusion-advection equation [20]

$$\frac{\partial \phi}{\partial t} + \mathbf{v} \cdot \nabla \phi = \nabla \cdot (\gamma \nabla \mu), \quad (10)$$

where  $\gamma$  is the order-parameter mobility and  $\mathbf{v}(\mathbf{r})$  is the fluid velocity.

The dynamics of the fluid are described by the Navier–Stokes equations, whose exact forms depend on the system considered. The equations describing a viscous, compressible, isothermal flow are the conservation of mass:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0, \quad (11)$$

and conservation of momentum:

$$\rho \left[ \frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla) \mathbf{v} \right] = -\nabla \cdot \mathbf{P}^{\text{th}} + \nabla \cdot \boldsymbol{\tau} + \rho \mathbf{g}, \quad (12)$$

where  $\mathbf{g}$  is the acceleration of gravity vector,  $\mathbf{P}^{\text{th}}$  is the thermodynamic pressure tensor with components given by Eqs. (8) and (9), and  $\boldsymbol{\tau}$  is the viscous stress tensor, which for a newtonian fluid is given by

$$\tau_{ij} = \lambda e_{kk} \delta_{ij} + 2\eta e_{ij}, \quad (13)$$

where  $\eta$  is the coefficient of viscosity (dynamic viscosity) and  $\lambda$  is the second coefficient of viscosity generally defined as  $\lambda = -(2/3)\eta$  [130]. Summation over repeated indices is implied. In this equation  $e_{ij}$  is the strain rate tensor given by:

$$e_{ij} = \frac{1}{2} \left( \frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i} \right). \quad (14)$$

The field variables are the order parameter  $\phi$ , the three components of fluid velocity  $\mathbf{v}$ , density  $\rho$ , and bulk pressure  $P$ . Under isothermal conditions  $P$  may be specified through a van der Waals equation of state. The diffusion-advection Eq. (10), continuity Eq. (11), and the three momentum Eqs (12), along with an equation of state for  $P$ , therefore provide a complete description of the physico-chemical evolution of the binary fluid mixture. All other quantities can be obtained from the relations (2), (3), (7), (8), (9), (13), and (14). For non-isothermal flow, a complete description is obtained from Eqs. (10–12), along with two additional equations describing the conservation of energy and entropy production [8].

The thermodynamic properties of the two-phase system enter the above equations through the chemical potential  $\mu = \delta \mathcal{F} / \delta \phi$ , and pressure tensor  $\mathbf{P}^{\text{th}}$ , which are both obtained from the free energy Eq. (4). As a result, interfaces between the different fluid phases appear naturally within the model and do not need to be put in as boundary conditions. The procedure is quite general and may conceivably be applied to any complex fluid with a properly defined free energy.

The coupled Cahn–Hilliard/Navier–Stokes system of equations representing binary fluid systems is analytically intractable, and efforts have mainly been directed toward the development of robust, stable numerical schemes. The description of the fluid in terms of a smooth variation of the composition  $\phi(\mathbf{r})$  provides a coarse-grained representation of the fluid where the smallest length scale is larger than the average distance between molecules. Under dynamical conditions, care must be taken to ensure that this coarse-grained description adequately represents the underlying microscopic physics. For the nominal interfacial thickness representing 90% of the variation of  $\phi$  this typically requires 7–10 grids to resolve [146]. Numerical applications include the Lattice Boltzmann Method (LBM) [16,71,112,113,127,128,141,142], and various finite-difference approaches based on an Implicit-Explicit (IMEX) discretization [12], semi-implicit discretization [15,146], fully implicit discretization [73], or fully explicit central-differenced staggered-grid discretization [58].

An example of application of this method that may provide a useful metaphor for the process of Vulcanian degassing discussed in Sect. “**Coupled Diffusive-Elastic Pressurization at Popocatepetl Volcano**” is the 3D spinodal decomposition of a density-matched binary fluid mixture in a channel under shear [15]. A remarkable feature of this type of flow is the formation of string-like structures. Similar structures have been observed in immiscible viscoelastic systems subjected to complex flow fields [93], and such process may also play a role in the formation of degassing channels in a rhyolitic bubbly magma driven by slow pressurization in response to crystallization and degassing. A 2D application of the method to the dynamics of a gas slug ascending in a liquid-filled conduit featuring a flare is described in the chapter by *D’Auria and Martini*. The numerical results obtained by these authors reproduce the basic behavior observed in the experimental simulations of James et al. [59] and James et al. [60] and hold great potential as a key link between the seismic source mechanisms imaged at Stromboli (see Sect. “**Slug Disruption at Stromboli**”) and the underlying volcanic fluid dynamics.



## Acknowledgment

I am grateful to Phil Dawson for his assistance in drafting figures. I am indebted to Robert Tilling and David Hill for careful reviews and helpful suggestions.

## Bibliography

1. Acocella V, Neri M, Scarlato P (2006) Understanding shallow magma emplacement at volcanoes: Orthogonal feeder dykes during the 2002–2003 Stromboli (Italy) eruption. *Geophys Res Lett* 33:L17310. doi:10.1029/2006GL026862
2. Aki K, Fehler M, Das S (1977) Source mechanism of volcanic tremor: Fluid-driven crack models and their application to the 1963 Kilauea eruption. *J Volcanol Geotherm Res* 2:259–287
3. Aki K, Richards PG (1980) *Quantitative Seismology*. Freeman, New York, p 932
4. Almendros J, Chouet B, Dawson P (2001) Spatial extent of a hydrothermal system at Kilauea Volcano, Hawaii, determined from array analyses of shallow long-period seismicity. *J Geophys Res* 106:13581–13597
5. Almendros J, Chouet B, Dawson P, Huber C (2002) Mapping the sources of the seismic wave field at Kilauea Volcano, Hawaii, using data recorded on multiple seismic antennas. *Bull Seismol Soc Am* 92:2333–2351
6. Almendros J, Chouet B, Dawson P, Bond T (2002) Identifying elements of the plumbing system beneath Kilauea Volcano, Hawaii, from the source locations of very-long-period signals. *Geophys J Int* 148:303–312
7. Almendros J, Chouet B, Dawson P (2008) Shallow magma transport pathway under Kilauea Volcano, Hawaii, imaged from waveform inversions of very-long-period seismic data. 2-Source modeling *J Geophys Res*, submitted
8. Anderson DM, McFadden GB, Wheeler AA (1998) Diffuse-interface methods in fluid mechanics. *Annu Rev Fluid Mech* 30:139–165
9. Arciniega-Ceballos A, Chouet BA, Dawson P (1999) Very-long-period signals associated with Vulcanian explosions at Popocatepetl Volcano, Mexico. *Geophys Res Lett* 26:3013–3016
10. Arciniega-Ceballos A, Chouet B, Dawson P (2003) Long-period events and tremor at Popocatepetl Volcano (1994–2000) and their broadband characteristics. *Bull Volcanol* 65:124–135
11. Arciniega-Ceballos A, Chouet B, Dawson P, Asch G (2008) Broadband seismic measurements of degassing activity associated with lava effusion at Popocatepetl Volcano, Mexico. *J Volcanol Geotherm Res* 170:12–23
12. Ascher UM, Ruuth SJ, Wetton BTR (1995) Implicit-explicit methods for time-dependent partial differential equations. *SIAM J Numer Anal* 3:797–823
13. Aster R, Mah S, Kyle P, McIntosh W, Dunbar N, Johnson J, Ruiz M, McNamara S (2003) Very long period oscillations of Mount Erebus Volcano. *J Geophys Res* 108(B11). doi:10.1029/2002JB002101
14. Auger E, D'Auria L, Martini M, Chouet B, Dawson P (2006) Real-time monitoring and massive inversion of source parameters of very long period seismic signals: An application to Stromboli Volcano, Italy. *Geophys Res Lett* 33:L04301. doi:10.1029/2005GL024703
15. Badalassi VE, Cenicerros HD, Banerjee S (2003) Computation of multiphase systems with phase field models. *J Comput Phys* 190:371–397
16. Balazs AC, Verberg R, Pooley CM, Kuksenok O (2005) Modeling the flow of complex fluids through heterogeneous channels. *Soft Matter* 1:44–54
17. Bray AJ (1994) Theory of phase-ordering kinetics. *Adv Phys* 43:357–459
18. Burton M, Allard P, Muré F, La Spina A (2007) Magmatic gas composition reveals the source depth of slug-driven Strombolian explosive activity. *Science* 317(5835):227–230
19. Cahn JW, Hilliard JE (1958) Free energy of a nonuniform system. I. Interfacial free energy. *J Chem Phys* 28:258–267
20. Cahn JW, Hilliard JE (1959) Free energy of a nonuniform system. III, Nucleation in a two-component incompressible fluid. *J Chem Phys* 31:688–689
21. Cervelli PF, Miklius A (2003) The shallow magmatic system of Kilauea Volcano. *US Geol Surv Prof Paper* 1676:149–163
22. Chouet B, Julian BR (1985) Dynamics of an expanding fluid-filled crack. *J Geophys Res* 90:11187–11198
23. Chouet B (1986) Dynamics of a fluid-driven crack in three dimensions by the finite difference method. *J Geophys Res* 91:13967–13992
24. Chouet B (1988) Resonance of a fluid-driven crack: Radiation properties and implications for the source of long-period events and harmonic tremor. *J Geophys Res* 93:4375–4400
25. Chouet B (1992) A seismic source model for the source of long-period events and harmonic tremor. In: Gasparini P, Scarpa R, Aki K (eds) *IAVCEI Proceedings in Volcanology*, vol 3. Springer, New York, pp 133–156
26. Chouet BA, Page RA, Stephens CD, Lahr JC, Power JA (1994) Precursory swarms of long-period events at Redoubt Volcano (1989–1990), Alaska: Their origin and use as a forecasting tool. *J Volcanol Geotherm Res* 62:95–135
27. Chouet B (1996) Long-period volcano seismicity: its source and use in eruption forecasting. *Nature* 380:309–316
28. Chouet B (1996) New methods and future trends in seismological volcano monitoring. In: Scarpa R, Tilling RI (eds) *Monitoring and Mitigation of Volcano Hazards*. Springer, New York, pp 23–97
29. Chouet B, Saccorotti G, Martini M, Dawson P, De Luca G, Milana G, Scarpa R (1997) Source and path effects in the wave fields of tremor and explosions at Stromboli Volcano, Italy. *J Geophys Res* 102:15129–15150
30. Chouet B, Saccorotti G, Dawson P, Martini M, Scarpa R, De Luca G, Milana G, Cattaneo M (1999) Broadband measurements of the sources of explosions at Stromboli Volcano, Italy. *Geophys Res Lett* 26:1937–1940
31. Chouet B (2003) Volcano seismology. *PAGEOPH* 160(3–4):739–788
32. Chouet B, Dawson P, Ohminato T, Martini M, Saccorotti G, Giudicepietro F, De Luca G, Milana G, Scarpa R (2003) Source mechanisms of explosions at Stromboli Volcano, Italy, determined from moment-tensor inversions of very-long-period data. *J Geophys Res* 108(B1):2019. doi:10.1029/2002JB001919
33. Chouet B, Dawson P, Arciniega-Ceballos A (2005) Source mechanism of Vulcanian degassing at Popocatepetl Volcano, Mexico, determined from waveform inversions of very long period signals. *J Geophys Res* 110:B07301. doi:10.1029/2004JB003524

34. Chouet B, Dawson P, Nakano M (2006) Dynamics of bubble growth and pressure recovery in a bubbly rhyolitic melt embedded in an elastic solid. *J Geophys Res* 111:B07310. doi:10.1029/2005JB004174
35. Chouet B, Dawson P, Martini M (2008) Shallow-conduit dynamics at Stromboli Volcano, Italy, imaged from waveform inversion. In: Lane SJ, Gilbert JS (eds) *Fluid Motions in Volcanic Conduits A Source of Seismic and Acoustic Signals*. *Geol Soc Lond Special Pub* 307:57–84
36. Collier L, Neuberg JW, Lensky N, Lyakhovsky V, Navon O (2006) Attenuation in gas-charged magma. *J Volcanol Geotherm Res* 153:21–36
37. Commander KW, Prosperetti A (1989) Linear pressure waves in bubbly liquids: Comparison between theory and experiments. *J Acoust Soc Am* 85:732–746
38. Dawson PB, Dietel C, Chouet BA, Honma K, Ohminato T, Okubo P (1998) A digitally telemetered broadband seismic network at Kilauea Volcano, Hawaii. *US Geol Surv Open-File Rep* 98–108:1–121
39. Dawson P, Whilldin D, Chouet B (2004) Application of near real-time radial semblance to locate the shallow magmatic conduit at Kilauea Volcano, Hawaii. *Geophys Res Lett* 31:L21606. doi:10.1029/2004GL021163
40. Dzurisin D, Vallance JW, Gerlach TM, Moran SC, Malone SD (2005) Mount St. Helens reawakens. *Eos Trans AGU* 86:25,29 doi:10.1029/2005EO030001
41. Evans R (1979) The nature of the liquid-vapour interface and other topics in the statistical mechanics of non-uniform, classical fluids. *Adv Phys* 28:148–200
42. Ferrazini V, Aki K (1987) Slow waves trapped in a fluid-filled infinite crack: Implications for volcanic tremor. *J Geophys Res* 92:9215–9233
43. Fujita E, Ida Y, Oikawa J (1995) Eigen oscillation of a fluid sphere and source mechanism of harmonic volcanic tremor. *J Volcanol Geotherm Res* 69:365–378
44. Fujita E, Ukawa M, Yamamoto E (2004) Subsurface cyclic magma sill expansions in the 2000 Miyakejima Volcano eruption: Possibility of two-phase flow oscillation. *J Geophys Res* 109:B04205. doi:10.1029/2003JB002556
45. Gil Cruz F, Chouet BA (1997) Long-period events, the most characteristic seismicity accompanying the emplacement and extrusion of a lava dome in Galeras Volcano, Colombia, in 1991. *J Volcanol Geotherm Res* 77:121–158
46. Hamada N, Jingu H, Ikumoto K (1976) On the volcanic earthquake with slowly decaying coda wave (in Japanese with English abstract). *Bull Volcanol Soc Jpn* 21:167–183
47. Hasegawa A, Zhao D, Hori S, Yamamoto A, Horiuchi S (1991) Deep structure of the northeastern Japan arc and its relationship to seismic and volcanic activity. *Nature* 352:683–689
48. Heliker C, Mattox T (2003) The first two decades of the Pu'u O'o-Kupaianaha eruption: Chronology and selected bibliography. *US Geol Surv Prof Paper* 1676:1–27
49. Hess K, Dingwell DB (1996) Viscosities of hydrous leucogranitic melts: A non-Arrhenian model. *Am Mineral* 81:1297–1300
50. Hidayat D, Voight B, Langston C, Ratdomopurbo A, Ebeling C (2000) Broadband seismic experiment at Merapi Volcano, Java, Indonesia: Very-long-period pulses embedded in multi-phase earthquakes. *J Volcanol Geotherm Res* 100:215–231
51. Hidayat D, Chouet B, Voight B, Dawson P, Ratdomopurbo A (2002) Source mechanism of very-long-period signals accompanying dome growth activity at Merapi Volcano, Indonesia. *Geophys Res Lett* 29(23):2118. doi:10.1029/2002GL015013
52. Hidayat D, Chouet B, Voight B, Dawson P, Ratdomopurbo A (2003) Correction to Source mechanism of very-long-period signals accompanying dome growth activity at Merapi Volcano, Indonesia. *Geophys Res Lett* 30(10):2118. doi:10.1029/2003GL017211
53. Hill DP, Dawson P, Johnston MJS, Pitt AM, Biasi G, Smith K (2002) Very-long-period volcanic earthquakes beneath Mammoth Mountain, California. *Geophys Res Lett* 29(10):1370. doi:10.1029/2002GL014833
54. Hill DP, Prejean S (2005) Magmatic unrest beneath Mammoth Mountain, California. *J Volcanol Geotherm Res* 146:257–283
55. Hirabayashi J (1999) Formation of volcanic fluid reservoir and volcanic activity. *J Balneol Soc Jpn* 49:99–105
56. Ichihara M, Kameda M (2004) Propagation of acoustic waves in a visco-elastic two-phase system: Influences of the liquid viscosity and the internal diffusion. *J Volcanol Geotherm Res* 137:73–91
57. Iverson RM, Dzurisin D, Gardner CA, Gerlach TM, LaHusen RG, Lisowski M, Major JJ, Malone SD, Messerich JA, Moran SC, Palister JS, Qamar AI, Schilling SP, Vallance JW (2006) Dynamics of seismogenic volcanic extrusion at Mount St. Helens in 2004–05. *Nature* 444:439–443. doi:10.1038/nature05322
58. Jacqmin D (1999) Calculation of two-phase Navier–Stokes Flows using phase-field modeling. *J Comput Phys* 155:96–127
59. James MR, Lane SJ, Chouet B, Gilbert JS (2004) Pressure changes associated with the ascent and bursting of gas slugs in liquid-filled vertical and inclined conduits. *J Volcanol Geotherm Res* 129:61–82
60. James MR, Lane SJ, Chouet BA (2006) Gas slug ascent through changes in conduit diameter: Laboratory insights into a volcano-seismic source process in low-viscosity magmas. *J Geophys Res* 111:B05201. doi:10.1029/2005JB003718
61. Jones RAL (2002) *Soft Condensed Matter*. Oxford University Press, Oxford, p 195
62. Julian BR (1994) Volcanic tremor, Nonlinear excitation by fluid flow. *J Geophys Res* 99:11859–11877
63. Kanamori H, Given JW (1982) Analysis of long-period waves excited by the May 18, 1980, eruption of Mount St. Helens – A terrestrial monopole? *J Geophys Res* 87:5422–5432
64. Kanamori H, Given JW, Lay T (1984) Analysis of seismic body waves excited by the Mount St. Helens eruption of May 18:1980. *J Geophys Res* 89:1856–1866
65. Kaneshima S, Kawakatsu H, Matsubayashi H, Sudo Y, Tsutsui T, Ohminato T, Ito H, Uehira K, Yamasato H, Oikawa J, Takeo M, Iidaka T (1996) Mechanism of phreatic eruptions at Aso Volcano inferred from near-field broadband seismic observations. *Science* 273:642–645
66. Kawakatsu H (1989) Centroid single force inversion of seismic waves generated by landslides. *J Geophys Res* 94:12363–12374
67. Kawakatsu H, Ohminato T, Ito H, Kuwahara Y (1992) Broadband seismic observation at Sakurajima Volcano, Japan. *Geophys Res Lett* 19:1959–1962
68. Kawakatsu H, Ohminato T, Ito H (1994) 10-s-period volcanic tremors observed over a wide area in southwestern Japan. *Geophys Res Lett* 21:1963–1966
69. Kawakatsu H, Kaneshima S, Matsubayashi H, Ohminato T, Sudo Y, Tsutsui T, Uehira K, Yamasato H, Ito H, Legrand D (2000)

- Aso94: Aso seismic observation with broadband instruments. *J Volcanol Geotherm Res* 101:129–154
70. Kedar S, Sturtevant B, Kanamori H (1996) The origin of harmonic tremor at Old Faithful Geyser. *Nature* 379:708–711
  71. Kendon VM, Cates ME, Pagonabarraga I, Desplat J-C, Bladon P (2001) Inertial effects in three-dimensional spinodal decomposition of a symmetric binary fluid mixture: A lattice Boltzmann study. *J Fluid Mech* 440:147–203
  72. Kieffer SW (1977) Sound speed in liquid-gas mixtures: Water-air and water-steam. *J Geophys Res* 82:2895–2904
  73. Kim J, Kang K, Lowengrub J (2004) Conservative multigrid methods for Cahn–Hilliard fluids. *J Comput Phys* 193:511–543
  74. Klein FW, Koyanagi RY, Nakata JS, Tanigawa WR (1987) The seismicity of Kilauea's magma system. In: Decker RW, Wright TL, Stauffer RW (eds) *Volcanism in Hawaii*. US Geol. Surv. Prof. Pap., 1350. US government printing office, Washington, pp 1019–1185
  75. Koyanagi RY, Chouet B, Aki K (1987) Origin of volcanic tremor in Hawaii. Part I: Compilation of seismic data from the Hawaiian Volcano Observatory, 1972 to 1985. In: Decker RW, Wright TL, Stauffer RW (eds) *Volcanism in Hawaii*, US Geol. Surv. Prof. Pap., 1350. US government printing office, Washington, pp 1221–1257
  76. Kumagai H, Chouet BA (1999) The complex frequencies of long-period seismic events as probes of fluid composition beneath volcanoes. *Geophys Int* 138:F7–F12
  77. Kumagai H, Chouet BA (2000) Acoustic properties of a crack containing magmatic or hydrothermal fluids. *J Geophys Res* 105:25493–25512
  78. Kumagai H, Ohnato T, Nakano M, Ooi M, Kubo A, Inoue H, Oikawa J (2001) Very-long-period seismic signals and caldera formation at Miyake Island, Japan. *Science* 293:687–690
  79. Kumagai H, Chouet BA, Nakano M (2002) Temporal evolution of a hydrothermal system in Kusatsu-Shirane Volcano, Japan, inferred from the complex frequencies of long-period events. *J Geophys Res* 107(B10):2236. doi:10.1029/2001JB000653
  80. Kumagai H, Chouet BA, Nakano M (2002) Waveform inversion of oscillatory signatures in long-period events beneath volcanoes. *J Geophys Res* 107(B11):2301. doi:10.1029/2001JB001704
  81. Kumagai H, Miyakawa K, Negishi H, Inoue H, Obara K, Suet-sugu D (2003) Magmatic dyke resonances inferred from very-long-period seismic signals. *Science* 299:2058–2061
  82. Kumagai H, Chouet BA, Dawson PB (2005) Source process of a long-period event at Kilauea Volcano, Hawaii. *Geophys J Int* 161:243–254
  83. Kumagai H (2006) Temporal evolution of a magmatic dike system inferred from the complex frequencies of very long period seismic signals. *J Geophys Res* 111:B06201. doi:10.1029/2005JB003881
  84. Kumagai H, Yepes H, Vaca M, Caceres V, Nagai T, Yokoe K, Imai T, Miyakawa K, Yamashina T, Arrais S, Vasconez F, Pinajota E, Cisneros C, Ramos C, Paredes M, Gomezjurado L, Garcia-Arztizabal A, Molina I, Ramon P, Segovia M, Palacios P, Troncoso L, Alvarado A, Aguilar J, Pozo J, Enriquez W, Mothes P, Hall M, Inoue I, Nakano M, Inoue H (2007) Enhancing volcano-monitoring capabilities in Ecuador. *Eos Trans Am Geophys Union* 88:245–246
  85. Lane SJ, Chouet BA, Phillips JC, Dawson P, Ryan GA, Hurst E (2001) Experimental observations of pressure oscillations and flow regimes in an analogue volcanic system. *J Geophys Res* 106:6461–6476
  86. Legrand D, Kaneshima S, Kawakatsu H (2000) Moment tensor analysis of near-field broadband waveforms observed at Aso Volcano, Japan. *J Volcanol Geotherm Res* 101:155–169
  87. Lensky NG, Lyakhovsky V, Navon O (2002) Expansion dynamics of volatile-saturated liquids and bulk viscosity of bubbly magmas. *J Fluid Mech* 460:39–56
  88. Lensky NG, Navon O, Lyakhovsky V (2004) Bubble growth during decompression of magma: experimental and theoretical investigation. *J Volcanol Geotherm Res* 129:7–22
  89. Lowengrub J, Truskinovsky L (1998) Quasi-incompressible Cahn–Hilliard fluids and topological transitions. *Proc Royal Soc Lond A* 454:2617–2654
  90. Lyakhovsky V, Hurwitz S, Navon O (1996) Bubble growth in rhyolitic melts: experimental and numerical investigation. *Bull Volcanol* 58:19–32
  91. Major JJ, Scott WE, Driedger C, Dzurisin D (2005) Mount St. Helens erupts again – Activity from September 2004 through March 2005. US Geological Survey Fact Sheet FS2005-3036. US government printing office, Washington, p 4
  92. Meier GEA, Grabitz G, Jungowsky WM, Wittczak KJ, Anderson JS (1978) Oscillations of the supersonic flow downstream of an abrupt increase in duct crosssection. *Mitteilungen aus dem Max-Planck Institut für Strömungsforschung und der Aerodynamischen Versuchsanstalt* 65:1–172
  93. Migler KB (2001) String formation in sheared polymer blends: coalescence, breakup, and finite size effects. *Phys Rev Lett* 86(6):1023–1026
  94. Moran SC, Malone SD, Qamar AI, Thelen W, Wright AK, Caplan-Auerbach J (2007) 2004–2005 seismicity associated with the renewed dome-building eruption of Mount St. Helens. In: Sherrod DR, Scott WE, Stauffer PH (eds) *A Volcano rekindled: the first year of renewed eruptions at Mount St. Helens, 2004–2006*. US Geological Survey Prof. Pap. 2007-XXXX
  95. Morrissey MM, Chouet BA (1997) A numerical investigation of choked flow dynamics and its application to the triggering mechanism of long-period events at Redoubt Volcano, Alaska. *J Geophys Res* 102:7965–7983
  96. Nakano M, Kumagai H, Kumazawa M, Yamaoka K, Chouet BA (1998) The excitation and characteristic frequency of the long-period volcanic event: An approach based on an inhomogeneous autoregressive model of a linear dynamic system. *J Geophys Res* 103:10031–10046
  97. Nakano M, Kumagai H, Chouet BA (2003) Source mechanism of long-period events at Kusatsu-Shirane Volcano, Japan, inferred from waveform inversion of the effective excitation functions. *J Volcanol Geotherm Res* 122:149–164
  98. Navon O, Lyakhovsky V (1998) Vesiculation processes in silicic magmas. *Geol Soc Lond Special Pub* 145:27–50
  99. Neuberg J, Luckett R, Ripepe M, Braun T (1994) Highlights from a seismic broadband array on Stromboli Volcano. *Geophys Res Lett* 21:749–752
  100. Neuberg JW, Tuffen H, Collier L, Green D, Powell T, Dingwell D (2006) The trigger mechanism of low-frequency earthquakes on Montserrat. *J Volcanol Geotherm Res* 153:37–50
  101. Nishimura T, Nakamichi H, Tanaka S, Sato M, Kobayashi T, Ueki S, Hamaguchi H, Ohtake M, Sato H (2000) Source process of very long period seismic events associated with the 1998 activity of Iwate Volcano, northeastern Japan. *J Geophys Res* 105:19135–19147

102. Nishimura T, Chouet B (2003) A numerical simulation of magma motion, crustal deformation, and seismic radiation associated with volcanic eruptions. *Geophys. J Int* 153: 699–718
103. Nishimura T, Ueki S, Yamawaki T, Tanaka S, Hasino H, Sato M, Nakamichi H, Hamaguchi H (2003) Broadband seismic signals associated with the 2000 volcanic unrest of Mount Bandai, northeastern Japan. *J Volcanol Geotherm Res* 119:51–59
104. Nishimura T (2004) Pressure recovery in magma due to bubble growth. *Geophys Res Lett* 31:L12613. doi:10.1029/2004GL019810
105. Ochs FA, Lange RA (1999) The density of hydrous magmatic liquids. *Science* 283:1314–1317
106. Ohba T, Hirabayashi J, Nogami K (2000) D/H and  $^{18}\text{O}/^{16}\text{O}$  ratios of water in the crater lake of Kusatsu-Shirane Volcano, Japan. *J Volcanol Geotherm Res* 97:329–346
107. Ohminato T, Chouet BA (1997) A free-surface boundary condition for including 3D topography in the finite-difference method. *Bull Seismol Soc Am* 87:494–515
108. Ohminato T, Ereditato D (1997) Broadband seismic observations at Satsuma-Iwojima Volcano, Japan. *Geophys Res Lett* 24:2845–2848
109. Ohminato T, Chouet BA, Dawson PB, Kedar S (1998) Waveform inversion of very-long-period impulsive signals associated with magmatic injection beneath Kilauea volcano, Hawaii. *J Geophys Res* 103:23839–23862
110. Ohminato T (2006) Characteristics and source modeling of broadband seismic signals associated with the hydrothermal system at Satsuma-Iwojima volcano, Japan. *J Volcanol Geotherm Res* 158:467–490
111. Ohminato T, Takeo M, Kumagai H, Yamashina T, Oikawa J, Koyama E, Tsuji H, Urabe T (2006) Vulcanian eruptions with dominant single force components observed during the Asama 2004 volcanic activity in Japan. *Earth Planets Space* 58:583–593
112. Orlandini E, Swift MR, Yeomans JM (1995) A Lattice Boltzmann Model of binary fluid mixtures. *Europhys Lett* 32:463
113. Pooley CM, Kuksenok O, Balasz AC (2005) Convection-driven pattern formation in phase-separating binary fluids. *Phys Rev E* 71:030501(R)
114. Power JA, Stihler SD, White RA, Moran SC (2004) Observations of deep long-period (DLP) seismic events beneath Aleutian Arc Volcanoes, 1989–2002. *J Volcanol Geotherm Res* 138:243–266
115. Proussevitch AA, Sahagian DL, Anderson AT (1993) Dynamics of diffusive bubble growth in magmas: Isothermal case. *J Geophys Res* 98:22283–22307
116. Proussevitch AA, Sahagian DL (1996) Dynamics of coupled diffusive and decompressive bubble growth in magmatic systems. *J Geophys Res* 101:17447–17455
117. Proussevitch AA, Sahagian DL (1998) Dynamics and energetics of bubble growth in magmas: Analytical formulation and numerical modeling. *J Geophys Res* 103:18223–18251
118. Ripepe M, Poggi P, Braun T, Gordeev E (1996) Infrasonic waves and volcanic tremor at Stromboli. *Geophys Res Lett* 23: 181–184
119. Rowe CA, Aster RC, Kyle PR, Schlue JW, Dibble RR (1998) Broadband recording of Strombolian explosions and associated very-long-period seismic signals on Mount Erebus Volcano, Ross Island, Antarctica. *Geophys Res Lett* 25:2297–2300
120. Saccorotti G, Chouet B, Dawson P (2001) Wavefield properties of a shallow long-period event and tremor at Kilauea Volcano, Hawaii. *J Volcanol Geotherm Res* 109:163–189
121. Scriven LE (1959) On the dynamics of phase growth. *Chem Eng Sci* 10:1–13
122. Shaw HR, Chouet B (1989) Singularity spectrum of intermittent seismic tremor at Kilauea Volcano, Hawaii. *Geophys Res Lett* 16:195–198
123. Shaw HR, Chouet B (1991) Fractal hierarchies of magma transport in Hawaii and critical self-organization of tremor. *J Geophys Res* 96:10191–10207
124. Shimomura Y, Nishimura T, Sato H (2006) Bubble growth processes in magma surrounded by an elastic medium. *J Volcanol Geotherm Res* 155:307–322
125. Sparks R (1978) The dynamics of bubble formation and growth in magmas: A review and analysis. *J Volcanol Geotherm Res* 3:1–38
126. Stephens CD, Chouet BA (2001) Evolution of the December 14, 1989 precursory long-period event swarm at Redoubt Volcano, Alaska. *J Volcanol Geotherm Res* 109:133–148
127. Swift MR, Osborn WR, Yeomans JM (1995) Lattice Boltzmann simulation of nonideal fluids. *Phys Rev Lett* 75(5):0–833
128. Swift MR, Orlandini E, Osborn WR, Yeomans JM (1996) Lattice Boltzmann simulations of liquid-gas and binary fluid systems. *Phys Rev E* 54(5):5041–5052
129. Takei Y, Kumazawa M (1994) Why have the single force and torque been excluded from seismic source models? *Geophys J Int* 118:20–30
130. Tannehill JC, Anderson DA, Pletcher RH (1997) *Computational Fluid Mechanics and Heat Transfer*, 2nd edn. Taylor and Francis, Philadelphia, p 792
131. Toramaru A (1989) Vesiculation process and bubble size distributions in ascending magmas with constant velocities. *J Geophys Res* 94:17523–17542
132. Toramaru A (1995) Numerical study of nucleation and growth of bubbles in viscous magmas. *J Geophys Res* 100:1913–1931
133. Tuffen H, Dingwell DB, Pinkerton H (2003) Repeated fracture and healing of silicic magma generates flow banding and earthquakes? *Geology* 31:1089–1092
134. Uhira K, Yamasato H, Takeo M (1994) Source mechanism of seismic waves excited by pyroclastic flows observed at Unzen Volcano, Japan. *J Geophys Res* 99:17757–17773
135. van der Waals JD (1979) The thermodynamic theory of capillarity under the hypothesis of a continuous variation of density. *Verhandel Konink. Acad. Wet. Amsterdam, (Sect. 1)*, 1, pp 1–56. Translation by Rowlingson JS. *J Statist Phys* 20: 197–244
136. Vergnolle S, Brandeis G, Mareschal JC (1996) Strombolian explosions, 2. Eruption dynamics determined from acoustic measurements. *J Geophys Res* 101:20449–20466
137. Waite GP, Chouet BA, Dawson PB (2008) Eruption dynamics at Mount St. Helens imaged from broadband seismic waveforms: Interaction of the shallow magmatic and hydrothermal systems. *J Geophys Res* 113, B02305. doi:10.1029/2007JB005259
138. Wallis GB (1969) *One-dimensional Two-phase Flow*. MacGraw-Hill, New York, pp 408
139. White RA (1996) Precursory deep long-period earthquakes at Mount Pinatubo: Spatio-temporal link to a basalt trigger.

- In: Newhall CG, Punongbayan RS (eds) *Fire and Mud: Eruptions and Lahars of Mount Pinatubo, Philippines*. University of Washington Press, Seattle, pp 307–326
140. Wilson L, Head JW (1981) Ascent and eruption of basaltic magma on the earth and moon. *J Geophys Res* 86:2971–3001
141. Xu A, Gonnella G, Lamura A (2003) Phase-separating binary fluids under oscillatory shear. *Phys Rev E* 67:056105
142. Xu A, Gonnella G, Lamura A (2004) Phase separation of incompressible binary fluids with lattice Boltzmann methods. *Physica A* 331:10–22
143. Yamamoto M, Kawakatsu H, Kaneshima S, Mori T, Tsutsui T, Sudo Y, Morita Y (1999) Detection of a crack-like conduit beneath the active crater at Aso Volcano, Japan. *Geophys Res Lett* 26:3677–3680
144. Yamamoto M, Kawakatsu H, Yomogida K, Koyama J (2002) Long-period (12 sec) volcanic tremor observed at Usu 2000 eruption: seismological detection of a deep magma plumbing system. *Geophys Res Lett* 29(9):1329. doi:10.1029/2001GL013996
145. Yoon SW, Crum LA, Prosperetti A, Lu NQ (1991) An investigation of the collective oscillations of a bubble cloud. *J Acoust Soc Am* 89:700–706
146. Yue P, Feng JJ, Liu C, Shen J (2004) A diffuse-interface method for simulating two-phase flows of complex fluids. *J Fluid Mech* 515:293–317
147. Zhang Y, Xu Z, Liu Y (2003) Viscosity of hydrous rhyolitic melts inferred from kinetic experiments: A new viscosity model. *Am Mineral* 88:1741–1752

# Wedge Mechanics: Relation with Subduction Zone Earthquakes and Tsunamis

KELIN WANG<sup>1,2</sup>, YAN HU<sup>2</sup>, JIANGHENG HE<sup>1</sup>

<sup>1</sup> Pacific Geoscience Centre, Geological Survey of Canada, Sidney, Canada

<sup>2</sup> School of Earth and Ocean Sciences, University of Victoria, Victoria, Canada

## Article Outline

Glossary

Definition of the Subject

Introduction

Stable and Critical Coulomb Wedges

Dynamic Coulomb Wedge

Stress Drop and Increase in a Subduction Earthquake

Tsunamigenic Coseismic Seafloor Deformation

Future Directions

Acknowledgments

Bibliography

## Glossary

**Subduction zone earthquake cycle** Megathrust fault, the interface between the two converging lithospheric plates at a subduction zone, moves in a stick-slip fashion. In the “stick” phase, the fault is locked or slips very slowly, allowing elastic strain energy to be accumulated in both plates around the fault. Every few decades or centuries, the fault breaks as high-rate “slip” to release the strain energy, causing a large or great earthquake, usually accompanied with a tsunami. An interseismic period and the ensuing earthquake together is called a subduction zone earthquake cycle. The word cycle by no means implies periodicity. Neighboring segments of the same subduction zone may exhibit different temporal patterns of earthquake cycles.

**Accretionary wedge (prism)** At some subduction zones, as one plate subducts beneath the other, some sediment is scraped off the incoming plate and accreted to the leading edge of the upper plate. Because of its wedge shape, the accreted sedimentary body is called the accretionary wedge (or accretionary prism). If all the sediment on the incoming plate is subducted, there is still a sedimentary wedge in the frontal part of the upper plate, but it is usually very small and consists of sediments derived from the upper plate by surface erosion.

**Coulomb plasticity** Coulomb plasticity is a macroscopic, continuum description of the most common type of permanent deformation of Earth materials such as sand, soil, and rock at relatively low temperature and pressure and is widely used in civil engineering and Earth science. In detail, the deformation mechanism is actually brittle shear failure, with or without emitting elastic wave energy. The macroscopic yield criterion is the Coulomb’s law, in which shear strength increases linearly with confining pressure. If the strength does not change with permanent deformation, the material is said to be perfectly plastic. Note that in Earth science the word plasticity is also used to indicate thermally activated creep, but it is very different from the meaning used herein.

**Velocity-weakening and strengthening** These are macroscopic descriptions of dynamic frictional behavior of contact surfaces. Velocity-weakening, featuring a net decrease in frictional strength with increasing slip rate, is the necessary condition for a fault to produce earthquakes. It differs from slip-weakening in that a velocity-weakened fault will regain its strength when the slip slows down or stops. Velocity-strengthening is the opposite of velocity-weakening and is the necessary condition for a fault to resist earthquake rupture. Detailed physical processes on the contact surfaces or within the fault zones controlling their frictional behavior are still being investigated.

## Definition of the Subject

Mechanics of wedge-shaped geological bodies such as accretionary prisms at subduction zones and fold-and-thrust belts at collision zones is of great scientific interest, mainly because it enables us to use the observed morphology and deformation of the wedge-shaped bodies to constrain properties of the thrust faults underlying them. Davis et al. [12] drew analogy between these wedge-shaped bodies and the sand wedge in front of a moving bulldozer and established a mathematical model. Their work triggered wide applications of wedge mechanics to geology. The fundamental process described in wedge mechanics is how gravitational force, in the presence of a sloping surface, is balanced by basal stress and internal stress. The internal state of stress depends on the rheology of the wedge. The most commonly assumed wedge rheology for geological problems is perfect Coulomb plasticity [12], and the model based on this rheology is referred to as the Coulomb wedge model.

The Coulomb wedge model was designed to explain geological processes of timescale of hundreds of thousands

of years. The state of stress is understood to be an average over time, and the wedge is assumed to be in a critical state, that is, uniformly at the Coulomb yield stress. In the application of the model to the sedimentary wedges at subduction zones, attention is now being paid to the temporal changes in stress and pore fluid pressure associated with great subduction earthquakes which have recurrence intervals of decades to centuries. To account for the short-term stress changes, Wang and Hu [50] expanded the Coulomb wedge model by introducing the elastic – perfectly Coulomb plastic rheology. The expanded Coulomb wedge model links long-term geology with coseismic processes and provides a new perspective for the study of subduction zone earthquakes, tsunami generation, frontal wedge structure, and forearc hydrogeology.

**Introduction**

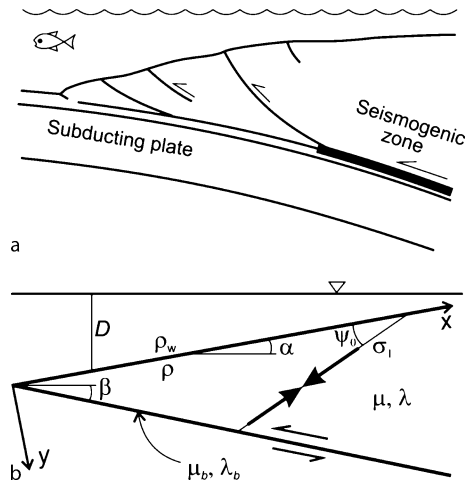
Sloping surfaces are commonly dealt with in engineering problems such as dam design and landslide hazard mitigation. In the presence of a sloping surface, materials under gravitational force tend to “flow” downhill and thus generate tensile stress, but whether collapse actually occurs depends on the strength of the material. Materials used in construction can easily sustain stresses caused by the presence of a vertical surface, but a material of no shear strength such as stationary water cannot support any surface slope. A geological wedge, such as the accretionary prism at a subduction zone (see Fig. 1a), overlies a dipping fault, and hence its internal stress is controlled by the strength of the fault as well. If the basal fault is a thrust fault and is strong relative to the strength of the wedge material, the wedge can undergo compressive failure. Wedge mechanics thus consists of two aspects: the frictional behavior of the basal fault and the deformation of the wedge itself. Given a wedge with density  $\rho$ , surface slope angle  $\alpha$ , and basal dip  $\beta$  (see Fig. 1b), Elliot [16], Chapple [8], and later workers all considered the following equations of force balance (exact form varies between publications depending on the assumed coordinate systems)

$$\frac{\partial \sigma_x}{\partial x} + \frac{\partial \tau_{xy}}{\partial y} - \rho g \sin \alpha = 0, \tag{1a}$$

$$\frac{\partial \tau_{xy}}{\partial x} + \frac{\partial \sigma_y}{\partial y} + \rho g \cos \alpha = 0, \tag{1b}$$

where  $g$  is gravitational acceleration,  $\sigma_x$  and  $\sigma_y$  are normal stresses, and  $\tau_{xy}$  is shear stress.

It was Davis et al. [12] who introduced Coulomb plasticity into wedge mechanics. Coulomb plasticity was initially proposed by French engineer Coulomb in 1773 to describe the mechanical strength of soil and sand. The



Wedge Mechanics: Relation with Subduction Zone Earthquakes and Tsunamis, Figure 1

**a** Schematic illustration of a subduction zone accretionary wedge. **b** Coordinate system used in this article ( $x, y$ ). The surface of the wedge and the subduction fault are simplified to be planar, with slope angle  $\alpha$  and dip  $\beta$ , respectively, and the  $x$  axis is aligned with the upper surface.  $\psi_0$  is the angle between the maximum compressive stress  $\sigma_1$  and the upper surface.  $D$  is water depth.  $\rho$  and  $\rho_w$  are densities of the wedge material and overlying water, respectively.  $\mu, \lambda, \mu_b$  and  $\lambda_b$  are as defined in Eqs. (2) and (3)

Coulomb failure criterion in its simplest form is

$$\tau = S - \mu(\sigma_n + P) = S - \mu \bar{\sigma}_n, \tag{2}$$

where  $\tau$  is shear strength,  $\sigma_n$  is normal stress,  $S$  is cohesion, and  $\mu = \tan \varphi$  is the coefficient of internal friction with  $\varphi$  called the internal friction angle.  $P$  is the pressure of fluids present in the pore space between solid grains and in various small fractures and is loosely referred to as the pore fluid pressure. Note that the effective stress  $\bar{\sigma}_n = \sigma_n + P$  is normal stress with  $P$  subtracted. The plus sign is due to the custom in mechanics (except rock mechanics) that compressive stress is defined to be negative, but pressure, although also compressive, is defined to be positive. Here the plane on which the stress is evaluated is oriented in any arbitrary direction, but failure will start on the set of planes that meets the above criterion. This is a generalization of the Coulomb friction criterion for a fixed fault plane

$$\tau_b = S_b - \mu_b(\sigma_{n_b} + P_b) = S_b - \mu_b \bar{\sigma}_{n_b}, \tag{3}$$

where  $S_b$  is the cohesion of the fault,  $\mu_b = \tan \varphi_b$  is its friction coefficient, and  $P_b$  is fluid pressure along the fault.  $S_b$  is usually negligibly small and almost always taken to be

zero, and  $\mu_b$  is normally significantly lower than  $\mu$ . A well developed fault such as a plate boundary fault is a zone of finite thickness filled with gouge material, so that the “friction” described by (3) or other friction laws is actually the shear deformation of the gouge in the fault zone, and  $P_b$  is actually pore fluid pressure of the gouge. Fault gouge is often made very weak by the presence of hydrous minerals [5,42], such that the collective strength of the fault zone material is much less than the strength of the rocks on both sides. Another process that may weaken the fault is that the local hydrogeological regime may dynamically maintain  $P_b$  in the fault zone to stay higher than  $P$  on both sides [14]. For both Coulomb plasticity and Coulomb friction, strength increases with depth because of the increasing pressure thus normal stress.

It has long been recognized that Coulomb plasticity, featuring strong depth dependence, applies to the shallow part of Earth’s lithosphere. The most common example is the use of Byerlee’s law of rock friction [6] to describe brittle strength in “Christmas-tree”-like vertical strength-depth profiles of the lithosphere. The Byerlee’s law is an empirical Coulomb friction law. By assuming that faults, i. e., potential failure planes, are oriented in all directions, we regard the brittle part of the lithosphere as being Coulomb plastic. This example also illustrates how a system of numerous discrete structures can be regarded as a continuum at a much larger scale. Similarly, although a geological wedge actually consists of numerous blocks divided by fractures, Coulomb plasticity can be used to describe its overall rheology. However, specific values of friction parameters for submarine wedges may be quite different from those in the Byerlee’s law.

The Coulomb wedge model explains how the geometry (taper) of the wedge is controlled by the interplay between the gravitational force, the strength of the wedge material, and the strength of the basal fault. The wedge strength and fault strength are both strongly influenced by pore fluid pressure, and the most popular application of the model is to estimate pore fluid pressure from observed wedge geometry. Since the work of Davis et al. [12], more rigorous analytical solutions have been derived [9,10,18,50,55], and some extensions have been proposed, e. g., [2,3,17,54]. Lithospheric scale numerical models are often used to study the evolution of geological wedges including such effects as erosion and sedimentation in collision or subduction zone settings, e. g., [20,53]. Utilizing the bulldozer – sand wedge analogy, important physical insights have been obtained from sandbox experiments [7,11,29,30,31,35,41,52]. For a list of applications of the Coulomb wedge model to subduction-zone accretionary prisms, see [50].

### Stable and Critical Coulomb Wedges

Depending on the state of stress, a Coulomb wedge can be at a critical state, that is, everywhere at Coulomb failure, or a stable (also referred to as supercritical) state, that is, everywhere not at failure (see Fig. 2). The taper of a critical wedge will not change if the stress does not change; note that a critical wedge of stable geometry should not be confused with a stable wedge. Stress solutions have been derived for both critical and stable states. Here we only summarize the simplest, exact solution for a cohesionless wedge ( $S = 0$ ) derived in the coordinate system shown in Fig. 1b, because of the convenience of its application as compared to other solutions. The wedge is assumed to be elastic – perfectly Coulomb plastic. If it is at the critical state, it obeys (2); if it is at the stable state, it obeys the Hooke’s law of elasticity. The basal thrust fault obeys (3). The wedge is assumed to be under water of density  $\rho_w$  and depth  $D$  (a function of  $x$ ; see Fig. 1b). By defining a Hubbert–Rubey fluid pressure ratio within the wedge [9]

$$\lambda = \frac{P - \rho_w g D}{-\sigma_y - \rho_w g D}, \tag{4}$$

and a similar parameter along the basal fault [51]

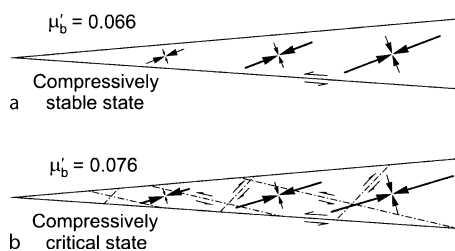
$$\lambda_b = \frac{P_b - \rho_w g D + \lambda H}{-\sigma_y - \rho_w g D + H}, \tag{5a}$$

where

$$H = \frac{\sigma_y - \sigma_n}{1 - \lambda}, \tag{5b}$$

and assuming  $S_b = 0$ , we can rewrite (3) into

$$\tau_b = -\mu_b \bar{\sigma}_{n-b} = -\mu'_b \bar{\sigma}_n, \tag{6a}$$



Wedge Mechanics: Relation with Subduction Zone Earthquakes and Tsunamis, Figure 2

An example to show how stresses in an elastic – perfectly Coulomb-plastic wedge, with  $\alpha = 5^\circ$ ,  $\beta = 4^\circ$ ,  $\mu = 0.6$ , and  $\lambda = 0.86$ , are affected by basal friction  $\mu'_b = \mu_b(1 - \lambda_b)$ . Converging arrows represent principal stresses, with the larger one being  $\sigma_1$ . **a** Compressively stable state. **b** Compressively critical state. *Dot-dashed* lines are plastic slip lines (potential failure planes)



where

$$\mu_b'' = \frac{1 - \lambda_b}{1 - \lambda} \mu_b = \frac{\mu_b'}{1 - \lambda}. \tag{6b}$$

The strength of the basal fault is represented by  $\mu_b' = \mu_b(1 - \lambda_b)$  which always appears as a single parameter and is commonly referred to as the effective friction coefficient. The hydrological process in the fault zone may differ from that in the wedge and thus cause a sharp gradient in fluid pressure across the wedge base. By allowing  $\lambda_b$  to be different from  $\lambda$ , we use a pressure discontinuity to approximate the sharp gradient. Because stress is continuous across the basal fault, the discontinuity in pore fluid pressure leads to a discontinuity in effective stress. The second equality of Eq. (6a) shows the relation between the effective stress along the fault ( $\bar{\sigma}_{n,b}$ ) and the effective stress just above the fault ( $\bar{\sigma}_n$ ). The establishment of (6) allows the following exact stress solution to be derived. In this expression, all stress components are normalized by  $\rho gy$  (e.g.,  $\bar{\sigma}'_x = \bar{\sigma}_x/\rho gy$ ).

$$\bar{\sigma}'_x = -m(1 - \lambda) \cos \alpha, \tag{7a}$$

$$\bar{\sigma}'_y = -(1 - \lambda) \cos \alpha, \tag{7b}$$

$$\tau'_{xy} = (1 - \rho') \sin \alpha, \tag{7c}$$

where  $\rho' = \rho_w/\rho$ , and the effective stress ratio  $m = \bar{\sigma}_x/\bar{\sigma}_y$  depends on whether the wedge is stable or critical. If the wedge is in a stable state (elastic) [50],

$$m = 1 + \frac{2(\tan \alpha' + \mu_b'')}{\sin 2\theta(1 - \mu_b'' \tan \theta)} - \frac{2 \tan \alpha'}{\tan \theta}, \tag{8}$$

where  $\theta = \alpha + \beta$ , and,

$$\tan \alpha' = \frac{1 - \rho'}{1 - \lambda} \tan \alpha. \tag{9}$$

The angle  $\psi_0$  between the most compressive stress  $\sigma_1$  and the upper surface is uniform (see Fig. 1b). A more general solution for a purely elastic wedge can be found in [23]. If the wedge is in a critical state (perfectly plastic) [9]

$$m = m^c = 1 + \frac{2 \tan \alpha'}{\tan 2\psi_0^c}, \tag{10}$$

where  $\psi_0^c$  is the value of  $\psi_0$  in the critical state and is given by the following relation

$$\frac{\sin \varphi \sin 2\psi_0^c}{1 - \sin \varphi \cos 2\psi_0^c} = \tan \alpha'. \tag{11}$$

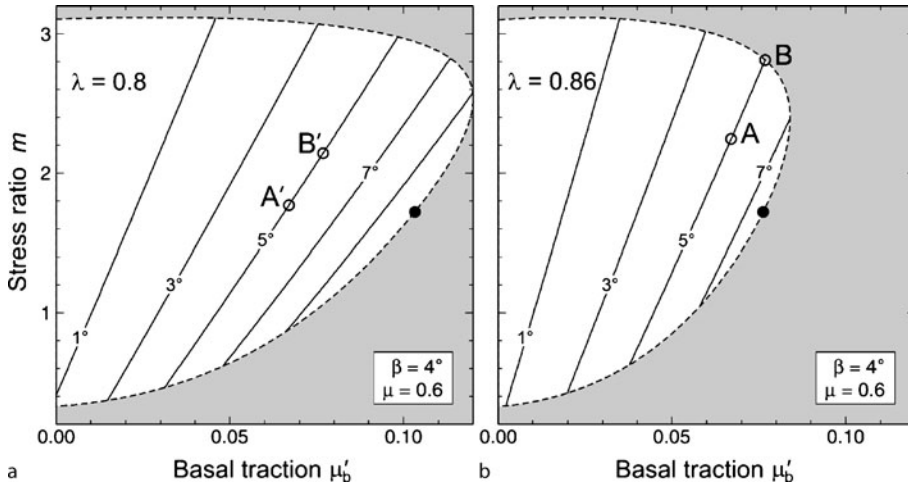
In the above expressions, superscript c indicates critical state. If the wedge has a cohesion that is proportional to depth, (10) and (11) will be modified only slightly [50,55].

A wedge of fixed geometry has two  $m^c$  values. The lower one defines the extensionally critical state, in which the wedge is on the verge of gravitational collapse. This occurs if friction along the basal fault is very low relative to the strength of the wedge material. The higher one defines the compressively critical state, in which the wedge is everywhere on the verge of thrust failure. If  $m$  lies between these two critical values, the wedge is in a stable state and only experiences elastic deformation. A change of basal friction  $\mu_b'$  will cause a change in  $m$  and thus may potentially cause the wedge to switch between the stable and critical states (see Fig. 3). An example of a wedge being in a stable or compressively critical state as controlled by basal friction is shown in Fig. 2. The plastic slip lines (potential failure planes) in the critical wedge (see Fig. 2b) are reminiscent of the out-of-sequence faults in a real accretionary prism (see Fig. 1a).

If we fix the values of all material properties, the critical-wedge solution ( $m = m^c$ ) defines a relation between  $\alpha$  and  $\beta$  representing all possible geometries of a critically tapered wedge (dashed line in Fig. 4). This is a very commonly used diagram, in which the lower branch of the  $\alpha - \beta$  curve represents compressively critical states, and the upper branch represents extensionally critical states. Combinations of  $\alpha$  and  $\beta$  outside of the stability region comprise unstable geometries and cannot exist in steady state. If sedimentary wedges of subduction zones are compressively critical, their observed  $\alpha - \beta$  pairs should line up with the lower branch. However, it has been shown that most of them fall in the stable or even extensionally unstable region of this type of diagram [30,44]. To fit observations, we need to move the lower branch upward by a significant amount. In order to do this, we need to assume either a weaker wedge or a stronger basal fault, or both. This is more simply illustrated by Fig. 3. Given wedge geometry and internal friction, if the state of the wedge is to be changed from stable to compressively critical, we need to have a higher friction ( $\mu_b'$ ) along the basal fault (i.e., greater stress coupling) and/or higher pore fluid pressure within the wedge (i.e., greater  $\lambda$  weakening the wedge material by reducing effective pressure). Wang and Hu [50] proposed that higher  $\mu_b'$  and  $\lambda$  can occur at the time of a great earthquake and introduced the concept of dynamic Coulomb wedge.

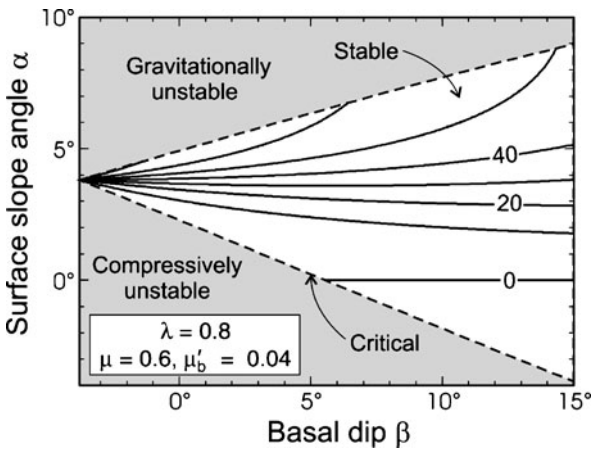
### Dynamic Coulomb Wedge

The concept of dynamic Coulomb wedge is based on the widely recognized frictional behavior of subduction faults. Ignoring the presence of along-strike variations of frictional properties, we can summarize the frictional



Wedge Mechanics: Relation with Subduction Zone Earthquakes and Tsunamis, Figure 3

Effective stress ratio  $m$  (Eq. (7)) as a function of basal friction  $\mu'_b = \mu_b(1 - \lambda_b)$  for wedges of the same basal dip ( $\beta = 4^\circ$ ) but different surface slope angles ( $\alpha$ ) as labelled on the curves (solid lines). a and b are for two different pore fluid pressure ratio values within the wedge. Each curve is terminated at the extensionally critical state at a lower  $\mu'_b$  and the compressively critical state at a higher  $\mu'_b$ . The end points (connected by a dashed line) outline the stable region (white). No solution exists outside this region. The solid circle marks the state in which the surface slope is at the angle of repose. It divides the line of critical states (dashed line) into the compressive part (above) and extensional part (below). Open circles in b labelled A and B mark the states shown in Fig. 2a and b, respectively. State A' in (a) is for the same wedge with the same basal friction as state A in (b) except for a lower pore fluid pressure ratio, and state B' in (a) corresponds to state B in (b) in the same fashion. Comparison of B with B' shows how an increase in pore fluid pressure weakens the wedge



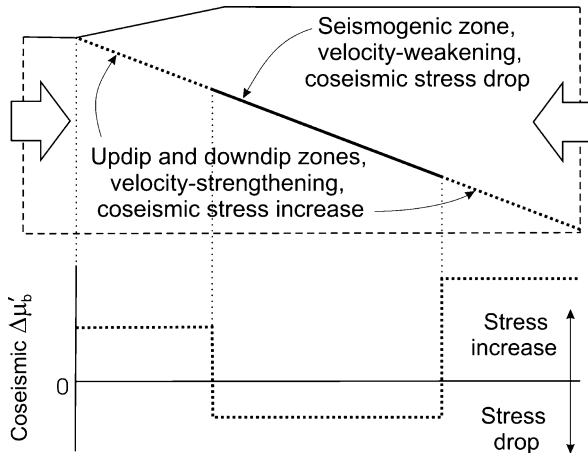
Wedge Mechanics: Relation with Subduction Zone Earthquakes and Tsunamis, Figure 4

Possible wedge geometry ( $\alpha$  and  $\beta$ ) given material strength and basal friction. The dashed line indicates critical wedge geometry, with the upper and lower branches representing extensionally and compressively critical states, respectively. Contours of  $\psi_0$  (in degrees) are shown in the stable region (white)

behavior in a simplified cross-section view (see Fig. 5). The seismogenic zone exhibits velocity-weakening behavior: It weakens in response to high-rate slip, resulting in

slip instability, that is, earthquakes. The segments updip and downdip of the seismogenic zone exhibits velocity-strengthening: They strengthen at the time of the earthquake to develop higher stress to resist rupture, but they may slip aseismically after the earthquake to relieve the high stress attained during the earthquake. We assume that the actively deforming sedimentary wedge overlies the updip segment (also see Fig. 1a).

Microscopic mechanisms for the velocity-weakening behavior of the seismogenic zone and the velocity-strengthening behavior of the aseismic zones are subjects of intense research. The aseismic behavior of the deeper part of any fault is intuitively easy to comprehend; higher temperature at greater depths increasingly enhances viscous deformation and inhibits brittle faulting. There are different physical explanations for the velocity-weakening behavior of the seismogenic zone as summarized in [37]. The mechanism responsible for the velocity-strengthening behavior of the updip segment is yet to be identified, although it is widely accepted that the presence of clay minerals has something to do with it [24,33,34,45,46]. Laboratory experiments indicate that dilatancy of granular fault zone material during fast slip can lead to velocity-strengthening [32]. We think part of the reason for the velocity-strengthening behavior of the updip segment may be its



Wedge Mechanics: Relation with Subduction Zone Earthquakes and Tsunamis, Figure 5

Schematic illustration of the subduction zone model considered in this work. Large arrows represent interseismic strain accumulation. An earthquake is represented by a sudden decrease in the effective friction coefficient  $\mu'_b$  of the seismogenic zone by  $\Delta\mu'_b$ . Coseismic strengthening of the updip and downdip zones is represented by a sudden increase in their  $\mu'_b$  values

inability to localize into a very thin slip zone. Seismic rupture occurs along very thin slip zones of a few millimeters thickness that are parts of a thicker fault zone [40]. Fast slip of the updip segment, if triggered by the rupture of the deeper seismogenic zone, may tend to drag along fault zone materials over a more distributed band and thus meet greater resistance. This view is different from the velocity-strengthening process described by laboratory-derived rate- and state-dependent friction laws [15,39] in which dynamic changes in the thickness of the slip zone plays no role.

Regardless of the microscopic mechanisms, the downdip variation of the frictional behavior is expected to bring direct consequences to wedge deformation. In an earthquake, thrust motion of the seismogenic zone causes the frontal wedge to be pushed from behind, and velocity-strengthening of the updip megathrust segment gives rise to higher stress at its base. If the wedge is originally in a stable state, the coseismic strengthening of the basal fault may increase the  $m$  value in (7a) from subcritical to critical. After the earthquake, when the seismogenic zone has returned to a locked state, the stress along the updip segment will relax. The decrease in  $\mu'_b$  leads to a smaller  $m$ , and therefore the frontal wedge returns to a stable state.

If we ignore the change in fluid pressure, the above described process can be seen as the stress ratio  $m$  moving up-and-down along one of the solid lines in Fig. 3 in re-

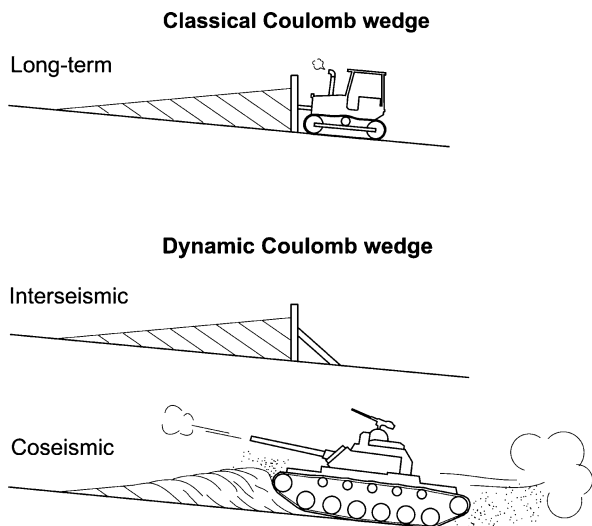
sponse to changes in  $\mu'_b$  in earthquake cycles. During a big earthquake, it will hit the upper end of the line ( $m^c$ ). The state of stress for a stable wedge before the earthquake ( $m < m^c$ ) and that for a compressively critical wedge during the earthquake ( $m = m^c$ ) are illustrated by examples in Fig. 2a and b, respectively. Fluid pressure variation should not be ignored, however. For example, the same basal friction as shown in Fig. 2b (also see point B in Fig. 3b) will not drive the wedge to failure if the pore fluid pressure is lower, as shown by point B' in Fig. 3a.

During an earthquake, the sudden compression of the frontal wedge will cause its internal pore fluid pressure to increase, coseismically weakening the wedge material. By comparing Fig. 3a and b, we can see that if the pore fluid pressure ratio in the wedge is higher, the increase in basal friction required to push the wedge into a critical state can be smaller. We may envisage the following scenario. The pore fluid pressure in a frontal wedge may decrease to some degree over the interseismic period due to fluid drainage through fractures and stress relaxation, and the mechanical state of the wedge before an earthquake can be represented by point A' in Fig. 3a as opposed to point A in Fig. 3b. An earthquake not only causes the basal friction to increase but also the pore fluid pressure within the wedge to rise, such that the wedge enters a critical state represented by point B in Fig. 3b. Therefore, the coseismic strengthening of the basal fault and coseismic weakening of the wedge both work toward bringing the wedge to failure.

All previous applications of the Coulomb wedge model to subduction zones assume  $m = m^c$ . The dynamic Coulomb wedge model of [50] explains the meaning of this long-term  $m^c$ : At least as an end-member scenario, it is the value of  $m$  briefly achieved in numerous large earthquakes. The “average” basal stress that determines the wedge geometry in long-term Coulomb wedge models is actually the peak stress achieved at the time of large earthquakes. Thus, the common illustration of the peaceful scene of a bulldozer pushing a sand wedge in classical Coulomb wedge papers (see Fig. 6a) should be modified to reflect the unpleasant reality of the world (see Fig. 6b).

### Stress Drop and Increase in a Subduction Earthquake

It is important to know the possible amount of stress increase in the frontal wedge for a given earthquake. The increase cannot be arbitrarily large; it is limited by the level of the “push” on the wedge from behind during an earthquake. For this purpose, a numerical model of a larger scale embracing the essential components of the subduc-



Wedge Mechanics: Relation with Subduction Zone Earthquakes and Tsunamis, Figure 6

Cartoon illustrating the difference between the classical Coulomb wedge model and the dynamic Coulomb wedge model for subduction zone accretionary prisms. In the classical wedge model,  $m = m^c$ . In the dynamic wedge model,  $m < m^c$  in the interseismic period but  $m = m^c$  at the time of a large earthquake. See Eqs. (7) and (10) for definition of  $m$  and  $m^c$

tion fault as shown in Fig. 5 must be considered, because the stress interaction between the frontal wedge and the material overlying the seismogenic zone cannot be handled by the analytical Coulomb wedge solutions. For an illustration, we consider the following model geometry, representative of most subduction zones. The subduction fault has a constant dip  $\beta = 4^\circ$ . The frontal 50 km of the upper plate has a surface slope  $\alpha = 5^\circ$ , representing the sedimentary wedge, and the rest of the upper plate has a flat surface. We wish to focus on the process of stress transfer from the seismogenic zone to the updip segment during an earthquake, and a static, uniform, and purely elastic model suffices. For simplicity, the effect of pore fluid pressure change on deformation is neglected.

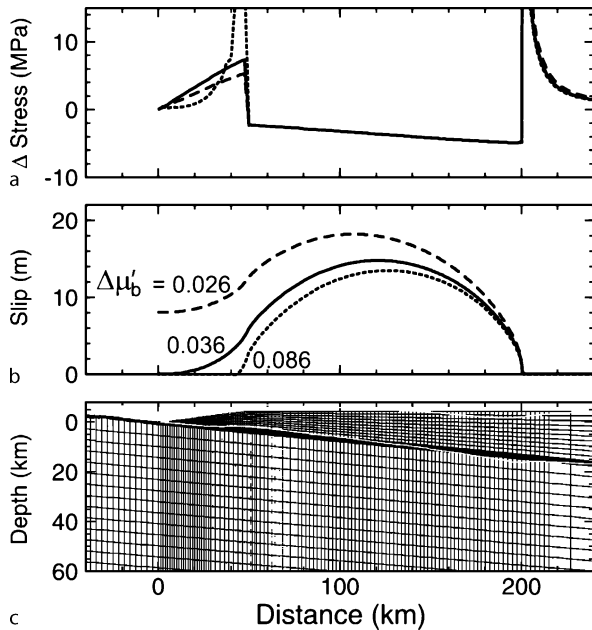
We use a 2D plane-strain finite element model and simulate Coulomb friction along the fault using the method of Lagrange-multiplier Domain Decomposition [48]. The model boundaries are set sufficiently far away so that the model resembles a half-space. For numerical stability, we invoke gravity (assuming a rock density of  $2800 \text{ kg/m}^3$  only when determining yield stress along the fault but exclude it from the deformation calculation. The effect of gravity on coseismic elastic deformation is very small and is neglected in most earthquake cycle deformation models, but gravity is important in the calculation of frictional slip of the fault.

We first generate a pre-stress field by moving the remote seaward and landward model boundaries toward each other against a locked fault. At this stage of “interseismic” strain accumulation, we use a  $\mu'_b$  of 0.04 for the seismogenic and updip segments and 0.004 for the deeper segment. The low strength of the subduction fault is based on the weak fault argument as summarized in Wang and Hu (2006). The nearly zero strength of the deeper part represents a relaxed state after a long time of locking of the seismogenic zone. However, the absolute strength of the fault has no effect on our results. It is the incremental change in fault strength ( $\Delta\mu'_b$ ) at the time of the earthquake that is relevant. A negative  $\Delta\mu'_b$  represents the net effect of weakening, and a positive  $\Delta\mu'_b$  represents the net effect of strengthening. The velocity-dependent evolution of  $\Delta\mu'_b$  through time is not explicitly simulated.

Three examples are shown in Fig. 7. In all cases,  $\Delta\mu'_b = -0.01$  is assigned to the 150-km wide seismogenic zone. This value is chosen to produce an average stress drop of a few MPa (see Fig. 7a), typical of values observed for great subduction earthquakes. The stress drop releases elastic strain energy initially stored in the system, leading to fault slip that represents an earthquake rupture. The deepest segment is assigned a sufficiently large positive  $\Delta\mu'_b$  so that it cannot slip. The examples differ in the  $\Delta\mu'_b$  values assumed for their 50 km wide updip segment, which is the coseismic increase in basal friction in the dynamic Coulomb wedge model.

*Example 1* No trench-breaking rupture (solid line in Fig. 7b). In this case, the strengthening of the updip segment is  $\Delta\mu'_b = 0.036$ . This particular value of  $\Delta\mu'_b$  creates a situation in which the entire updip segment is at failure but is just short of breaking the trench. This is the minimum value of the  $\Delta\mu'_b$  of the updip segment required to prevent trench-breaking rupture and is denoted  $\Delta\mu'_{b,t}$ . The value of  $\Delta\mu'_{b,t}$  depends on the product of the stress drop and the area of the seismogenic zone, a quantity we refer to as “force drop”. That is, if the upper edge of the seismogenic zone is fixed, increasing its downdip width or stress drop gives the same result. For the same model geometry as used for this example,  $\Delta\mu'_{b,t}$  as a function of force drop per unit strike length is shown in Fig. 8, assuming the upper edge of the seismogenic zone is fixed. Using a different model geometry or position of the seismogenic zone upper edge will change the slope of this function.

*Example 2* Trench-breaking rupture (dashed line in Fig. 7b). Given the same force drop in the seismogenic zone, if  $\Delta\mu'_b$  of the updip segment is less than  $\Delta\mu'_{b,t}$ , the rupture will break the trench. Knowing whether coseismic trench-breaking rupture exists or is common awaits future



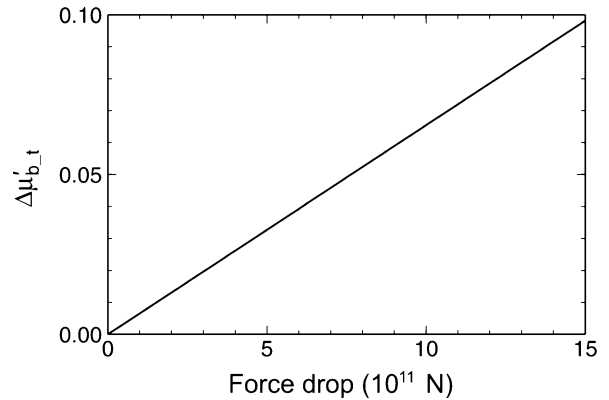
Wedge Mechanics: Relation with Subduction Zone Earthquakes and Tsunamis, Figure 7

Three examples of the stress transfer model. The examples differ in the coseismic strength increase of the updip fault segment, indicated in (b) as  $\Delta\mu'_b$ . a Shear stress drop (or increase) along the fault. b Slip distribution along the fault. c Central portion of the finite element mesh (*thin lines*). The “*thick line*” along the plate interface is actually a group of very densely spaced elements. *Thick gray lines* indicate deformed fault and surface after the earthquake (exaggerated by a factor of 2000)

seafloor monitoring observations. This example shows that a trench-breaking rupture does not necessarily indicate the updip segment exhibits velocity-weakening. Conceivably, the slip of the velocity-strengthening updip segment may be slower than that of the seismogenic zone and may not generate much seismic waves.

**Example 3** Fully buried rupture (dotted line in Fig. 7b). If  $\Delta\mu'_b$  of the updip segment is greater than  $\Delta\mu'_{b,t}$ , rupture may only extend into its lower part. For a very high  $\Delta\mu'_b$ , most of the segment does not slip at all, because a tiny portion immediately updip of the seismogenic zone is sufficient to stop the rupture. This is just the buried-rupture scenario of the crack model commonly used in earthquake simulation [21]. Because most of the updip segment is “protected” and does not experience coseismic stress increase, this scenario is not applicable to the dynamic Coulomb wedge model. A very large stress increase just updip of the seismogenic zone is considered unrealistic.

These examples show the consequences of changes in the strength of the basal fault of the frontal wedge result-



Wedge Mechanics: Relation with Subduction Zone Earthquakes and Tsunamis, Figure 8

Minimum increase in  $\Delta\mu'_{b,t}$  of the updip megathrust segment (denoted  $\Delta\mu'_{b,t}$ ) required to prevent trench-breaking rupture as a function of force drop along the seismogenic zone for the model shown in Fig. 7c, with the upper edge of the seismogenic zone fixed at 50 km from the trench

ing from the rupture of the seismogenic zone. Whether the given strength increase  $\Delta\mu'_b$  can drive the wedge from a stable state into a critical state depends on two factors. First, it depends on the value of  $\mu'_b$  before the earthquake. The value of 0.04 used in the above examples is only one of the numerous possible values. If  $\mu'_b$  is already near a critical value, that is,  $m$  in Fig. 3 for a given  $\alpha$  is already near  $m^c$ , a small increase will do. Conceivably,  $\mu'_b$  before an earthquake may be relatively high if the strengthened state of the fault caused by the previous earthquake has not fully relaxed. Second, given  $\mu'_b$ , it depends on the strength of the wedge material. A weaker wedge becomes critical at a lower  $\Delta\mu'_b$ . The average internal friction value  $\mu$  of an actively deforming frontal prism is lower than the rest of the lithosphere because of its low degree of consolidation and high degree of fracturing. The pore fluid pressure within it, represented by  $\lambda$  in the Coulomb wedge model, may increase due to coseismic compression of the prism, further weakening the wedge, as discussed in Sect. “**Dynamic Coulomb Wedge**”.

### Tsunamiogenic Coseismic Seafloor Deformation

To understand the process of tsunami generation by a great subduction zone earthquake, we must know how the seafloor deforms at the time of the earthquake. Despite the dramatic worldwide improvement of geodetic, seismological, and oceanographic monitoring networks over the past few decades, our knowledge of coseismic seafloor deformation (CSD) is surprisingly poor and is based almost

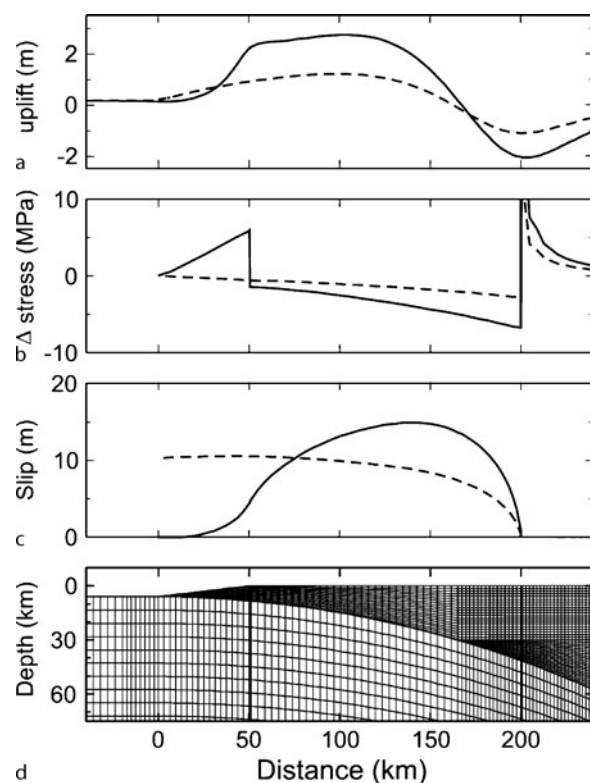
entirely on theoretical models. The problem is the rarity of near-field observations. Except for seafloor pressure sensor records at the time of the M8.2 Tokachi-oki earthquake of 2003 [1] and continuous GPS measurements from islands very near the Sumatra trench at the time of the M8.7 Nias-Simeulue earthquake of 2005 [4,22], most observations are made at sites too remotely located to resolve reliably CSD within about 100 km of the trench. The lack of near-field CSD information causes severe nonuniqueness in the inversion of tsunami, seismic, and geodetic data to determine coseismic slip patterns of the shallow part of the subduction fault, for which the only cure is to introduce a priori constraints on the basis of theoretical models. Until the situation is improved by the establishment of seafloor monitoring systems, we must continue to resort to what we are able to deduce from these models.

The largest uncertainty in our knowledge of the processes that control CSD is how coseismic slip along the subduction fault varies in the downdip direction [21,49]. It may overshadow uncertainties in our knowledge of the timescale of the deformation and the spatial variation of mechanical properties of the rock medium. In comparison, along-strike variations of the coseismic slip, usually described in terms of “asperities”, are much easier to determine using high-density terrestrial monitoring networks. Theoretical models discussed above can help us understand the downdip slip distribution. Using the same type of model as shown in Fig. 5 and Fig. 7 but with a realistic, curved fault, we illustrate how the frictional behavior of the updip segment affects the CSD (see Fig. 9). The simulated earthquakes in the two shown examples have the same “size”, quantified by the seismic moment – the product of rigidity, slip area, and average slip, but they cause very different CSD patterns.

*Example 4* No trench-breaking rupture (solid line in Fig. 9a). This model is similar to the first example in the preceding section in that the updip segment is assumed to strengthen by  $\Delta\mu'_{b,t}$ , and the rupture is on the verge of breaking the trench.

*Example 5* Full trench-breaking rupture (dashed line in Fig. 9a). In this model, there is no velocity-strengthening updip segment, and the seismogenic zone extends to the trench. The stress drop of the resultant much wider seismogenic zone features a monotonic increase from the trench.

The model of full trench-breaking rupture yields much smaller vertical CSD than does the model of no trench-breaking rupture. The reason is two-fold. First, without a velocity-strengthening updip segment to resist rupture, the maximum slip occurs in the most shallowly dipping



Wedge Mechanics: Relation with Subduction Zone Earthquakes and Tsunamis, Figure 9

Two examples showing how the frictional behavior of the updip segment (see Fig. 5) affects CSD. In one example (solid line), the segment strengthens by  $\Delta\mu'_{b,t}$ , and in the other example (dashed line), the segment weakens and thus becomes part of the seismogenic zone. The simulated earthquakes in both examples have the same seismic moment. a Surface uplift. b Stress drop (or increase) along the fault. c Slip distribution along the fault. d Central part of the finite element mesh (thin lines). The two vertical “thick lines” at distances 50 km and 200 km bracketing the seismogenic zone of the model of no trench-breaking rupture are actually groups of very densely spaced elements

near-trench part of the fault where seafloor displacement is predominantly horizontal. This effect would not be obvious had a straight fault geometry been used in the model. Second, without the resistance of an updip segment, the upper plate does not experience horizontal compression and the resultant vertical expansion. If we rescale the two models so that they have the same maximum slip, the model of full trench-breaking rupture will have a greater seismic moment but still a much lower vertical CSD [49]. This result demonstrates the importance of the frictional behavior of the shallowest fault segment in affecting seafloor uplift. However, it addresses only one aspect of tsunamigenic CSD. Many other factors con-

tribute to tsunami generation. For example, although the full trench-breaking model yields very low vertical CSD, its horizontal CSD may contribute to tsunami generation. If the seafloor slopes at angle  $\alpha$ , its horizontal motion  $D$  in the slope direction should raise the seafloor by  $D \tan \alpha$  relative to a fixed water column above, an effect addressed by Tanioka and Satake [43]. The speed of the coseismic slip is also an important factor in tsunami generation. In some rare cases, the rupture is too low to generate much seismic wave energy yet fast enough to generate rather large tsunamis, giving rise to a class of earthquake called tsunami earthquakes [27].

Elastic deformation of the ocean floor as discussed above is the primary cause of tsunami generation during subduction earthquakes, but inelastic deformation can be locally important. For example, although the lower continental slope of active margins is on average the expression of a critically tapered Coulomb wedge, seafloor topography at these margins is rugged at smaller scales due to sedimentation, erosion, and deformation processes, and where the local surface slope is sufficiently high earthquake shaking may trigger gravitational failure. Such submarine “landslides” may have a locally significant effect on tsunami generation. Another potentially important inelastic process is the coseismic activation of out-of-sequence thrust faults (splay faults) in the accretionary prism (Fig. 1a). Splay faults are much more steeply dipping, and their thrust motion will serve to “redirect” the low-angle slip of the megathrust to a higher angle and thus may greatly enhance local seafloor uplift and contribute to tsunami generation [19,36]. As mentioned in the Introduction, from the continuum perspective, such faulting is a manifestation of Coulomb plasticity. At the local scale, it is actually frictional sliding of a contact surface with elastically deforming rocks on both sides. By comparison of the splay faults schematically illustrated in Fig. 1a and the potential failure planes of the critical wedge in Fig. 2b, we can see that some of the splay faults are oriented optimally for thrust failure if the frontal wedge is compressed during a megathrust earthquake.

### Future Directions

The connection between wedge mechanics and great earthquakes and tsunamis at subduction zones is an emerging new field of study. It leads to challenges in both theoretical development and experimental design and thus excellent research opportunities. We need better constraints on how stresses along different downdip segments of the subduction fault evolve with time throughout an earthquake cycle and how the evolution impacts wedge

and seafloor deformation. A number of outstanding questions are to be addressed: Can we constrain the updip limit of the seismogenic zone using wedge morphology? What is the timescale of stress relaxation along the updip segment of the megathrust after an earthquake? Does the seismogenic zone stay locked in the interseismic period? How does pore fluid pressure evolve in an earthquake cycle? How is the transfer of material from the incoming plate to the upper plate (accretion), from the subducted plate to the upper plate (underplating), or from the upper plate to the subducted plate (tectonic erosion) accomplished? What determines the dominant mode of material transfer? What does the spatial change in wedge morphology tell us about changes in the mechanical state of the wedge and the megathrust fault? These questions should be put in the proper context of larger-scale processes such as the viscoelastic relaxation of the mantle following a megathrust earthquake and the deformation of the subducting plate in earthquake cycles [47].

Sandbox experiments designed to study wedge mechanics and dynamic friction experiments designed to study fault mechanics are traditionally separate research activities addressing processes of vastly different timescales. The linkage between subduction earthquakes and submarine wedge evolution suggests the need to combine these experiments. Rapid motion used to simulate earthquakes has begun to be introduced into sandbox experiments [38].

The most promising type of field observation is continuous monitoring of deformation, such as strain and tilt, and fluid pressure using submarine borehole and seafloor observatories. Seafloor elevation change in response to the 2003 Tokachi-oki, northeast Japan, earthquake (M8.2), continuously recorded by two seafloor pressure sensors, clearly indicated coseismic strengthening behavior of the shallowest segment of the subduction fault [1]. Formation fluid pressure changes detected at subsea borehole observatories at the Nankai Trough subduction zone, southwest Japan, have been interpreted to indicate transient aseismic motion of a part of the locked seismogenic zone and/or dynamics of the incoming plate [13]. A number of very-low-frequency earthquakes have been remotely detected within the Nankai Trough accretionary prism using land-based seismic networks [25], revealing the need for near-field observation using seafloor systems. Submarine monitoring in conjunction with land-based monitoring at subduction zones that are currently in different phases of earthquake cycles will allow us to understand the evolution of fault and wedge stresses during the interseismic period. In this regard, cabled seafloor monitoring networks including borehole observatories, being designed or implemented at dif-

ferent subduction zones [26,28] will surely yield valuable data in the near future.

## Acknowledgments

We thank EE Davis, N Kukowski, SE Lallemand, and K Satake for reviewing the article and providing valuable comments. This work is Geological Survey of Canada contribution 20070221.

## Bibliography

### Primary Literature

- Baba T, Hirata K, Hori T, Sakaguchi H (2006) Offshore geodetic data conducive to the estimation of the afterslip distribution following the 2003 Tokachi-oki earthquake. *Earth Planet Sci Lett* 241:281–292
- Barr TD, Dahlen FA (1990) Constraints on friction and stress in the Taiwan fold-and-thrust belt from heat flow and geochronology. *Geology* 18:111–115
- Breen NA, Orange DL (1992) The effects of fluid escape on accretionary wedges 1. Variable porosity and wedge convexity. *J Geophys Res* 97:9265–9275
- Briggs RW, Sieh K, Meltzner AJ, Natawidjaja D, Galetzka J, Suwargadi B, Hsu Y-J, Simons M, Hananto N, Suprihanto I, Prayudi D, Avouac J-P, Prawirodirdjo L, Bock Y (2006) Deformation and slip along the Sunda megathrust in the great 2005 Nias-Simeulue earthquake. *Science* 311:1897–1901
- Brown K, Kopf A, Underwood MB, Weinberger JL (2003) Compositional and fluid pressure controls on the state of stress on the Nankai subduction thrust: A weak plate boundary. *Earth Planet Sci Lett* 241:589–603
- Byerlee JD (1978) Friction of rocks. *Pure Appl Geophys* 116:615–626
- Byrne DE, Davis DM, Sykes LR (1988) Local and maximum size of thrust earthquakes and the mechanics of the shallow region of subduction zones. *Tectonics* 7:833–857
- Chapple WM (1978) Mechanics of thin-skinned fold-and-thrust belts. *Geol Soc Am Bull* 89:1189–1198
- Dahlen FA (1984) Noncohesive critical Coulomb wedges: An exact solution. *J Geophys Res* 89:10125–10133
- Dahlen FA, Suppe J, Davis DM (1984) Mechanics of fold-and-thrust belts and accretionary wedges: Cohesive Coulomb theory. *J Geophys Res* 89:10087–10101
- Davis DM (1990) Accretionary mechanics with properties that vary in space and time. In: Debout GE et al. (eds) *Subduction: Top to Bottom*. AGU Monograph 96, Washington, DC, pp 39–48
- Davis DM, Suppe J, Dahlen FA (1983) Mechanics of fold-and-thrust belts and accretionary wedges. *J Geophys Res* 88:1153–1172
- Davis EE, Becker K, Wang K, Obara K, Ito Y (2006) A discrete episode of seismic and aseismic deformation of the Nankai subduction zone accretionary prism and incoming Philippine Sea plate. *Earth Planet Sci Lett* 242:73–84
- Dewhurst DN, Clennell MB, Brown KM, Westbrook GK (1996) Fabric and hydraulic conductivity of sheared clays. *Géotechnique* 46:761–768
- Dieterich JH (1979) Modeling of rock friction: 1. Experimental results and constitutive equations. *J Geophys Res* 84:2161–2168
- Elliot D (1976) The motion of thrust sheets. *J Geophys Res* 81:949–963
- Enlow RL, Koons PO (1998) Critical wedges in three dimensions: Analytical expressions from Mohr–Coulomb constrained perturbation analysis. *J Geophys Res* 103:4897–4914
- Fletcher RC (1989) Approximate analytical solutions for a cohesive fold-and-thrust wedge: Some results for lateral variation in wedge properties and for finite wedge angle. *J Geophys Res* 94:10347–10354
- Fukao Y (1979) Tsunami earthquakes and subduction processes near deep-sea trenches. *J Geophys Res* 84:2303–2314
- Fuller CW, Willett SD, Brandon MT (2006) Formation of forearc basins and their influence on subduction zone earthquakes. *Geology* 34:65–68
- Geist EL, Dmowska R (1999) Local tsunamis and distributed slip at the source. *Pure Appl Geophys* 154:485–512
- Hsu Y-J, Simons M, Avouac J-P, Galetzka J, Sieh K, Chlieh M, Natawidjaja D, Prawirodirdjo L, Bock Y (2006) Frictional afterslip following the 2005 Nias-Simeulue earthquake, Sumatra. *Science* 312:1921–1926
- Hu Y, Wang K (2006) Bending-like behavior of wedge-shaped thin elastic fault blocks. *J Geophys Res* 111; doi:10.1029/2005JB003987
- Hyndman RD, Wang K (1993) Thermal constraints on the zone of major thrust earthquake failure: The Cascadia subduction zone. *J Geophys Res* 98:2039–2060
- Ito Y, Obara K (2006) Dynamic deformation of the accretionary prism excites very low frequency earthquakes. *Geophys Res Lett* 33; doi:10.1029/2005GL025270
- Juniper K, Bornhold B, Barnes C (2006) NEPTUNE Canada community science experiments. *Eos Transactions, American Geophysical Union* 87(52), Fall Meeting Supplement: Abstract OS34F-04
- Kanamori H (1972) Mechanism of tsunami earthquakes. *Phys Earth Planet Interior* 6:246–259
- Kaneda Y (2006) The advanced dense ocean floor observatory network system for mega-thrust earthquakes and tsunamis in the Nankai Trough – precise real-time observatory and simulating phenomena of earthquakes and tsunamis. *Eos Transactions, American Geophysical Union* 87(52), Fall Meeting Supplement: Abstract OS34F-01
- Kukowski N, von Hune R, Malavieille J, Lallemand SE (1994) Sediment accretion against a buttress beneath the Peruvian continental margin at 12° S as simulated with sandbox modeling. *Geol Rundsch* 83:822–831
- Lallemand SE, Schnürle P, Malavieille J (1994) Coulomb theory applied to accretionary and nonaccretionary wedges: Possible causes for tectonic erosion and/or frontal accretion. *J Geophys Res* 99:12033–12055
- Lohrmann J, Kukowski N, Adam J, Oncken O (2003) The impact of analogue material properties on the geometry, kinematics, and dynamics of convergent sand wedges. *J Struct Geol* 25:1691–1711
- Morone C (1998) Laboratory-derived friction laws and their application to seismic faulting. *Annu Rev Earth Planet Sci* 26:649–696



33. Moore DE, Lockner DA (2007) Friction of the smectite clay montmorillonite: A review and interpretation of data. In: Dixon T, Moore JC (eds) *The Seismogenic Zone of Subduction Thrust Faults*, Columbia University Press, New York
34. Moore JC, Saffer D (2001) Updip limit of the seismogenic zone beneath the accretionary prism of southwest Japan: An effect of diagenetic to low grade metamorphic processes and increasing effective stress. *Geology* 29:183–186
35. Mourgues R, Cobbold PR (2006) Thrust wedges and fluid overpressures: Sandbox models involving pore fluids. *J Geophys Res* 111; doi:10.1029/2004JB003441
36. Park JO, Tsuru T, Kodaira S, Cummins PR, Kaneda Y (2002) Splay fault branching along the Nankai subduction zone. *Science* 297:1157–1160
37. Rice J (2006) Heating and weakening of faults during earthquake slip. *J Geophys Res* 111; doi:10.1029/2005JB004006
38. Rosenau M, Melnick D, Brookhagen B, Echtler HP, Oncken O, Strecker MR (2006) About the relationship between forearc anatomy and megathrust earthquakes. *Eos Transactions, American Geophysical Union* 87(52), Fall Meeting Supplement: Abstract T12C-04
39. Ruina A (1983) Slip instability and state variable friction laws. *J Geophys Res* 88:10359–10370
40. Sibson RH (2003) Thickness of the seismic slip zone. *Bull Seismol Soc Am* 93:1169–1178
41. Smit JHW, Brun JP, Sokoutis D (2003) Deformation of brittle-ductile thrust wedges in experiments and nature. *J Geophys Res* 108; doi:10.1029/2002JB002190
42. Takahashi M, Mizoguchi K, Kitamura K, Masuda K (2007) Effects of clay content on the frictional strength and fluid transport property of faults. *J Geophys Res* 112; doi:10.1029/2006JB004678
43. Tanioka Y, Satake K (1996) Tsunami generation by horizontal displacement of ocean bottom. *Geophys Res Lett* 23:861–864
44. von Huene R, Ranero CR (2003) Subduction erosion and basal friction along the sediment-starved convergent margin off Antofagasta, Chile. *J Geophys Res* 108; doi:10.1029/2001JB001569
45. Vrolijk P (1990) On the mechanical role of smectite in subduction zones. *Geology* 18:703–707
46. Wang CY (1980) Sediment subduction and frictional sliding in a subduction zone. *Geology* 8:530–533
47. Wang K (2007) Elastic and viscoelastic models of subduction earthquake cycles. In: Dixon T, Moore JC (Eds) *The Seismogenic Zone of Subduction Thrust Faults*, Columbia University Press, New York
48. Wang K, He J (1999) Mechanics of low-stress forearcs: Nankai and Cascadia. *J Geophys Res* 104:15191–15205
49. Wang K, He J (2007) Effects of Frictional Behaviour and Geometry of Subduction Fault on Coseismic Seafloor Deformation. *Bull Seismol Soc Am* 98:571–579
50. Wang K, Hu Y (2006) Accretionary prisms in subduction earthquake cycles: The theory of dynamic Coulomb wedge. *J Geophys Res* 111; doi:10.1029/2005JB004094
51. Wang K, He J, Hu Y (2006) A note on pore fluid pressure ratios in the Coulomb wedge theory. *Geophys Res Lett* 33; doi:10.1029/2006GL027233
52. Wang WH, Davis DM (1996) Sandbox model simulation of forearc evolution and noncritical wedges. *J Geophys Res* 101:11329–11339
53. Willett S, Beaumont C, Fullsack P (1993) Mechanical model for the tectonics of doubly vergent compressional orogens. *Geology* 21:371–374
54. Xiao HB, Dahlen FA, Suppe J (1991) Mechanics of extensional wedges. *J Geophys Res* 96:10301–10318
55. Zhao W L, Davis DM, Dahlen FA, Suppe J (1986) Origin of convex accretionary wedges: Evidence from Barbados. *J Geophys Res* 91:10246–10258

### Books and Reviews

- Dahlen FA (1990) Critical taper model of fold-and-thrust belts and accretionary wedges. *Annu Rev Earth Planet Sci* 18:55–99
- Dixon T, Moore JC (eds) *The Seismogenic Zone of Subduction Thrust Faults*, Columbia University Press, New York
- Scholz CH (2003) *The Mechanics of Earthquakes and Faulting*, 2nd edn. Cambridge University Press, Cambridge, 471 p

# List of Glossary Terms

## A

A parameter estimate 1104  
Abrupt climate change 2  
Absolute time 1104  
Accretionary wedge (prism) 1207  
Active fault 278  
Agglomerative (hierarchical) clustering algorithm 127  
Andesite 1082  
Anthropogenic climate change 2  
Anthropogenic emissions 31  
Apparent stress 364  
Aqueous fluid 462  
Arbo viruses 42  
Arbo viruses transmitted by Aedes mosquitoes 42  
Arrival time 230  
Aseismic 79  
Assortative mixing and disassortative mixing 313  
Atmosphere 1  
Atmospheric boundary layer 844  
Attenuation factor  $Q^{-1}$  788  
Attenuation relation (ground-motion prediction equation) 623  
Autoregressive equation 1146

## B

Basalt 1082  
“Baseline” monitoring data 1135  
Bayesian analysis 805  
Beach profile 1008  
Binary forecast 218  
Bingham liquid 1082  
Blue-tongue disease in Europe 42  
Body waves 278  
Bore 1008  
Bore solitons 873  
Bottleneck 517  
Boussinesq equations 1008  
Branching process 338

Bubble collapse 731  
Bubbly flow 1082  
Bubbly liquid 1179

## C

Carbon dioxide fertilization and quality of food 43  
Cascade model 320  
Cascadia 79  
Celerity 663  
Centroid 230  
Centroid moment tensor solution 982  
CGCM 908  
Changes in the pollen season and new pollen 43  
Chaos 573, 825  
Choked flow 1179  
Climate change 109  
Climate change related direct health effects 42  
Climate models 2, 753  
Climate simulation 2  
Climate time scales 1  
Climate variability pattern 2  
Climate variables and forcing 1  
Cluster 127  
Clustering 127  
Clustering coefficient 312  
CMT centroid moment tensor 406  
CO<sub>2</sub> 908  
Coda attenuation factor  $Q_C^{-1}$  788  
Coda waves 278, 788  
Collective intelligence 697  
Community 484  
Community empowerment 484  
Community safety group 484  
Complex and evolving structural systems 201  
Complexity 551, 573  
Computational fluid dynamics 858  
Conduit 1035, 1082  
Confidence ellipse 589

Connectivity distribution or degree distribution 312  
 Continuous phase transitions 825  
 Continuum forecast 218  
 Convolution 406  
 Corner frequency 255, 364  
 Coseismic 79  
 Coulomb plasticity 1207  
 Crack wave 1146, 1179  
 Critical earthquake concept 805  
 Critical exponents 825  
 Critical phenomena 825  
 Critical point 680  
 Crowd 517, 697  
 Crowd disaster 517  
 Crowd turbulence 697  
 Cryosphere 1, 95  
 Crystallization 1082

**D**

Data mining 126  
 Data space 127, 230  
 Data transmission system 175  
 Data vector 127  
 “Decision window” 1136  
 Declustering 825  
 Deconvolution 406  
 Deep earthquake 406  
 Degassing n. (degas v.) 1082  
 Differentiation 1082  
 Diffuse interface 1179  
 Diffuse interface theory 858  
 Dimensionless density 717  
 Direct costs 52  
 Direct search 230  
 Disaster 484  
 Disaster lead time 484  
 Disaster threat 484  
 Discontinuous Galerkin method 765  
 Discount factor 52  
 Discount rate 52  
 Dispersion 255  
 Dispersion, amplitude 1008  
 Dispersion, frequency 1008  
 Diurnal cycle 844  
 Dome 1082  
 Downscaling 753  
 Ductile shear zone 22  
 Dyke 1082  
 Dynamic rupture model 623  
 Dynamic stress change 425

Dynamic stress drop 320, 364  
 Dynamical scaling and exponents 825

**E**

Earthquake 22, 573  
 Earthquake catalog 126  
 Earthquake early warning 982  
 Earthquake early warning system (EEMS) 175, 278  
 Earthquake early-warning 230  
 Earthquake forecasting 573  
 Earthquake forecast/prediction 805  
 Earthquake location 175, 230  
 Earthquake magnitude 175  
 Earthquake mechanism 338  
 Earthquake precursor 278  
 Earthquake prediction 278, 447, 573  
 Earthquake preparedness 573  
 Earthquake quantities 680  
 Earthquake size 255  
 Earthquake source 255  
 Earth-system models of intermediate complexity 2  
 Economics 52  
 Effective risk communication 484  
 El Niño-southern oscillation (ENSO) 31  
 Elastic 79  
 Electro-kinetic effect 447  
 Emergence 697  
 Emergency 484  
 Energy release rate 364  
 Envelope broadening 788  
 Epicenter 230, 278  
 Equilibrium 109  
 Equity weights 52  
 Error 230  
 Evacuation 484, 517  
 Evolutionary optimization 697  
 Excluded volume 717  
 Explosive eruption 1082  
 External factors 1  
 Extreme events 551, 573  
 Extrusive flow or eruption 1082

**F**

Factual statement 1135  
 Failure to predict 218  
 False alarm 218  
 Far-field 278  
 Fast acceleration of unemployment (FAU) 551  
 Faster-is-slower effect 697  
 Fault 22, 278

Fault model 805  
 Fault parameters 1022  
 Fault slip 278  
 Fault strength 320  
 Feature extraction 127  
 Feature space 127  
 Feature vector 127  
 Features 127  
 Feedbacks 3  
 Finite size scaling 825  
 Flow 517  
 Focal mechanism 278  
 Forecast 1135  
 Forecast point 982  
 Forward model 79  
 Fractal 568, 825  
 Fractal dimension 568  
 Fracture network 717  
 Freezing-by-heating effect 697  
 Frequency bands of electromagnetic waves 447  
 Fundamental diagram 517  
 Fundamental modes 255

## G

Game horizon 109  
 Gas slug 858, 1179  
 GCM 908  
 Geodesy 79  
 GHG 908  
 Global climate models or General circulation models 2  
 Global expansion of tropical diseases 42  
 Global search 230  
 Global seismographic network (GSN) 126  
 GNSS 589  
 GPS 79, 889  
 Greenhouse gases 31  
 Grid 126  
 Grid generation 765  
 Ground deformation 731  
 Ground motion intensity measures 623  
 Ground motion uncertainty 623  
 GT 908

## H

Hazard 484  
 Hazard zone 484  
 Health 42  
 Hierarchical organization 312  
 Human bioclimate 42  
 Hypocenter 230, 278, 406

## I

Importance sampling 230  
 Impulse 731  
 Independent events 338  
 Indirect costs 52  
 Infrasound 663  
 Infrasound array 663  
 Initial rupture process 320  
 InSAR 79  
 Intensity, earthquake 278  
 Interactive visualization 126  
 Interferometric synthetic aperture radar (InSAR) 589  
 Intergovernmental coordination group 982  
 Internal solitons 873  
 International GNSS service (IGS) 589  
 Inter-onset time 1104  
 Interseismic deformation 79  
 Inter-story drift 150  
 Inverse problem 1022  
 Inverse problem, inversion 230  
 Inverse theory 79  
 Ionosphere 447, 589  
 IPCC 908  
 IPCC emission scenario 753

## K

Kinematic GPS 589  
 Kinematic rupture model 623  
 k-Means clustering 127

## L

Land surface 1  
 Lane formation 517  
 Leveling 79  
 Likelihood function 230  
 Lithosphere 573  
 Local tsunami 406  
 Locked zone 79  
 Long period seismic events 731  
 Long-period (LP) event 1179  
 Long-term climate statistics 3  
 LP 663

## M

Macroscopic models 517  
 Mafic 1083  
 Magma 1035, 1083, 1179  
 Magma chamber 731, 1083  
 Magma intrusion 1135

Magma, melt, liquid 731  
 Magnitude 255, 406  
 Magnitude, earthquake 278  
 Magnitude saturation 967  
 Malaria and global warming 42  
 Marginal costs 52  
 Marogram 406  
 Mean field theory 680  
 Mean-Field 826  
 Mechanisms for power laws 826  
 Mega-thrust earthquake 589  
 Melt 462, 1083  
 Meta-stability of man-made structures 201  
 Microlite 1083  
 Microscopic models 517  
 Misfit function 230  
 Mitigation 484  
 Model categories 2  
 Model space 230  
 Model test 338  
 Modeling 858  
 Moment magnitude ( $M_w$ ) 589  
 Moment tensor 278, 1146, 1179  
 Moment-tensor 858  
 Monetary valuation 52  
 Multidimensional scaling 127  
 Multiple equilibria 3  
 Multi-resolutional clustering analysis 127  
 Mylonite 22

## N

Navier–Stokes equations 1008  
 N-body solver 127  
 Near-field 278  
 Neo-classical economics 52  
 Net present value 52  
 Network or graph 312  
 Newtonian liquid 1083  
 Newtonian viscous body 95  
 Nonlinear process 1179  
 Non-equilibrium phase transitions 826  
 Non-hierarchical clustering algorithm 127  
 North atlantic Oscillation (NAO) 31  
 Nucleation process 320  
 Nucleation zone 320  
 Numerical methods 765

## O

Objective function 230  
 Oceans 1

Onset 1104  
 Onset time 1104  
 OpenGL 126  
 Origin time 230

## P

Paleoclimate 3  
 Paleoseismology 79  
 Panic 697  
 Parallel algorithms 765  
 Partially molten rock 462  
 Path effects 623  
 Pattern recognition 126  
 Pattern recognition of rare events 551  
 Payoffs 109  
 PDF 218  
 Pedestrian 517  
 Percolation and percolation threshold 717  
 Phase transitions 826  
 Phase-field method 1179  
 Phenocryst 1083  
 Piezo-electric effect 447  
 Plane convex fractures 717  
 Plate tectonics 79, 278  
 Players 109  
 Point process 338  
 Polygenetic vent 1104  
 Posterior pdf 231  
 Postseismic deformation 79  
 Precursory signal 338  
 Prediction 1135  
 Preferential attachment rule 312  
 Premonitory patterns 551  
 Premonitory seismicity patterns 573  
 Preslip model 320  
 Prevention 484  
 Prior pdf 230  
 Probabilistic seismic hazard analysis 278  
 Probability 805  
 Probability density function – PDF 176  
 Probability density function – pdf 230  
 Probability forecast 338  
 Problem solving environment (PSE) 126  
 Projection 753  
 Pyroclastic flow or surge 1083

## R

Radiated or seismic energy 364  
 Radiative transfer theory 788  
 Ramp-up preparations 484

- Random media 788  
 Ray path 231  
 Receiver or station 231  
 Recession 551  
 Rectified diffusion and rectified heat transfer 731  
 Reference frame 589  
 Refraction and inverse refraction diagrams (travel timemap) 1022  
 Regime shifts 2  
 Regional tsunami 406  
 Renormalization group (RG) 680  
 Renormalization group theory 826  
 Repose 1104  
 Resonant frequencies of vibration 150  
 Rheology 1035  
 Rise time 623  
 Risk treatment options 484  
 Rossby solitons 873  
 Run-up height 967  
 Runup, or runup height 1008  
 Rupture 589  
 Rupture duration 624  
 Rupture velocity 624
- S**
- Satellite laser ranging (SLR) 589  
 Saturation 255  
 Scaling 1035  
 Scattering coefficient  $g$  788  
 Scientific visualization 126  
 Seafloor geodesy 889  
 Seismic body waves 406  
 Seismic cycle 79  
 Seismic data-logger 176  
 Seismic energy 255  
 Seismic hazard 176, 279, 805  
 Seismic hazard analysis 279  
 Seismic hazard map 279  
 Seismic magnitude 967  
 Seismic moment 279, 320, 364, 406, 680, 967  
 Seismic moment  $M_0$  256  
 Seismic phase 231  
 Seismic potency 680  
 Seismic radiation 364  
 Seismic risk 279  
 Seismic sensors 176  
 Seismic spectrum 364  
 Seismic surface waves 407  
 Seismic tomographic image 462  
 Seismic wave 279  
 Seismic waves 407  
 Seismicity 231  
 Seismogenic 79  
 Seismogram 231, 279  
 Seismograph 279  
 Seismometer 279  
 Self-organization 697  
 Self-organized criticality 805  
 Self-organized criticality (SOC) 826  
 Shallow and deep water waves 873  
 Shallow earthquake 407  
 Shallow water (long) waves 1022  
 Shallow water wave equations 1008  
 Silicic 1083  
 Simulation point 982  
 Site effects 624  
 Skin effect 447  
 Slip 589  
 Slip distribution 624  
 Slip velocity 320  
 Slip-velocity function (Slip-rate function) 624  
 Slow earthquake 889  
 Social force 697  
 Soil organic carbon 31  
 Soil-structure interaction (SSI) 150  
 Soil-structure interaction 201  
 Soliton 873  
 Source 231  
 Source effects 624  
 Source mechanism 256  
 Source parameters of an earthquake 279  
 Soybean cyst nematode 31  
 Spectral elements 765  
 Spinodal decomposition 826  
 Stable North America 589  
 Start of the homicide surge (SHS) 551  
 Static stress change 425  
 Static stress drop 364, 624  
 Station 231  
 Statistical physics 826  
 Stochastic 338  
 Stochastic process 338  
 Strainmeter 590  
 Strategies 109  
 Stress drop (static stress drop) 320  
 Strike slip fault 680  
 Strombolian activity 858  
 Strong motion seismograph 590  
 Structural health monitoring 150  
 Subduction 79  
 Subduction zone earthquake cycle 1207

Submarine landslide 889  
 Sumatra-Andaman earthquake 126  
 Suprafault 22  
 Surface waves 279

**T**

Tectonics 279  
 Teleseism 279  
 Teleseismic 590  
 Teletsunami 407  
 Telluric current 447  
 Thermal stress and mortality 42  
 Thermohaline circulation 3  
 Total costs 52  
 Trace velocity 663  
 Transfer function 256  
 Travel time 231  
 Tremor 663  
 Triangulation 79  
 Trilateration 79  
 Troposphere 590  
 Tsunami 407, 873  
 Tsunami amplitude 982  
 Tsunami early warning system 982  
 Tsunami earthquake 407, 967  
 Tsunami inundation 1008  
 Tsunami magnitude 967  
 Tsunami warning system 407  
 Tuning parameters 680  
 Turbulence 826, 844  
 2004 Indian Ocean tsunami, the 1022

**U**

Uncertainty 231  
 Universality 680, 826

**V**

Validation 753  
 Vector-borne diseases 42  
 Vegetative index 31  
 Velocity-weakening and strengthening 1207  
 Very long baseline interferometry (VLBI) 590  
 Very-long-period events (VLP) 858  
 Very-long-period (VLP) event 1179  
 Viscoelastic 79  
 Viscosity 79  
 Volatile 731, 1083, 1179  
 Volatiles 1035  
 Volcanic crisis 1135  
 Volcano hazards 1135  
 Volcano monitoring 1135  
 Volcano risk 1135  
 Volcano status 1135  
 Volcano unrest 1135  
 Vulnerability 484, 1135

**W**

Warning 1136  
 Waveform inversion 1146, 1179  
 WEB-IS 126

# Index

## A

- Abrupt climate changes 43
- Accelerating moment release (AMR) 222
- Accelerograph 281
- Accretionary lapilli 1052
- Accretionary prism 977, 1208, 1209, 1216
- Accretionary wedge (prism) 1207
- Acoustic signal 1068
- Acoustic speed, infrasound 665
- Acoustic surveillance for hazardous eruptions (ASHE) 674
- Acoustic velocity, infrasound 665
- Action research 500
- Adaptation 32, 35, 927
  - p*-Adaptation 778
- Adaption
  - autonomous 36
  - planned 36
- ADER-DG method 777
- Adiabatic limit 683
- Adjoint calculations 783
- Aftershock 147, 430, 681, 684, 692, 806, 972, 978
- Agent-based model 698
- Aggregate 1051
- Agricultural production 31
- Air pollution, allergic reactions 47
- Akaike Bayesian information criterion 88
- Akaike information criterion (AIC) 1164
- Akaike's information criterion 988
- Alarm 500, 552
- Albedo feedback 9
- Algorithm 575
  - M8 575
  - MSc 575
  - prediction 580
  - RTP 575
  - SSE 575
- Alignments 1124
  - Hough transform 1124
  - two-point azimuth method 1124
- Allergenic neophytes 47
- Allergies, caused by pollen 47
- Analogue experimentation 1036
- Analysis method 682
- Anelasticity 481
- Anisotropy 683, 701
- Anomalous transmission of EM waves 448, 449
  - anomalous reception of VHF waves 455
- Anticipation 713
- Apennines 179
- Apparent duration time 790
- Apparent stress 264, 267, 368
- APS operator 114
  - property 115
- Arching 525
- Arenal volcano, tremor signal 673
- Arequipa earthquake analysis 667
- Arrest stress 684
- Ash 1191
- ASHE *see* Acoustic surveillance for hazardous eruptions (ASHE)
- Aspect ratio 474
  - equivalent aspect ratio 477
- Asperity 684, 978, 1030, 1215
  - contact 684
  - model, earthquake waveforms 332
- Asymmetric dynamic commons model 120
- Atmospheric model intercomparison project (AMIP) 97
- Atmospheric noise 91
- Atmospheric water vapor 91
- Attenuation 263
- Attenuation relation (Ground-motion prediction equation, GMPE) 623
  - directivity effect 638



empirical 650, 656  
 ground-motion scaling relations 628, 650  
 velocity-pulses 638  
 Australia 485, 500, 507, 509, 512  
 Autocorrelation function 793  
 Autoregressive (AR) model 988, 1160  
 Avalanche 680  
 Avalanche models 97

**B**

b-Values 800  
 Back-arc 798  
 Backscattering enhancement 794  
 Barkhausen noise 680  
 Barriers 499  
   bureaucratic process 500  
   innovator 500  
 Basin and range province 437  
 Bathymetry 977  
 Bayes' theorem 192, 947  
   application of Bayes' theorem to inverse  
   problems 948  
 Bayesian probability theory 807  
 Ben-Zion and Rice model 683  
 Benjamin–Bona–Mahoney (BBM) equation 878  
 Beta statistic 429  
 Bifurcation 9  
 Biofuel 38  
 Block density, fracture networks 728  
 Block-slider 358  
 Blockage 525, 526  
 Blocks 575  
 Blowing snow models 96  
 Body-wave magnitude 260, 268  
   cumulative 269  
   earthquake 413  
 Bonding distance 721  
 Bonds  
   average per object 721  
 Boolean grid 713  
 Born approximation 793  
 Bottleneck 708  
 Boundary conditions 544  
 Boussinesq equation 878  
 Branching models 343  
 Brine shrimp population 912  
 Brine solution 1074  
 Brittle-ductile transition 22  
 Brittle failure 1146  
 Brittle tectonics 72

Broadband body wave, earthquake 413  
 Broadband P-wave moment magnitude 415  
 Broadband record 263  
 Broadband vertical velocity seismogram 415  
 Brune's model 366  
 Bubble cloud 747  
 Bubble dynamics 732  
 Bubble growth 742  
 Bubble nucleation 1041  
 Bubbles 441, 1181  
 Bucharest 178  
 Buildings damaged by earthquakes  
   Imperial County Services Building 156  
   7-story RC building in Van Nuys 156  
 Bulk modulus  
   bubble 737  
   effective 737, 738  
   of a bubble 738  
 Bulk viscosity 468  
 Burgers vector 389  
 Burridge–Knopoff 683  
 Burst 681  
 BZR model 685  
 Båth's Law 342  
   Griffiths' crack theory 342  
   'go-game' model 343  
 Båth's law  
   elastic rebound model 342

**C**

Cairns 501  
 Cairo 178  
 Caldera collapse 1074  
 Calibration 701  
   function, regional 275  
 California 177  
 Caliper diameter, fracture networks 726  
 Camassa–Holm equation 878  
 Campania 178  
 Campi Flegrei Caldera, Italy 1141  
   Monte Nuovo eruption 1141  
 Cap-and-trade 923  
 Capacitors 1099  
 Capacity  
   bottleneck 523  
 Capillary number 1039  
 Carbon cycle 909  
 Carbon market 924, 926  
 Cascade model, earthquake 322  
 Catchments 759

- Cell model 741
- Cellular automata (CA) 344, 806
- Cellular automaton (CA) 533, 683
- Cellular models 734
- Cenozoic climate 13
- Centroid moment tensor 269
  - solution 982
- CFL criterion 769
- CGPS networks 593
- Change
  - in regime 1118
  - in regime, statistically identifiable 1119
- Change-point models
  - CUSUM 356
  - hidden Markov models 356
  - stress-shadowing 356
- Change-point problems 1118
- Chaos 219
- Chaos diagram 203
- Chaotic dynamics 202
- Chaotic motions 202
- Chaotic vibration 202
- Characteristic frequency 739
  - earthquake 413
- Characteristic time 742
- Characteristic time scale 732
- Characterization, dynamic game 114
- Checkerboarding 87
- Chelungpu fault, Taiwan 601
- Chords 726
- Cinder cone 1072
- Circular crack 374
- Circular fault 373
- Circum-Antarctic seaways 13
- Climate 754
  - climate simulations 755
  - key parameters 43
- Climate change 31, 763
  - abrupt 43
  - impact on human health 42, 43, 46
  - impacts on plants 48
  - indirect health effects caused 47
  - migration 49
  - natural 43
  - vector-borne tropical infectious diseases 48
- Climate feedbacks 909
- Climate modeling 916
- Climate simulation 2
- Climate statistics 3
- Climate time scales 1
- Climate variability 6
- Clogging 519
- Cloncurry 502
- Cluster analysis 1124
  - distance metric 1124
  - statistic 1124
- Clustering 129
  - algorithms 135
  - coefficient 312, 316–318
- CMT 976
- CO<sub>2</sub>-concentration, elevated 49
- Coarse graining 147
- Coda attenuation factor 789
- Coda interferometry 801
- Coda waves 788
- Collapse
  - bubble 731
    - of a bubble 745, 746
- Collective behavior 713
- Collective dynamics 697
- Collective event 682
- Collective intelligence 703
- Collective phenomenon 707
- Communication 484, 485, 509, 713
  - action statement 491
  - barrier to change 499
  - core goal 491
  - electronic mass media 488
  - indigenous community 485
  - information flow 491
  - institutional barrier 499
  - integrating theory 493
  - language 492
  - make the threat real 493
  - media 491, 493
  - media support 489
  - medium and message 488
  - meteorological knowledge 492
  - non-English speaking household 496
  - paradigm shift 500
  - phone tree disaster warning model 496
  - policy and laws 490
  - politics 493
  - precautionary principle 492
  - risk communication theory 507
  - safety group 500
  - safety triangle 486, 497, 500, 513
  - simulation 488
  - siren 493
  - target audience 492
  - web 498
  - world view 491

- Community 484, 505, 509
  - awareness 484
  - cultural knowledge base 509
  - empowerment 486, 512
  - indigenous 494
  - passivity to partnerships 490
  - personal contact 498
  - resilience 488
  - safety group 504
  - self-help 505
  - seven steps to community safety 498
- Compact 688
- Compaction length 471
- Complex amplitude 1162
- Complex behavior 683
- Complex frequencies 1162
- Complex social system 713
- Compressibility 735
- Computation of local travel times 940
- Computerized tomographie
  - backprojection 933
  - iterative solutions 934
- Computerized tomography
  - analytical solutions 931
  - filtered projection 933
  - historical overview 929
  - parallel projection 930
- Conceptual model 15
- Conduit 1035, 1082, 1085, 1187
- Configurational entropy 692
- Connectivity
  - geological fractures 717
- Constitutive relationship 467
- Contingency table 225
- Continuum
  - asymmetric 388
  - percolation 719, 721
  - thresholds 721
- Control theory 908
- Conversion relations 258
- Cooperation 698
- Coordination 698
- Coordination problem 709
- Coriolis parameter 994
- Corner frequency 255, 267, 269, 366, 367, 374, 681
- Corner period 264
- Correlation 582
  - long-range 582
- Correlation dimension 1116
- Correlation distance 793
- Correlation functions 356
  - Ripley's K-function 356
- Correlation length 686, 687
- Correspondence
  - compact-valued 115
- Coseismic seafloor deformation (CSD) 1214
- Coseismic slip distribution 84
- Coulomb failure 440
- Coulomb failure function 219
- Coulomb failure stress 433
- Coulomb plasticity 1207–1209
- Coulomb wedge 1207–1209, 1212–1214, 1216
  - dynamic 1210, 1213, 1214
- Counterflow 519, 710
- Crack 1184
- Crack model 688, 1167
- Crack oscillation 1170
- Crack stiffness 1169
- Crack wall 1169
- Crack wave 1165
- Crackling noise 681
- Crater 1195
- Critical dimension 688
- Critical point 680, 807
- Critical slip distance
  - earthquake 334
  - earthquake nucleation process 321
- Criticality 822
  - scale-free characteristics 818
- CROCUS 96
- Cross-correlation 1129
- Cross-validation (CV) 88
- Crowd 697
- Crowd disaster 520, 528, 704
- Crowd turbulence 710
- Crushing 709
- Crustal deformation 80
  - block modeling 607, 609
  - dislocation sources 597
  - inferring source characteristics 600
  - modelling with GPS data 596
  - Mogi deformation source 597
  - sensitivity of GPS data to 600, 609
  - source geometry 597–600
- Crustal deformation monitoring 73
- “Cry wolf” issue 197
- Cryosphere 95
- Cryosphere models 95
- Crystal growth 1090
- Crystallization 1082, 1084, 1087
  - latent heat 1094

- Cutoff function 687  
 Cyclicity 1090  
 Cyclomatic number  
   fracture network 728  
 Cyclone Ingrid 502  
 Cyclone Larry 490, 504  
 Cyclone Tracy 508  
 Cylindrical source 1152
- D**
- Dacite 1049  
 Damage 150, 584  
   diversity 584  
   reduction 584  
 Damping 737, 745  
   acoustic 736  
   viscous 736  
 Dansgaard–Oeschger cycle 11  
 DART 984, 1026  
 Data assimilation 1032  
 Data-logger 176  
 Data transmission system 175  
 Database 182  
   events 184  
   instrumental 182  
   waveforms 184  
 DCG *see* Dynamic commons game  
 Deadly earthquakes 281  
 Death spiral 921, 922  
 Debye equation 480  
 Decaying harmonic oscillations 1160  
 Decision window 1136, 1137  
 Deep earthquake 274  
 Deformation 891  
   basic 386  
 Deformation nucleus 389  
 Deformation of the crack wall 1171  
 Degassing 1086  
 Deglaciation 10  
 Degree of fracturation 720  
 Degree of freedom 683  
 Delauney triangulation 771  
 DEMTER 458  
 Denali fault earthquake 427  
 Dengue fever, effects of climate change 48  
 Density 521, 522  
   dimensionless 728  
   isotropic fracture networks 728  
 Density dependent control 914  
 Density wave 519
- Depinning transition 685  
 Depth phases 268  
 Desired velocity 708  
 Deterministic 683  
 Developing countries 199  
 Differential game 699  
 Differential ground motion 209  
 Differential motion 206  
 Diffusion 1042, 1046  
 Diffusion solution 791  
 Dihedral (torsion) angles 475  
 Dike 1192  
 Dimensionless density 728  
   percolation properties 726  
 Dirac tensors, invariant forms 392  
 Directed search 239  
 Directivity 437  
   effects 641  
 Disaster 484, 518  
   impact severity 488  
   lead time 488  
   management 486  
   mitigation 487  
 Discharge rate 1084  
 Disclination density 393  
 Discontinuities  
   discrete 717  
   in rock 717  
 Discontinuous Galerkin method 765, 770  
 Discrete Fourier transform 1158  
 Discrete wave-number method 193  
 Discreteness 683  
 Discretization 554, 581  
 Diseases 34  
 Dislocation density  
   dislocation-relation 393  
   dislocation-stress 393  
 Dislocation patch 686  
 Dislocation theory 806  
 Disorder 682  
 Dispersion 255  
 Displacement  
   spectrum 263, 267  
 Distance  
   epicentral distance 260  
   teleseismic distance 260  
 Distribution 680  
 Domain  
   problem, sequential compactness 114  
 Dome 1082, 1182  
 Dome growth 1092

- Drainage networks 569
  - Duration magnitude 258
  - Dusty gas 1057
  - Dyke 1085, 1098
  - Dynamic circular crack 375
  - Dynamic commons game (DCG) 120
  - Dynamic friction 684
  - Dynamic game
    - aggregative growth model 118
    - argument 112
    - asymmetry 119
    - Bellman optimality equation 114
    - business-as-usual equilibrium 122
    - characterization 114
    - climate change 109
    - climate change model 109
    - emission factor 124
    - equilibrium 109
    - equilibrium payoff vector 123
    - feasible payoff 116, 117
    - finite argument 112
    - Folk theorem 116, 118
    - game horizon 109
    - global common 119
    - global warming 119
    - individually rational payoff 117
    - model 109
    - most rapid approach (MRAP) 122
    - near-irreversibility 119
    - nonlinearity 120
    - payoff 109, 116
    - player 109
    - second-best problem 123
    - state 109
    - state space 118
    - strategic setting 120
    - strategy 109
    - time cycle strategy 117
  - Dynamic instability 204, 207
  - Dynamic regime 680
  - Dynamic triggering 73
  - Dynamic weakening 685
  - Dynamical downscaling 753
  - Dynamical strengthening 684
  - Dynamical weakening 684
  - Dynamics 1096
    - of earthquakes 681
- E**
- Early warning 150
  - Early warning systems 71
  - Earth
    - orbital parameters 6
  - Earthquake 68, 230, 257, 448, 569, 573, 680, 711
    - a possible model 333
    - Aegean Sea EQ, 2001 452
    - as a critical phenomenon 454
    - Athens EQ, 1999 454
    - characteristic frequency 413
    - Chi-Chi EQ, 1999 454
    - critical slip distance 334
    - direct-search location 238
    - early-warning 230, 249, 251
    - emergency response 231
    - epicenter 230
    - Eratini–Egio EQ, 1995 452
    - estimate of source processes 327
    - estimate of source time functions 326
    - forecasting 573
    - Great Chilean EQ, 1960 453
    - Guam (Marianas) EQ, 1993 448, 453
    - Hawaii region 412
    - hazard assessment 231
    - hypocenter 230
    - infrasound 663
    - Kobe EQ, 1995 449, 453–455
    - Kozani–Grevena EQ, 1995 452, 454
    - large earthquake 324
    - linearized location 238
    - local magnitude methods 413
    - location 230
    - Loma-Prieta EQ, 1989 452
    - Loma-Prieta EQ, 1988 448
    - magnitudes 411
    - main phase 324
    - measurements of initial parts 324
    - model with small nucleation zones 322
    - moment magnitude 413
    - nucleation zone 321
    - objective function 230
    - observations of initial rupture processes 324
    - observatory message 409
    - ordinary phase 324
    - origin time 230
    - P-wave magnitude scale 413
    - P-wave velocity pulses 321, 322
    - P-waves 411, 414
    - preparedness 573
    - probabilistic location 235
    - real-time location 251
    - rupture growth 323

- rupture velocity 321
- S-waves 411
- schematic source models 323
- seismic swarm in Izu Island region, 2000 450, 451
- size 231
- slow 273, 889
- small 325
- source location computing 669
- source parameters 406, 411
- source time functions 332
- Spitak (Armenia) EQ, 1988 453
- Spitak (Armenia) EQ, 1989 448
- strength profile 334
- Strofades EQ, 1997 452
- submarine earthquake 257
- theoretical models 321
- theta 411
- Tokachi-Oki EQ, 2003 453
- tsunami warning 416, 418
- tsunamis 407, 408
- velocity pulse 321
- waveforms 321
- Earthquake catalogs 296, 811
- Earthquake clustering 128, 130, 131, 133, 134, 136, 137, 139, 141, 143–145
- Earthquake clusters 70, 142
  - waveform 331
- Earthquake complexity 265
- Earthquake damage 72
- Earthquake damage detection 152
  - literature review 152
- Earthquake doublets 801
- Earthquake dynamics 137, 219, 375
- Earthquake early warning (EEW) systems 175, 297, 323, 982
  - regional 175
  - site-specific 175
- Earthquake engineering 71, 201, 203
- Earthquake forecasting 71
- Earthquake geodesy 80, 81
- Earthquake-induced disasters 199
- Earthquake infrasound
  - case study Chile 2005 668
  - case study China 2001 667
  - case study Peru 2001 666
- Earthquake light 448, 449, 453
- Earthquake location 69, 293
- Earthquake magnitude 69, 288, 295
- Earthquake monitoring 69
  - in the digital era 289
  - instrumentation 280
    - regional and local networks 289
- Earthquake networks 70, 312–314, 316–318
  - connectivity distribution 312, 314
  - hierarchical organization 312, 316
  - scale-free 313–317
  - small-world 313, 315, 317
  - small-worldness 317
- Earthquake nucleation 72
- Earthquake nucleation process 320
  - critical slip distance 321
  - initial crack 321
  - models 322
  - preslip model 321
- Earthquake physics 72
- Earthquake prediction 320, 447, 448, 573, 805
- Earthquake prediction and forecasting 69
- Earthquake quantity 680
- Earthquake rupture process 260, 624, 634
  - angle of slip, or rake angle 629
  - asperities 634, 635
  - branching fault 637
  - constitutive law (i. e. a friction model) 634
  - directivity 638
  - directivity effect 638
  - dynamic rupture model 623, 633, 634
  - dynamic stress drop 624
  - dynamics of earthquake rupture 627
  - earthquake magnitude 629
  - earthquake scaling laws 634
  - earthquake source dynamics 637
  - earthquake-source inversions 633
  - earthquake source-scaling 652
  - elasto-dynamic equations of motion 634
  - fault dimensions 633
  - fault-zone maturity 636
  - finite-source rupture model 633, 634
  - geometrical fault complexity 636
  - hypocenter 633, 638
  - hypocenter location 635
  - kinematic rupture model 623, 633
  - left-lateral strike-slip 629
  - local slip-functions 633
  - moment magnitude 631
  - moment magnitude,  $M_w$  653
  - multi-scale dynamic simulations 656
  - normal faulting 629
  - physics of earthquake rupture 624
  - point of rupture nucleation 638
  - radiation pattern of P- and S-waves 633
  - representation theorem 633

- right-lateral strike-slip 629
- rise time 623, 633
- rupture duration 624
- rupture finiteness 628
- rupture nucleation 638
- rupture velocity 624, 633
- scale-invariant static stress 631
- seismic moment 624, 629
- seismic potency 631
- self-similar constant stress-drop scaling 624
- self-similar earthquake scaling 653
- self-similar earthquake source scaling 631
- slip distribution 624, 633
- slip heterogeneity 634
- slip vectors 629, 633
- slip-velocity function (slip-rate function) 624, 633
- source properties 625
- source-scaling relations 634, 653
- static stress drop 624, 631
- style-of-faulting 633
- thrust faulting 629
- velocity-pulses 638
- Earthquake scaling laws 73
- Earthquake size 255–258
- Earthquake source 73, 255, 260
  - seismic methods 410
- Earthquake spectra 366
- Earthquake statistics 129
- Earthquake swarms 132
- Earthquake thermodynamics 395
- Earthquake visualization 139
- Earthquake waveforms *see also* Earthquake slip area 332
- Earth's interior 69
- Earthworm 180
- Ecologically sustainable development 486
- Economic impact of climate change 61
- Eddy current 690
- EEPAS model, precursory swarm 350
- Effective medium theory 468
- Effective stress 467, 1208, 1210
- Effects of snow cover on frozen soil 104
- Efficiency 398
- Eigenvalue problem 1162
- El Niño 4
- El Niño-Southern oscillation (ENSO) 35
- ElarmS 178
- Elastic expansion 1088
- Elastic half space 687
- Elastic rebound 607
- Elastic rebound theory 83
- Elastic wave 473
  - attenuation 479
  - dispersion 479
- Elastic wave propagation 304
- Elections 554
  - presidential 556
  - senatorial 555
- Electricity market 909
- Electro-kinetic effect 447, 458
- Electromagnetic signals 70
- Electrostatic aggregation 1053
- Elevated CO<sub>2</sub>-concentration
  - changes in food composition 49
  - nitrogen content in plants 50
- Emergence 697
- Emergency 484
- Emissions from focal regions 448
  - co-seismic signal 448, 451, 456–458
  - co-seismic wave signal 448
  - higher frequency electromagnetic emission 453
  - pre-seismic appearance of EM signals 448
  - pre-seismic signal 457, 458
  - ultra low frequency (ULF) anomaly 452
  - VHF signal 448
  - VLF emission 453
- Empirical ground-motion prediction equations 628
- Empirical trajectory 700
- Energy
  - seismic energy 255
- Energy balance 370
- Energy magnitude 264, 267
- Energy moment ratio 369, 372
- Energy release 398
- Energy release rate 376
- Energy systems 908
- Engineering demand parameters 197
- Ensemble of random 793
- Entropy 224
- Envelope 190, 789
- Envelope broadening 790
- Environment
  - extensional; transtensional 433
  - transpressional; compressional 434
- Environmental systems 908
- Epicenter 255
- Equal differential-time 186
- Equation
  - Korteweg–de Vries (KdV) 877
- Equivalent body force 1148
- Error 230, 581
  - diagram 553, 581

- Eruption 1182, 1183  
   signature, volcanic activity 675  
 Eruptive volume 1105  
 Escape route 711  
 ETAS model  
   ground process 349  
   Hawkes' processes 349  
   process without ancestors 349  
 Etna 1038  
 Euler pole 595  
 Evacuation 484, 508, 511, 517, 704  
   arching 705  
   clogging 705  
   decision matrix 487  
   egress 705  
   jamming 705  
   planning process 505  
   precautionary 492, 504  
   pushing 705  
   route 501  
   social issue 512  
   stampede 705  
   trigger 488  
 Evacuation dynamics 517  
 Evacuation process 518, 540  
   simulation 539  
   simulation example 540, 542  
 Evacuation time, calculation 539, 540  
 Evaporation/condensation 737, 739  
 Event-event correlations 314  
 Evolutionary algorithm 712  
 Evolutionary approach 187  
 Evolutionary calibration 700  
 Evolutionary optimization 711  
 Evolutionary strategy 187  
 Excluded volume 717  
   calculation 720  
   continuum percolation 720  
   isotropic networks 726  
 Exit route 501  
 Expected losses 197  
 Experiment  
   evacuation 527  
 Experimental design 232, 237, 245  
 Experimental petrology 1045  
 Extreme value statistics 819
- F**
- Fabric weakening 22  
 FACE *see* Free air carbon dioxide enrichment  
 Failure stress 684  
 Fall velocity 1051  
 False alarm 197, 1143  
 Fast wavelet transform 161  
 Faster is slower effect 708  
 Fault 717  
   Fault creep 610  
   Fault dip 976  
   Fault impedance 684  
   Fault model 805  
   Fault parameters 984, 1023  
   Fault slip 264  
 Feedback 8, 755, 908  
 Feret diameter, fracture networks 726  
 Ferny Creek 509  
 Finite-difference 769  
 Finite-difference method 210, 1169  
 Finite element 1213–1215  
 Finite-element method 770  
 Finite-size scaling  
   fracture networks 722  
 Finite-source inversions 634  
 First international workshop on rotational seismology  
   and engineering applications 306  
 Floating ice 97  
 Floods 570  
 Floor field  
   dynamic 534  
   static 534  
 Floor field cellular automaton 534  
 Florida 511  
 Flow 521, 522  
   bidirectional pedestrian 523  
   bottleneck 523, 524  
   pedestrian stream 521  
   uni- and multidirectional 523  
 Fluid-filled crack 1165  
 Fluids 701, 977  
   analogies with 701  
 Fluid–solid coupled waves 1167  
 Focal depth 255  
 Fokker–Planck equation 5  
 Food security 31  
 Force chain 711  
 Forced vibration 202  
 Fore-arc 798  
 Forecast 223, 1135  
   binary 223  
   continuum 223  
   long-term forecasts 1136  
   short-term forecasts 1136



- Forecast verification 218
  - Forward problem 1022
  - Fourier slice theorem 932
  - Fractal 357, 568
    - distribution 568
  - Fractal dimension 357, 568, 1117
    - entropy scores 357
  - Fractal spatial structure 681
  - Fracto-emission 1052
  - Fracturation, degree 720
  - Fracture 1191
    - discrete families 727
  - Fracture mechanics 978
  - Fracture networks 718
    - density 728
    - geological 717
    - percolation 719
    - three-dimensional models 719
  - Fracture process, synchronization 384
  - Fractured porous media 717
  - Fractures in rock 717
  - Fragmentation 569, 1043
  - Fragmentation bomb 1048
  - Frame-relay protocol 182
  - Frank vector 389
  - Free air carbon dioxide enrichment (FACE) 49
  - Freezing by heating 708
  - Frequency band of electromagnetic waves 447
  - Frequency domain 7, 1157
  - Frequency-moment distribution 689
  - Frequency size event statistics 680
  - Frequency size statistics 681
  - Friction 440, 684, 809, 978
    - coefficient, effective 1210, 1212
    - law 683
    - parameter 535
  - Frozen ground temperatures 104
  - Fundamental diagram 522–524
  - Fundamental modes 255, 265
  - Funnel 713
  - Fuzzification 147
- G**
- Game
    - dynamic 109
  - Game horizon 116, 118
  - Game theory 109, 713
  - Gap propagation 710
  - Gas 1187
  - Gas volume fraction 1039
  - Gaussian ACF 795, 800
  - Gaussian distribution 683
  - General circulation models 2
  - Geo-complexity 70
  - Geodesy 894
    - submarine 897
  - Geodetic observations of ground deformation 80
  - Geological formations
    - transport processes 717
  - Geological fractures 717
  - Geometrical and material heterogeneity 683
  - Geostationary meteorological satellites 1001
  - Geothermal field 430
  - German Indonesian tsunami early warning system 985
  - Geyser 733, 744
  - Glaciers 101
    - glacier models 101
  - Glass 1191
  - Glass transition 1040
  - Global catalog 1105
  - Global climate change 121
    - issue 119
    - result 121
  - Global climate models 754
  - Global connectivity 723
  - Global mean temperature, change rates 43
  - Global Pareto optimum (GPO) 121
  - Global positioning system (GPS) 73, 81, 590
    - and earthquake response 615
    - and fault slip rates 607
    - and gravity data 605
    - and InSAR data 602
    - and seismic data 602, 615
    - and transient deformation 610
    - and volcanic deformation 601, 605, 614
    - applications of GPS data 594
    - error sources 592
    - high sampling rate 614
    - inferring deformation source from 601, 602, 605
    - modeling of data 596
    - postseismic deformation measurement using 612
    - real-time data 615
    - relative precision 594
    - strength and weaknesses for deformation measurement 607
    - strengths and weaknesses for deformation measurement 602, 614, 615
    - strengths of 594
    - use with other datatypes 602
  - Global-search 230
  - Global seismographic network 289

- Global warming 490
    - photochemical smog 45
  - Global warning system 145
  - Global welfare 121
  - Goudie 493, 505
  - GPO *see* Global Pareto optimum
  - GPS 889, 894, 1026
  - GPS data collection 592
    - campaign GPS measurements 592
    - campaign measurements 592
    - continuously recording GPS 593
    - semi-permanent instrument deployment 616
  - Granular media 711
  - Granular model 474
  - Graph
    - fracture networks 722
  - Greek fire 490
  - Green function 984
  - Greenhouse gases 754
  - Green's function 1031, 1148
  - Green's law 994, 1028
  - Grid environment 143
  - Grid generation 765, 774
  - Grid search in space 1158
  - Ground motion 71
  - Ground-motion complexity 624, 644
    - directivity effect 638
    - rotational motions 655
    - seismic coda 625
  - Ground motion intensity measures 195, 623, 650, 652
    - ground-motion scaling relations 628
    - peak ground 623
    - peak ground acceleration 623
    - peak ground acceleration close 654
    - PGA-value 654
    - precariouly balanced rocks 654
    - response spectra 623
    - rotational motions 655
    - $S_A$ -values 654
    - scalar intensity measures 623
    - strong-ground motions 624
    - strong-motion seismology 625
  - Ground motion uncertainty 623
    - directivity effect 638
    - epistemic uncertainty 653
    - ground-shaking variability 656
    - inter-event variability 653
    - intra-event 653
    - M 6.0 Parkfield earthquake 625
    - M 7.6 Taiwan (Chi-Chi) earthquake 625
    - maximum shaking levels 654
    - microzonation studies 648
    - physical limits to maximum ground motions 654
    - quantifying uncertainty 653
    - random (aleatory) variability 623
    - scientific (epistemic) uncertainty 623
    - velocity-pulses 638
  - GTS 1001
  - Gum rosin 1065
  - Gutenberg and Richter 681
  - Gutenberg–Richter (GR) law 25, 73, 139, 192, 341
    - tapered Pareto distribution 342
- ## H
- Hand calculation method 532, 539
    - complex 539
    - dynamic flow 539
    - fixed flow 539
    - simple 539
  - Harmonic tremor signal, volcanic sound 673, 674
  - Haskell model 379
  - Hawaiian fire-fountaining 1035
  - Hazard 484, 486, 1104
    - earthquake 505
    - fire zone 500
    - flood 491
    - future flood 502
    - map 218, 512
    - tsunami 493
    - type 486
  - Hazards warning 77
  - Health
    - consequences of climate change 43
    - World Health Organization (WHO) definition 43
  - Hector mine earthquake 426
  - Herding 528
    - behavior 705
  - Heterogeneity 713
  - Heterogeneous fault zone 680
  - Hexahedral grids 773
  - Hidden Markov model 1112, 1119
  - High performance computing 781
  - Historical earthquakes
    - characteristic earthquake 345
    - damage rheology 345
  - Historical records, incomplete 1106
  - History 684
  - Homicide, surge 551
  - Hooke's law 1209
  - Horst and graben 977

- Human health
    - climate change 42
    - ozone hole 46
  - Hurricane Katrina 506
  - Hurst exponent 1117
  - Hydraulic diffusivity 471
  - Hydrological cycle 759
  - Hydrological modeling 909
  - Hydrothermal system 733, 744
  - Hydrovolcanism 1044
  - Hypocenter 175, 255
  - Hysteresis 682
- I**
- Ice sheets 102
    - ice sheet dynamics 102
    - land ice models 103
  - Ill-conditioned problems 944
  - Ill-posed linear problems, solution of 943
    - Bayesian approach 947
  - Image 510
  - Imaging 766
  - Imitation 707
  - Impact 486
    - disaster 500
  - Imperial County Services (ICS) building 164
  - Imperial Valley earthquake 164, 166
  - Importance sampling 230, 239
  - Impulse 731, 732
  - Impulse response analysis 166, 169
  - Impulse responses 162
  - IMS *see* International monitoring system (IMS)
  - Inclusion model
    - crack model 474
    - oblate spheroid model 473
    - tube model 473
  - Incompatibility
    - geometric 579
    - kinematic 579
  - Indian Ocean earthquake 880
  - Indicator 554
    - economic 556
  - Inertial glut 1153
  - Information overload 512
  - Information technology 127
  - Infrasonic wave
    - propagation 666
    - propagation modeling 668
    - travel time curve 666
  - Infrasound 74, 663
    - acoustic speed 665
    - acoustic velocity 665
    - analysis method 665
    - basic principle 664
    - earthquake 663, 665
    - measuring 664
    - tsunami 663, 669
    - volcano 663, 672
  - Infrasound array 664
  - INGV 184
  - Inhomogeneous earthquake fault 683
  - Inhomogeneous fault zone 694
  - Initial crack, earthquake nucleation process 321
  - InSAR 602
  - Instability 575
    - geometric 578
    - physical 577
  - Institutional barrier 499, 509
  - Integration
    - limit problem 114
  - Intensity cross product 1128
  - Intensity of a process 1107
    - non-parametric 1107
  - Intensity spectral density 795
  - Inter-onset time 1104
  - Inter-person distance 521
  - Inter-story drift 150, 157, 167
    - outstanding issues 158
  - Interception models 96
  - Interdisciplinary modeling 926
  - Interferogram 82
  - Interferometric synthetic aperture radar 81
  - Intergovernmental coordination group 982
  - Intermittent flow 708
  - International monitoring system (IMS) 664
  - International working group on rotational seismology 306
  - Internet 512
  - Intersection 520
    - per fracture 723
  - Interseismic deformation 84, 607
  - Intra-plate earthquakes 972
  - Intrinsic absorption 792
  - Inverse problem 230, 1022
  - Inverse refraction diagram 983, 1022, 1028
  - Inverse theory 85
  - Inversion 230, 899
  - Inverted pendulum 204

- Ionosphere
  - GPS-TEC 454, 458
  - total electron content (TEC) 448
- IPCC
  - fourth assessment report 101
- Islands
  - Hawaiian 890
- ISNet 178
- Isotropic networks 726
- Isotropic scattering model 796
- Isotropic scattering process 790
- Istanbul 178
- Iwo Jima, Japan 430
  
- J**
- Jamming 519
- Japan 177
- Japanese Meteorological Agency 177, 982
- Jerky response 681
- JMA magnitude 988
  
- K**
- Kadomtsev–Petviashvili equation 878
- Kagan–Jackson models 349
- Katmai 432, 437, 439
- KdV equation 877
- Kilauea volcano, Hawaii (USA) 601, 1142
  - baseline monitoring 1142
- Kirchhoff coda migration 797
- Krakatau, 1883 1036
- Kyoto protocol 923
  
- L**
- L2-norm 186
- Lake ice 97
- Laminar 710
- Land surface schemes in GCMs 97
- Landers earthquake 426, 603
- Landslide 570
  - submarine 889
- Lane formation 702
- Language 485
- Large earthquakes 324
  - aftershock 332
  - estimate of source processes 331
  - initial rupture processes 331
  - large amplitude waveforms 324
  - main phase 324
  - mainshock 332
  - observed waveforms 325
  - ordinary phase 324
  - P-waves 324
  - phases 333
  - velocity waveforms 325
  - waveforms 331, 332
- Lattice coordination number 720
- Lattice percolation 719
- Lava domes 1083
- Lava flow 1035
- Law
  - Darcy 1094
- Lead time 185
- Learning
  - asymmetrical avoidance behavior 698
  - process 698
- Least-squares estimator
  - damped 945
  - generalized 949
  - ordinary 945
- Level
  - operational 530
  - strategic 529
  - tactical 530
- Likelihood function 230
- Likelihood test 224
- Linear dynamic system 1160
- Linear inverse theory 1155
- Linear momentum 1153
- Linking 509
- Liquid 731, 1181
- Lithosphere 573, 792
- Lithosphere-atmosphere-ionosphere (LAI) coupling
  - 449, 454, 456–458
- Load balancing 780
- Local control center 179
- Local earthquake 789
- Local earthquake magnitude methods 413
- Local site effects 627
- Local travel times, computation of
  - bending methods 940
  - finite-difference approximation 941
  - Huygens’s principle 941
  - pseudo-bending method 940
  - shooting method 941
  - shortest-path method 941
- Location 175
- Logarithmic correction 686, 687
- Logistic regression 351
- log(*t*) 684
- Loma Prieta 956

- Long period 731
  - Long-range elastic stress transfer 684
  - Long valley caldera 430, 435, 438, 442, 1141
  - Long-wave approximation 993
  - Longitudinal and transverse modes 1170
  - Love or Rayleigh wave 440
  - Love wave 436
  - LW *see also* APS operator
- M**
- M8 series 351
  - Macroscopic variable 464
  - Magma 441, 731, 1035, 1180
  - Magma chamber 731, 740, 1083, 1085, 1089, 1095
  - Magma discharge rate 1092
  - Magma extrusion 1096
  - Magma intrusion, subsurface intrusion 1143
  - Magmatic and hydrothermal fluids 1146
  - Magnet 680
  - Magnetic field 571
  - Magnification curve 256, 261
  - Magnification, static 258
  - Magnitude 175, 255, 256, 258, 262, 984
    - automatically determined 273
    - band-limited 263
    - broadband 263
    - earthquake 411
    - hawaii earthquake 414
    - local 175, 256, 258, 260
    - local earthquake 413
    - rapid 260
    - Richter 175
    - teleseismic 273
    - traditional amplitude based 413
  - Magnitude determination
    - automatic 275
    - real-time 275
  - Magnitude discrepancy 970
  - Magnitude saturation 265, 267, 268
  - Magnitude scale 257, 258, 265
  - Magnitude units 258
  - Malaria, effects of climate change 48
  - Malthus–Verhulst model 15
  - Man-made structure 201
  - Management science 908
  - Mantle magnitude 264
  - Mantle magnitude ( $M_m$ ) method
    - tsunami warning 416
  - Mantle surface-wave data 269
  - Markov approximation 789
  - Markov chain 1122
    - maximum likelihood 1122
  - Markov perfect equilibrium (MPE) 111
  - Markov process 1112
  - Markovian game 109
  - Markovian strategy 111
    - pure stationary 116
  - Markov–perfect equilibrium 122
  - Mass advection 1152
  - Mass density 684
  - Mass diffusion 736
  - Mathematical modeling 488
    - Eigen plane 490
  - Matrix 943
    - condition number 944
    - Moore–Penrose generalized inverse 943
    - null space 944
  - Mature localized fault 693
  - Maximum a posteriori estimator 948
  - Maximum nonlinear response 207
  - Mean field 684
    - theory 680, 686
  - Mean free path 791
  - Mean moment rate profile 690
  - Mechanism of pre-seismic EM phenomena 455
  - Media 502, 507
  - Medium resolution imaging spectrometer 91
  - Melt 731, 1193
  - Melt segregation 471
  - Mesh partitioning 780
  - Meta-stability 201
  - Methane 38
  - Mexico 177
  - Microearthquakes 325
    - velocity pulses 328, 329
  - Microlite 1083, 1093
  - Microscopic simulation 699
  - Microscopic variable 464
  - Mid-Pleistocene transition 4
  - Migration
    - climate change 49
  - Missed alarm 197
  - Mitigation 36, 484, 506, 927
    - disaster mitigation 487
  - Mixing 313, 315, 317
    - assortative 313, 317
    - disassortative 313, 317
  - Miyagi-Oki earthquake, tsunami 670
  - Mode-switching 686, 692, 694
  - Model 221
    - Blue–Adler 533

- fluid-dynamic 531
  - fracture networks 718
  - Fukui–Ishibashi 533
  - gas kinetic 531
  - Gipps–Marksjös 534
  - GPS measurements 359
  - lattice-gas 535
  - optimal-velocity 536
  - precursor events 359
  - seismicity based 221
  - social-force 532
  - stress release model 359
  - Modeling 511
    - forward and inverse approaches 74
  - Modeling approach
    - continuous 530
    - deterministic 530
    - discrete 530
    - force-based 530
    - high fidelity 531
    - low fidelity 531
    - macroscopic 530
    - microscopic 530
    - rule-based 530
    - stochastic 530
  - Modeling volcanic eruptions 77
  - Models 358
  - Moderate resolution imaging spectroradiometer 91
  - Molchan diagram 226
  - Molecular dynamics (MD) 136
  - Moment
    - seismic moment 256
  - Moment history 259
  - Moment magnitude 259, 265, 267, 377, 378, 988, 1024
    - earthquake 413
    - fundamental earth's modes 267
  - Moment measures 356
  - Moment rate 266
  - Moment-rate 269
  - Moment rate shape 688
  - Moment tensor 269, 365, 1150
  - Monitoring
    - real-time 895
  - Monodisperse fractures 723
  - Monodisperse objects, continuum percolation 721
  - Monotonic model 685
  - Monte-Carlo method 791
  - Moral 512
  - Mortality
    - thermal stress 45
    - weather extremes 45
  - Mount Pinatubo (Philippines) 1139
  - Mount St. Helens (USA) 1139
  - MPE *see* Markov perfect equilibrium (MPE)
  - MRAP *see* Most rapid approach (MRAP)
  - MS fractional fluctuation 793
  - Multi-dimensional feature space 132, 134
  - Multi-dimensional scaling 135
  - Multi-dimensional space 140
  - Multi-resolution analysis (MRA) 160
  - Multifractal analysis 1117
  - Multiple equilibria 3
  - Multiple rupture 269
  - Multiple vents 1122
  - Multiscale 130
  - Mwp 1001
  - $M_{wp}$  method, tsunami warning 415
  - Mylonite zones 399
- ## N
- Naples 178
  - Nash equilibrium (NE) 111
  - Natural climate changes 43
  - Natural time 448, 449, 451
  - Natural time analysis 449
  - Navier–Stokes equations 1088
  - NE *see* Nash equilibrium (NE)
  - Nearest neighbor method
    - average connectivity 313, 317
  - NEIC 259
  - Nervousness 709
  - Network 576
    - fault 576
    - seismic 260
  - Newtonian viscous body 95
  - NIED 192
  - Nitrogen content, plant tissue 50
  - Nodes 576
  - Noise-induced transition 708
  - Non-English-speaking household 511
  - Nonisotropic scattering 794
  - Nonlinear effect 731, 739, 744
  - Nonlinear oscillation 747
  - Nonlinear response 212
  - Nonlinear site effects 628
  - Nonlinear wave propagation in a building 209
  - Nonlinearity 207
  - Non-monotonic model 685
  - Non-stationary activity 1118
  - Normal fault 1004
  - North Atlantic oscillation (NAO) 35

- North west Pacific tsunami advisory center 986  
 Northridge earthquake 936, 950  
 Northumbria University 506  
 Novelty analysis 170  
 Novelty analysis 167  
 Novelty detection 160  
 Nozzle 1188  
 Nucleation 1090  
 Nucleation zone  
   earthquake 321  
 Numerical methods 765, 909  
 Numerical studies 683  
 Numerically 687
- O**
- Obligation 512  
 Observation 756  
 Observatory earthquake message 409  
 Oct-tree 189  
 Omega-squared model 367  
 Omori 439  
 Omori's law 341, 684  
 Onchocerciasis, effects of climate change 48  
 Opinion formation 713  
 Optimization 585, 712  
 Ordinary least-squares estimator (OLSE) method 949  
 Orientation distribution  
   continuum percolation 721  
 Orientation, fracture networks 718  
 Oscillations 520  
   pedestrian 520  
 Oscillatory flow 703  
 Outer rise 972, 978  
 Outlier 189  
 Over-crowding 707  
 Over-dispersed 1111  
 Ozone concentration, Germany 46  
 Ozone hole, human health 46
- P**
- P and S wave velocity models for Taiwan 953  
 P-Coda 800  
 P-Wave 411, 414  
   large earthquakes 324  
 P wave 257, 263, 432, 436  
 P-Wave magnitude  
   short-period 263  
 P-Wave magnitude scale 413  
 P-Wave velocity pulse 321, 322  
 Pacific tsunami warning center 259, 265, 986  
 Pacific tsunami warning center (PTWC) 408  
 Paleoclimate 3  
 Paleoseismological 344  
 Panic 520, 525, 528, 697  
 Parabolic wave equation 795  
 Paradigm 582  
 Parallel algorithms 765  
 Parallel computing 92  
 Parameter  
   optimization 701  
   space 86  
 Parkfield segment 806  
 Partial differential equations 766  
 Particle size distribution 1049  
 Partitioning of elastic energy 374  
 Path effects 623, 627  
   basins and other deterministic deviations from a  
     flat-layered model 623  
   broadband wave-propagation computations 656  
   far-field displacements of P- and S-waves 641  
   incoherently scattered wave energy 642  
   random heterogeneities in the three-dimensional  
     velocity-density structure 623  
   ray theory 641  
   wave-propagation in heterogeneous media 624  
   waves in a flat-layered attenuating earth 623  
 Pattern 574  
   earth-specific 583  
   premonitory 574  
   universal 583  
 Pattern informatics (PI) 221  
 Pattern recognition 564, 580, 1128  
 Payoff  
   Markovian strategy 116  
 Peak ground acceleration 179  
 Pedestrian 518, 697  
 Pedestrian area module 521  
 Perception 508, 713  
 Percolation 717  
   fracture networks 719  
   probability of 723  
   transition zone 725  
 Percolation cluster (IPC)  
   probability of 724  
 Percolation models 343  
 Percolation parameter 724  
 Percolation properties 726  
   fracture networks 722  
   geological fractures 717  
 Percolation threshold 722, 724, 725

- Performance 581  
   algorithm 581  
 Period distribution 315, 317  
 Periodicity 1108  
 Periodicity of earthquakes 340  
   earth tides 340  
 Permafrost  
   dynamics 105  
   models 105  
 Permeability 441, 465, 1042, 1086, 1094  
   fracture networks 728  
 Pest 34  
 Phantom panic 708  
 Phase diagram 693  
 Phase transition 683, 807  
 Phasic average 465  
 Philippine Institute of Volcanology and Seismology (PHIVOLCS) 1140  
 Photochemical smog, global warming 45  
 Photosynthesis  
   temperature dependence 49  
 Physical interaction force 706  
 Physical parameterizations 756  
 Physical processes in volcanoes 77  
 Picking 189  
 Piezo-electric effect 447, 449  
 Piezo-electric polarization 456  
 Plane convex fractures, excluded volume calculation 720  
 Plants  
   impacts of climate change 48  
 Plate motions, measurement using GPS 595  
 Plug 1088, 1188  
 Point source 266, 269  
 Point source model 267, 269  
 Poiseuille solution 1091  
 Poisson distribution 340  
   Poisson process 341  
 Poisson process 1108  
   nonhomogeneous 1109  
   simple (homogeneous) 1108  
   stationary 1109  
 Poisson ratio 1095  
 Politics 511  
 Pollen allergies 47  
 Polydisperse fractures 718, 723  
 Polygons, fracture networks 718  
 Population 490  
 Pore fluid pressure 1208, 1209, 1212  
   Hubbert–Rubey fluid pressure 1209  
 Pore geometry 473  
 Pore pressure 441  
 Poroelasticity 472  
 Porosity 473, 1043, 1099  
 Porosity wave 471  
 Porous media  
   fractured 717  
 Port Douglas 504  
 Positive hole (p-hole) 449, 455–458  
 Possible source geometries 1158  
 PostgreSQL 182  
 Postseismic deformation  
   afterslip 612  
   and high-rate GPS 615  
   poroelastic response 612  
   viscoelastic response 612  
 Power law 711  
   distribution 681  
   scaling behavior 686  
   size distribution 26  
 Powersim 909  
 Precipitation 437, 757  
   monthly mean precipitation 753  
   summer precipitation 761  
   winter precipitation 760  
 Precursors  
   complete model 351  
   of volcanic eruptions 1128  
   partial model 351  
 Predictability 574  
 Prediction 551, 1135  
   advance 565  
   algorithm 552  
   problem 551  
   targets 553  
 Predominant period 190  
 Preparation 484, 505  
   community 485  
   hazard zone 485  
   household 485  
   household safety preparedness and action index 488  
   individual 485  
   precautionary action 485  
   ramp-up 484, 486  
 Preparedness 488  
   conceptual shift 490  
   growth industry 490  
   holistic 497  
   household preparedness and safety action index 489, 497  
   institutional barrier to change 489  
   knowledge base 488



people's propensity to respond 488  
 possible travel mode 488  
 public participation 491  
 well-being 488  
 PreSEIS 178  
 Preslip model  
   earthquake nucleation process 321  
 Pressure 705  
 Pressure recovery 742  
 Pressure-stimulated current (PSC) 455, 458  
 Pressure-stimulated polarization effect 448  
 Principal component analysis 356  
 Probabilistic seismic hazard analysis (PSHA) 194,  
   623, 652  
   ground-motion prediction 652  
   microzonation studies 648  
 Probability density function 176, 179, 187, 230  
   Gaussian 948  
 Probability distribution 684  
 Probability distribution functions 763, 947  
 Probability forecasts 353  
   entropy scores 354  
   Kullback–Leibler distance 354  
   probability gain 354  
   R-score 355  
 Problem  
   dynamic game 113  
 Processing GPS 894  
 Propagation modeling, infrasonic wave 668  
 Pseudo-spectral method 769  
 Pseudogas 1058  
 Pseudospectrum 7  
 PSMS *see* Pure stationary Markovian strategy  
 Psychology 508  
   subjective uncertainty 507  
 PTWC *see* Richard H. Hagemeyer pacific tsunami  
   warning center  
 Pulsation 1092  
 Pure stationary Markovian strategy 116  
 Pyroclastic flow 1053, 1072

## Q

Q Factor 1162  
 Quantifying uncertainty 653  
 Quasi-monochromatic waves 795  
 Quasidynamic fault models 810  
 Quasistatic fault models 810  
 Quaternary volcanoes 798  
 Queensland 500, 511  
 Quenched random “pinning” force 684

## R

Rabaul Caldera (Papua New Guinea) 1140  
   false alarm 1140  
   stage-2 alert 1140  
 Radiated energy 368  
 Radiation models 370  
 Radiation pattern 264, 268, 366  
 Radiative transfer theory 790  
 Radiocarbon dating 1106  
 Radon transform  
   inverse 931  
 Random field Ising model (RFIM) 682  
 Random graph 312, 315, 316  
 Random inhomogeneity 789  
 Random stress drop 687  
 Random walk 686  
 Range of interactions 694  
 Rate/state variable friction law 22  
 Rayleigh–Plesset equation 734  
 Real estate 500  
 Real-time seismology 291  
 Rebound release 400  
 Recession 551  
   economic 556  
   end 559  
   start 558  
 Recharge 437, 438  
 Recording damaging ground shaking 292  
 Recovery 488  
 Rectified diffusion 731, 733, 739, 744  
 Rectified heat transfer 739, 747  
 Refraction diagram 1022, 1027  
 Regime shifts 2  
 Regional catalogues 347  
   declustering 347  
   residuals 348  
 Regional climate models 755  
 Relative intensity (RI) 221  
 Relaxation 480  
 Relaxation time 735, 1040  
 Reliability 581  
   algorithm 581  
 RELM test 224  
 Remote visualization 142  
 Renewal models 345  
 Renewal process 1110  
   gamma 1111  
   Gaussian 1111  
   likelihood 1110  
   log-logistic 1111

- lognormal 1111
  - mixture of Weibull 1112
  - power-law 1112
  - tests 1110
  - Weibull 1111
  - Renormalization group (RG) 680, 682
  - Repeated game 109
    - model 110
  - Repose time 1104
  - Resolution 87
    - matrix 946
  - Resonance 17
  - Resonance of a fluid-filled resonator 1146
  - Resonant frequency 734, 737
  - Result
    - dynamic game 112
    - stochastic game 112
  - REV (representative elementary volume) 464
  - Reverse fault 1004
  - Reversing 582
  - Rheology 735, 1035, 1090, 1095
    - Bingham 1096
    - Newtonian 1093, 1096
    - non-Newtonian 1093
  - Rhyolite 1046
  - Rigidity 974, 1095
  - Rising 582
  - Risk 484
  - Risk communication 484
  - River restoration 916
  - ROC diagram 225
  - Rock 717, 1180
    - discrete discontinuities 717
  - Romania 178
  - Rotation effects 384
    - seismic 384
  - Rotation motions, spin and twist 388
  - Rotational ground motions 305
  - Rotational seismology 304
  - Rough stress field 688
  - Runaway 691
  - Rupture area 264
  - Rupture duration 260, 263, 268, 269, 273
    - automatic estimates 272
  - Rupture evolution 811
  - Rupture length 265, 267
  - Rupture model 266, 267
  - Rupture process 263, 270, 971
  - Rupture slowness estimation, tsunami warning 418
  - Rupture velocity 267, 269, 376
    - earthquake 321
- ## S
- S-Coda waves 789
  - S wave 257, 259
  - S-wave 411
  - SAC 184
  - Safe shelter 508
  - Salmon population 909
  - San Andreas fault 437
  - San Fernando basin 949
    - P* wave velocity model 950, 951
    - SCEC model 950
  - San Fernando earthquake 950
  - Santorini 1049
  - Satellite altimeters 1026
  - Satellite positioning 591
    - GPS signal 591, 616
    - integer ambiguity 591
  - Saturation 255, 258, 263, 265
    - spectral saturation 266
    - time-window 266
  - Scale-invariant quantity 680
  - Scaling 1035
  - Scaling laws 222, 364, 687
    - of earthquakes 364
    - of seismic spectrum 371
  - Scaling result 686
  - Scalogram 421
  - Scarce resource 713
  - Scattering
    - slab 798
  - Schlafkrankheit, effects of climate change 48
  - Sea ice
    - ice dynamics 98
    - thermodynamic sea ice model 99
  - Seafloor deformation 1023
  - Seamount 977
  - Second-order phase transition 686
  - Sediments 975, 977
  - Segregation 704
  - Seismic 231
    - phase 231
    - stations 232
    - travel-times 234
    - velocity model 234
  - Seismic albedo 792
  - Seismic coupling 977
  - Seismic cycle 73, 79, 815
  - Seismic efficiency 368
  - Seismic energy 260, 264
  - Seismic gaps 84

- Seismic hazard 176, 218, 805
  - real-time 176
- Seismic inversion 974
- Seismic methods, earthquake source 410
- Seismic moment 175, 259, 262, 365, 397, 398, 680, 984, 1024, 1215
  - scalar 263, 266
  - small earthquakes 327
- Seismic monitoring of magmatic and hydrothermal activity 1171
- Seismic network 281
- Seismic nucleation phase 327
- Seismic phases
  - P-waves 176
  - S-waves 176
  - surface waves 176
- Seismic potency 680, 681
- Seismic sensors
  - accelerometer 176
  - seismometer 176
- Seismic stations
  - broadband 179
  - short period 179
  - strong motion 179
- Seismic tomography 478, 928
  - arrival-time 934
  - arrival-time tomography 928
  - checkerboard model 947
  - checkerboard test 955
  - comparison with X-ray tomography 935
  - decoupling of the earthquake location and tomography problems 938
  - examples 949
  - Fermat's principle 937
  - local slowness tomography 937
  - local velocity tomography 937
  - solution iteratively 946
  - solution roughness 945
  - teleseismic tomography 942
- Seismic wave propagation 69
- Seismic waves 69
- Seismicity 70
- Seismicity as a critical phenomenon 451
- Seismicity parameters 140
- Seismicity patterns 805
- SeismNet manager 182
- Seismo-electromagnetics 448
- Seismogenic zone 1211–1216
- Seismogram 1183
- Seismograph 257, 258, 262, 280
  - broadband seismograph 259
  - long-period 262
  - short-period 262
- Seismometer 256
- Self-affine fractal 571
- Self-help 485
- Self-organization 702
- Self-organized criticality (SOC) 28
- Self-organizing criticality (SOC) 344
- Self-similar solutions 375
- Self-similarity 358
- Semiotics 488
- Sequential compactness, domain problem 114
- Serial dependence 1109
- Seven steps to community safety 500, 513
- Shallow water theory 1025
- Shape
  - fractures 727
- Shape factor, polydisperse fractures 725
- Shear band 396
- Shear viscosity 468
- Shear wave 257
- Shetland 506
- Shinkansen 177
- Shockwave 711
- Signal identification, infrasound 664
- Signal-to-noise ratio 264
- Sill 1188
- Silly putty 1063
- Similarity 583
- Simple model 683
- Single force 1152
- Singular value decomposition
  - of a matrix 939
- Site effects 624, 625, 646
  - asperity 638
  - bedrock motions 646
  - crack propagation 638
  - directivity 638
  - dynamic rupture modeling 640
  - effects of local site conditions 624
  - energy and the energy radiated by seismic waves 638
  - high-frequency seismic radiation 639
  - hypocenter location 638
  - kinematic source parameters 640
  - liquefaction effects 650
  - local directivity effects 638
  - local sedimentary cover 624
  - local slip-velocity function 638
  - new crack surface (fracture energy) 638
  - nonlinearity 642, 646, 650
  - nucleation phase 639

- propagating crack tip 638
- representation theorem 641
- rupture velocity 639
- self-similarly expanding crack model 640
- site-amplification 646
- site-correction factors 649
- slip distribution 638
- slip-velocity function and rise time 640
- soil classification 647
- source-site geometry 651
- strong geologic contrasts 624
- strong topographic effects 644
- super-shear speeds 640
- temporal rupture evolution 638
- topography 624
- water-table variations 624
- Situation
  - competitive 525
  - normal 525
- Size-predictable 1115
- Skeleton 467
- Skin effect 447
  - skin depth 447, 452
- Slant stack 931, 933
- Sliderblock 683
- Sliding state 685
- Slip 373, 975
- Slip-velocity 269
- Slow earthquake 75, 274
- Slow slip events 610
  - and nonvulcanic tremor 610
  - and seismic hazard 611
  - geodetic observations 610
  - temporal characteristics 611
- Slow wave 1165
- Slowness 935
- Slug 1198
- Slug flow 1067
- Slumps 968
- Small earthquake
  - estimate of source processes 327
  - initial phase 325
  - main phase 325
  - waveforms 326, 328, 330
- Small-scale inhomogeneities 801
- Smoothing 559
- SMS *see* Stationary Markovian strategy (SMS)
- Snow cover 95
- SNOWPACK 96
- SNTHERM 95
- Social aspect 512
- Social cost of carbon 61
- Social field 699
- Social force 699
- Social phenomenon 484, 485
- Social self-help 486
- Soft magnet 686
- Software tool, comparison of commercial 543
- Soil carbon sequestration 37
- Soil-structure interaction 150, 201, 205, 213, 772
  - example 158
  - frequencies of vibration 160
  - Millikan library 160
  - shear wave velocity 159
- Soliton 878
- Sompi method 1160
- Source complexity 266, 269, 273
- Source depth 274
- Source durations 971
- Source effects 627
- Source mechanism 256, 264, 268
  - of VLP signals 1158
- Source parameters 179
- Source spectrum 266, 971, 973
  - seismic 264
- Source time function 366, 1150
  - earthquake 326, 332
- Spatial correlation 88
- Spatial heterogeneity 807
- Spatial periodicity, fracture networks 722
- Spatial random fields 636
  - correlation lengths 636
  - fractal dimension 636
  - Hurst number 636
  - von Karman auto-correlation function 636
- Spatio-temporal intensities
  - Bayesian 1121
  - kernel 1120
  - nearest-neighbor 1120
  - nearest-neighbor kernel 1120
- Spatio-temporal pattern 698
- Spatiotemporal stress field 807
- SPE *see* Subgame perfect equilibrium (SPE)
- SPE outcome path 118
  - dynamics 118
- SPE payoff 116
- Spearman correlogram 1129
- Spectral element scheme 770
- Spectral elements 765
- Spectroscopy, volcanic eruption 674, 675
- Spectrum 681

- Speed
  - free 526
  - horizontal upward walking 526
  - upstairs 526
- Spherical source 1152
- Spinodal 683
- Splay faulting 977
- SPLERT *see* System for processing local earthquakes in real time
- Squirt flow 480
- Stairs 526
- Stampede 528
- Static circular crack 373
- Static friction 684
- Static strain energy 373
- Stationary Markovian strategy (SMS) 111, 112
- Statistical control chart 1119
- Statistical models
  - for fault slip 379
- Statistical physics 683
- Statistics of earthquakes 681
- Steady state 1114
- Steam 1186
- Stella 909
- Step
  - collision 536
  - propagation 535
- Stepover 684
- Stern's behavioral explanation model 494
- Stochastic averaging of the wave equation 794
- Stochastic branching model 687
- Stochastic climate model 2
- Stochastic game 109
  - folk theorem 112
- Stochastic game model
  - equilibrium 111
  - history 110
  - set-up 110
  - state space 110
  - state variable 110
- Stochastic inverse 949
- Stochastic models 338
  - GPS measurements 340
  - instrumental catalogs 339
  - statistical seismology 339
- Stop-and-go wave 709
- Storm surge 501
- Story 496, 499
  - images tell a story 499
  - story telling bonds neighbor 499
- Strategy
  - stationary Markovian 111
- Strength profile, earthquake 334
- Strength threshold 682
- Stress 384, 972
  - and triggering, static 426
  - asymmetric 384
- Stress change
  - dynamic 425
  - static 425
- Stress concentration 376
- Stress corrosion 440
- Stress drop 373, 1212, 1213
- Stress glut 1148
- Stress intensity factor 376
- Stress release 711
- Stress-release model 346
  - conditional intensity 347
  - coupled stress release model 347
- Stress sensitivity 801
- Strike fault 1004
- Strike slip fault 680
- String-string deformation 383
- Stripe formation 703
- Stromboli 1037
- Strong ground motions 767
- Structural damage 150
- Structural damage detection 151
- Structural health monitoring 150
- Structural models 154
  - natural frequency of vibration 154
  - oscillator 154
  - wave guide 155
  - wave travel 155
- Structural system 201
- Structure 150, 212
  - excitation of 212
- Structure indices 390
  - phase shift 390
- Sub-nets 179
- Subduction 977, 978
- Subduction zone 265, 433
- Subduction zone earthquake 1207
- Subgame perfect equilibrium (SPE) 111
- Submarine landslides 75
- Sumatra earthquake 670
- Sumatra-Andaman earthquake 604
- Summer mean temperatures, observed an modeled 44
- Superlattice 395
- Surface tension 1039, 1067
- Surface wave 264, 426, 432

- Surface-wave magnitude 260, 261, 265  
 Sustainability 500, 512  
 Sustainability implementation research 485, 512  
 Sustainable development 506  
 Symmetry 694  
 Symmetry breaking 23  
 Synthetic aperture radar 81  
 Synthetic catalogs 138  
 Synthetic seismograms 974  
 System frequency 168  
 Systems  
   for processing local earthquakes in real time  
   (SPLERT) 413
- T**
- $\Delta t$ -Adaptation 779  
 Taiwan 177  
 Teleseismic magnitude methods  
   tsunami warning 413  
 Telluric current 447–449  
 Temperature 753  
   maximum temperature 761  
   temperature change 753  
 Tensile crack 1151  
 Tetrahedral grids 777  
 Thermal stress  
   global changes 44  
   mortality 45  
 Thermohaline circulation 3  
 Theta program, tsunami warning 418  
 Threat  
   internalize 485  
   internalizing 507  
   occasional 507  
   probable maximum flood 500  
   threat continuum 485  
 Three sisters volcanic center 1141  
 Threshold  
   continuum percolation 721  
 Threshold value, polydisperse fractures 726  
 Thrust earthquake 265  
 TIB *see* Tsunami information bulletin  
 Tide gauge 985, 1026  
 Tides 426  
 Time  
   pre-movement 527  
   response 527  
 Time- and slip-predictable models 346  
 Time frequency analysis 166, 171  
 Time-frequency analysis 167, 214  
 Time-predictable 1113  
 Time scales 1098  
 Time-window component 268  
 Tokachi-Oki earthquake 612  
 Tomography  
   computerized 929  
   conventional 929  
   seismic 928  
   X-ray 928  
 Topography 571  
 Tortuosity 473  
 Total scattering coefficient 789  
 Total scattering cross-section 789  
 Trace intersections density, fracture networks 726  
 Traces 726  
 Trajectory 710  
 Transfer function 161, 256  
 Transfer, heat 739  
 Transformation  
   preliminary 554  
 Transforming 582  
 Transient deformation 90  
 Transient growth 15  
 Transient pressure disturbance 1146  
 Transition zone  
   percolation 725  
 Transport 736  
   heat 736  
   mass 736  
 Transport coefficients  
   continuum percolation 719  
 Transport processes, geological 717  
 Travel time curve, infrasonic wave 666  
 Tremor 435  
 Trench 972, 974  
 Trend 1109  
 Triggering 733, 739  
   dynamic 425  
 Triggering mechanism 732  
 Truncated power law 694  
 Tsunameters 1026  
 Tsunami 74, 231, 257, 264, 273, 407, 891, 967, 1207,  
   1215  
   infrasound 663, 670  
   interaction with the coastline 671  
   Miyagi-Oki earthquake 670  
   propagation 407  
   Sumatra earthquake 670  
   tsunami wave heights 976  
 Tsunami alarms 273  
 Tsunami early warning 273, 274

- Tsunami early warning system 971, 979, 982
  - Tsunami earthquakes 265, 408, 418, 967, 984, 1028
    - model 977
    - slow character 970
  - Tsunami forecasting and warning 74
  - Tsunami generation 1023
  - Tsunami hazard 275
  - Tsunami information bulletin (TIB) 409
  - Tsunami infrasound 669, 672
  - Tsunami inundation 74
  - Tsunami magnitude 1028
  - Tsunami observations 1026
  - Tsunami propagation 1025
  - Tsunami warning 265
    - a fixed regional bulletin 409
    - an expanding watch/warning bulletin 409
    - future directions 420
    - local earthquake magnitude methods 413
    - mantle magnitude ( $M_m$ ) method 416
    - $M_{wp}$  method 415
    - rapid estimates 406
    - retroactive performance 410
    - rupture slowness estimation 418
    - teleseismic magnitude methods 413
    - theta program 418
  - Tsunami warning center operations 409
  - Tsunami warning centers 408
  - Tsunami waves 407
  - Tsunamigenesis 972, 975
  - Tsunamigenic earthquake 146, 408, 967, 1029
  - Tuning parameter 680
  - Turbulence 710
    - cluster 711
    - Eddy 710
    - rupture 711
    - stick-slip instability 711
    - vortex cascade 710
  - Turbulent flow 710
  - Turkey 178
  - Twist vector 383
  - Two-frequency mutual coherence function 795
- U**
- ULF electric signals 352
    - Cox regression model 353
    - RELM testing center 352
    - self-exciting (Hawkes type) process 352
  - Uncertainty 231
  - Under-dispersed 1111
  - Underground explosions 257
  - Underground nuclear explosion 267
  - Unemployment
    - acceleration 551
  - United States 177
  - Universal scaling function 681, 694
  - Universality 564, 680, 690
  - Universality class 685, 694
  - Update
    - parallel 533
    - synchronous 533
  - Upper critical dimension 687
  - Upper fractal limit 27
  - Urban 500
  - UrEDAS 177
  - US geological Survey (USGS) 1139
- V**
- Vacant dislocations 397
  - Validation 553
  - VAN method 448–450
    - conductive channel model 456
    - Ioannina station 450
    - seismic electric signal (SES) 449
    - seismic electric Signals (SES) 448
    - selectivity 449
    - selectivity map 450
    - SES 458
    - SES activity 449, 450, 452
    - VAN relation 449, 450
    - VAN-type monitoring 450
  - Van Nuys 7-story hotel 168
  - Van Nuys building 172
  - Vector-borne tropical infectious diseases, increase by
    - climate change 48
  - Velocity broadband record 263
  - Velocity pulse
    - earthquake 321
    - microearthquakes 328, 329
    - slow initial phas 325
  - Velocity-strengthening 1207, 1211, 1215
  - Velocity waveforms, large earthquakes 325
  - Velocity-weakening 1207, 1211
  - Vensim 909
  - Vent 1195
  - Very long baseline interferometry 81
  - Victoria 509
  - Video tracking 700
  - Viscoelastic 1040

- Viscosity 1037, 1084, 1090, 1095  
     Newtonian 1037  
     non-Newtonian 1037  
 Viscous expansion time 735  
 Viscous fingering 702  
 Visualization 136, 137, 139, 144, 145  
 Volatile 731, 1035, 1040  
 Volcanic 1035  
 Volcanic activity, eruption signature 675  
 Volcanic conduit 1088  
 Volcanic crisis 1135, 1138  
 Volcanic eruption 570  
     spectroscopy 675  
 Volcanic eruption detection 676  
 Volcanic explosivity index (VEI) 1035, 1105  
 Volcanic sound, modeling 673  
 Volcano 76  
     infrasound 663  
     Kilauea 893  
     Mount St. Helens 1084, 1092  
     Santiaguito 1092, 1093  
     Shiveluch 1096  
     Soufrière hills 1084  
 Volcano disaster assistance program (VDAP) 1140  
 Volcano hazards 1135  
 Volcano infrasound 672, 676  
     basic principle 672  
 Volcano monitoring 1135, 1142  
     baseline monitoring 1135  
 Volcano risk 1135  
 Volcano-seismic signal 1068  
 Volcano seismicity 1146  
 Volcano seismology 76, 1147  
 Volcano unrest 1135, 1137, 1139, 1142, 1143  
 Volume, excluded 717  
 von Kármán-type ACF 793  
 Voronoi cells 771  
 Vulcanian 1057  
 Vulnerability 484, 487  
     exit route 488
- W**
- W-Phase 971  
 Wandering effect 794  
 Warning 486, 1136  
     active warning image 508  
     cyclone 502  
     spreading 511  
     timeliness 508  
     traditional weather warning 509
- Warning map 510  
 Water 1180  
 Water availability 760  
 Water resource 34  
 Wave 257, 1199  
     density 529  
     longitudinal wave 257  
     shock 529  
 Wave magnitude calibration function 261  
 Wave propagation 767  
 Wave propagation in complex media: path and site  
     Effects 641  
     incoherently scattered wave energy 642  
     probabilistic seismic hazard analysis 652, 654  
     radiative transfer theory 645  
     scattering in inhomogeneous media 644  
     seismic coda 644  
     wave propagation in random media 643  
 Wave travel time analysis 171  
 Wave travel times 169  
 Waveform inversion 768, 1155  
 Wavelength 267  
     critical 266  
 Wavelet functions 160  
 Wavelet transform 162  
 Wavelets 7  
 Weak localization 794  
 Weak spring 685  
 Weather extremes  
     mortality 45  
 Wedge mechanics 75  
 Well log data 792  
 West Coast/Alaska tsunami warning center 986  
 Woodgate Beach 486, 490, 504, 508, 509  
 World meteorological organization 1001  
 World Organization of Volcano Observatories 1143  
 Worldwide standardized seismograph network  
     262, 283  
 WOVOdat 1144
- X**
- Xanthan gum 1059
- Y**
- Yellowstone 430, 431  
 Yellowstone caldera 1141  
 Yield stress 1038



**Z**

Zero temperature 682

Zigzag design 712

Zipper effect 523

Zones 575

boundary 575