Mathematics of Planet Earth Series 5

Hans G. Kaper
Fred S. Roberts
Editors

Mathematics of Planet Earth

Protecting Our Planet, Learning from the Past, Safeguarding for the Future

Springer

Mathematics of Planet Earth

Volume 5

Springer's Mathematics of Planet Earth collection provides a variety of well-written books of a variety of levels and styles, highlighting the fundamental role played by mathematics in a huge range of planetary contexts on a global scale. Climate, ecology, sustainability, public health, diseases and epidemics, management of resources and risk analysis are important elements. The mathematical sciences play a key role in these and many other processes relevant to Planet Earth, both as a fundamental discipline and as a key component of cross-disciplinary research. This creates the need, both in education and research, for books that are introductory to and abreast of these developments.

Springer's MoPE series will provide a variety of such books, including monographs, textbooks and briefs suitable for users of mathematics, mathematicians doing research in related applications, and students interested in how mathematics interacts with the world around us. The series welcomes submissions on any topic of current relevance to the international Mathematics of Planet Earth effort, and particularly encourages surveys, tutorials and shorter communications in a lively tutorial style, offering a clear exposition of broad appeal.

More information about this series at http://www.springer.com/series/13771

Hans G. Kaper • Fred S. Roberts
Editors

Mathematics of Planet Earth

Protecting Our Planet, Learning from the Past, Safeguarding for the Future

Springer

Editors
Hans G. Kaper
Mathematics and Statistics
Georgetown University
Washington, DC, USA

Fred S. Roberts
DIMACS Center
Rutgers University
Piscataway, NJ, USA

ISSN 2524-4264 ISSN 2524-4272 (electronic)
Mathematics of Planet Earth
ISBN 978-3-030-22043-3 ISBN 978-3-030-22044-0 (eBook)
https://doi.org/10.1007/978-3-030-22044-0

Mathematics Subject Classification: 60Gxx, 62–07, 62Pxx, 86Axx, 90Bxx, 91Dxx, 92Bxx, 92Dxx

This Springer imprint is published by the registered company Springer Nature Switzerland AG.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword

This book is an invitation. Since the winter of 2009, when I first had the idea of *Mathematics of Planet Earth*, it has become a passion for me to learn how the mathematical sciences can help us understand our planet, its ecosystems, and its organization, and how we as pure and applied mathematicians can contribute to protecting our planet from the effects of climate change, extreme events, and other risk factors. For a pure mathematician like me, it is not always easy to make one's way through the relevant literature, which—too often—is either targeting specialists or too elementary on the mathematical side. This book fills the gap. It is an invitation to our professional community to explore the new challenges for mathematics related to planet Earth and, at the same time, enrich the cultural heritage of science on our planet. The subjects—not so standard and very diverse—are likely to pique everyone's interest, as they did for me. Did you know that the surface of the Earth is *not so solid* and reacts to the sliding of glaciers? Or that the sea-level rise can vary substantially from one region of the globe to another? The book addresses these and many other interesting questions, like how to measure biodiversity and what mathematics can say about the sixth mass extinction, which is being driven by current human behavior. Other chapters are focused on how to optimize the long-term human use of natural capital or how to plan for infrastructure restoration after an extreme event. The reader is introduced to the mathematics of food systems and food security—new topics that are likely to become paramount as the world population experiences the limits of sustainable development. The subject of infectious diseases is treated with new examples, which can bring new ideas into a modeling course or motivate modeling projects for the students. I hope that this book will draw you in, as it did me.

Montreal, QC, Cananda — Christiane Rousseau
November 2018

Preface

Planet Earth offers a wealth of challenges to science. The planet is at risk, and as scientists, we have a responsibility to get involved and apply the particular skills of our discipline to face these challenges.

Mathematics of Planet Earth is a concerted effort to identify these challenges and bring them to the attention of the mathematical sciences research community. Conceived by Christiane Rousseau (Université de Montréal, Canada) in 2009, the first manifestation of this effort was a yearlong program, *Mathematics of Planet Earth 2013*. MPE2013 started out as a grassroots organization, which grew quickly into an international partnership of more than 150 scientific societies, universities, research institutes, and organizations. It brought the challenges facing our planet to the attention of the mathematics research community and organized many outreach activities to show the public how mathematics contributes to our understanding of planet Earth, the nature of the challenges our planet is facing, and how mathematical scientists address these challenges. It underscored the multidisciplinary nature of the problems facing the planet and emphasized multidisciplinary partnerships to address these problems. An anthology of blogs posted during MPE2013 was published in 2015 by the Society for Industrial and Applied Mathematics (SIAM).[1]

At the end of 2013, MPE2013 morphed into *Mathematics of Planet Earth* (MPE). A new structure was designed to support the ongoing research efforts and maintain the momentum created by MPE2013. A program of technical and educational workshops, MPE2013+ (supported by DIMACS at Rutgers University[2] and by the National Science Foundation, grant DMS-1246305), was instrumental in furthering the goals of MPE. So was the formation of a SIAM Activity Group on Mathematics of Planet Earth (SIAG/MPE).[3] The editors also recognize the initiative by Springer Verlag to initiate a book series on the themes of MPE.

[1] http://bookstore.siam.org/ot140.

[2] http://dimacs.rutgers.edu/archive/SpecialYears/2013_MPE/.

[3] https://www.siam.org/membership/Activity-Groups/detail/.

From the beginning, MPE has been interpreted in the broadest possible sense. While the geosciences have long been concerned with our planet as a physical system, there is a growing awareness of the human impact on the ecosystem and a gradual realization that natural resources are not infinite. Sustainability has become a concern, and risks, both social and economic, are receiving increased attention. These themes define the core activities of MPE.

The purpose of this book is to introduce challenging problems in MPE-related topic areas to the mathematical sciences research community, to demonstrate the application of a wide range of mathematical ideas to these challenges, and to raise awareness in the application disciplines that the mathematical sciences offer novel opportunities for quantitative and qualitative analysis. The book is potentially of interest to scientists in academia, the private sector, government, and nonprofit organizations active in application areas, such as the geophysical sciences, climate science, ecology, environmental science, public health, and socioeconomics.

The book covers some but by no means all topics of interest to MPE; it is meant to give a flavor of selected topics of current interest. As Professor Rousseau states in her Foreword, it is an invitation to explore new challenges. Among the topics covered are climate change, the spread of infectious diseases, multistability of ecosystems, biodiversity, infrastructure restoration after extreme events, urban environments and the Internet of Things, food security, and food safety. These topics illustrate the wide range of challenges for mathematical modeling. They also highlight the variety of mathematical techniques brought to bear on these challenges, from differential equations and dynamical systems theory, optimization, statistics, operations research, discrete mathematics, graph theory, and data analytics.

The prerequisite mathematics for the various chapters varies, but much of the material should be accessible to advanced undergraduate and graduate students. Selected chapters can be used as a text for seminars or self-study. Application scientists (including graduate students) and decision-makers with background knowledge in one or more of the mathematical topics listed in the previous paragraph will find a wealth of tools that they may wish to explore for practical purposes.

The chapters in this book were solicited from a diverse group of experts. Each chapter of the book was peer-reviewed. The editors worked with the authors to revise their chapters and to put them all into a common language and approach. The editors thank the (anonymous) reviewers for their extensive efforts to improve the quality of the presentations.

We hope that this volume will stimulate the readers to explore the challenges of MPE and apply the tools of the mathematical sciences to solve the problems of our planet.

Acknowledgments The editors express their appreciation for many enlightening discussions (technical and otherwise) with colleagues in the *Mathematics and Climate Research Network* (funded by the National Science Foundation, grant DMS-0940363). Fred Roberts acknowledges the support from the NSF through grant DMS-1246305.

Washington, DC, USA — Hans G. Kaper
Piscataway, NJ, USA — Fred S. Roberts
November 2018

Road Map

Mathematics of Planet Earth (MPE) views our planet through multiple lenses. Broadly speaking, it considers our planet as a physical system, as a system supporting life, as a system organized by humans, and as a system at risk. Each lens provides a different perspective and may require different tools and techniques from the mathematical sciences. But the common goal is to gain a better understanding of how the state of our planet influences and is influenced by human activities.

The chapters in this volume are grouped into four parts. The following is a brief introduction to each part and an overview of the chapters selected.

Part I: Geo- and Physical Sciences

In Part I, we consider planet Earth as a *physical system*. Here, the focus is on Earth's climate system, the physical processes that occur in the various components of the system (atmosphere, oceans, etc.), their dynamics, and mutual interactions. We have selected three case studies of physical systems which present typical challenges for mathematical and statistical modeling and analysis.

Chapter 1 discusses a conceptual model of the coupled atmosphere-ocean system which emphasizes the critical role of atmospheric carbon dioxide (CO_2) in the dynamics of glacial-interglacial cycles during the Pleistocene Epoch. The chapter also offers an interesting application of techniques from the theory of dynamical systems.

Chapter 2 presents a closely related problem from geophysics, namely, the Glacial Isostatic Adjustment problem—that is, the adjustment of the Earth's surface to a varying load due to the waxing and waning of ice sheets during a glacial-interglacial cycle.

Chapter 3 addresses a practical problem, namely, how to measure precipitation. Unlike temperature, precipitation is highly variable, both in space and in time. The chapter also describes various statistical methods to identify precipitation

patterns from data—a highly relevant problem, for example, for flood control and reservoir management, especially as precipitation patterns are changing due to climate change.

Part II: Life Sciences

In Part II and the following Part III, we consider planet Earth as a *system supporting life*. Topics of interest are many, including population dynamics, epidemiology, invasive species, the carbon cycle, natural resources, issues of sustainability and equity, mathematical ecology, evolution, and effects of climate change on living organisms. In Part II, we have collected three case studies in epidemiology; in Part III, we will turn our attention to ecology and evolution.

Environmental conditions have always been of profound importance in shaping the epidemiology of infectious diseases. Malaria is a good example; it is caused by plasmodium parasites and spread by the Anopheles mosquito, and the life cycle of both depends sensitively on temperature. Thus, global warming may shift and/or expand the geographic range of the disease. This topic is addressed in Chap. 4 with geographic focus on Africa, where the burden of malaria is greatest.

Chapter 5 is concerned with Buruli ulcer (BU), another infectious disease prevalent in Africa, especially Ghana. The disease is spread by a pathogen which shares its environment with humans. The mathematical model also accounts for the fact that not all individuals are equally susceptible to infection. The chapter illustrates the importance of data for model validation.

Chapter 6 presents three case studies from health science. They demonstrate some of the challenges a modeler is likely to encounter in attempting to develop and test a model that reflects real data. The first case is a source-attribution problem for food-related salmonellosis: identify possible pathways from food source to infected individual from a historical dataset listing human cases and Salmonella sources. Which food source(s) contribute the most infections? A second source-attribution problem arises for highly pathogenic avian influenza (HPAI), a communicable veterinary disease that affects poultry. Data on cell turnover rates are available but do not fit any of the standard population models for immune cells. The challenge is met with data-driven statistical modeling and bona fide detective work with genetic evidence.

Part III: Ecology and Evolution

This part continues the study of planet Earth as a *system supporting life*. Here, the emphasis is on mathematical ecology, evolution, and biodiversity.

Ecosystems are highly nonlinear dissipative systems characterized by multiple equilibria, some stable and some unstable. As the likelihood of extreme events

(floods, wildfires, storms, etc.) increases due to environmental change, an ecosystem may experience abrupt transitions from one stable state to another which may be desirable or undesirable. If the new state is undesirable, the effect can be disastrous; for example, the transition from a vegetated state to a desert state due to an extended period of drought is clearly undesirable. Chapter 7 gives examples of dry-land ecosystems which show distinct vegetation patterns on a path toward desertification due to different soil-water feedback mechanisms.

The health of an ecosystem is often judged by its biodiversity. But how do we measure biodiversity? The question is fundamental for a proper assessment of different intervention scenarios and progress toward a healthy ecosystem. Traditional attempts to measure biodiversity consider two components: richness (the number of species in the ecosystem) and evenness (the extent to which species are evenly distributed). Chapter 8 describes attempts to define richness and evenness mathematically in an effort to make the concept of biodiversity more precise.

Chapter 9 addresses the problem of estimating extinction risks across many levels ranging from an entire ecosystem or population to a single species. The chapter surveys current deterministic and stochastic methods of analysis and extends existing theory in two directions by considering the possibility of evolutionary rescue from extinction in a changing environment and the posthumous assignment of an extinction date from sighting records.

Part IV: Socioeconomics and Infrastructure

In Part IV, we focus on planet Earth both as a *system organized by humans* and a *system at risk*. Under the first theme, the focus is on infrastructure, ecosystem services, socioeconomics, social organization, public health, and rules and regulations. The second theme covers extreme events, risk assessment and risk management, emergency planning, and strategies for mitigation and adaptation. Again, too many topics to cover in a single volume, let alone a single part of a collection like the one facing us.

For this part, we have selected four topics which touch on both themes, namely, food systems and food security, ecosystem services and natural capital, infrastructure restoration after an extreme event, and infrastructure management enabled by the Internet of Things.

Chapter 10 introduces the topic of food systems and food security. The motivation is clear: enough food is being produced to provide enough nutrients for every person on Earth, but sizable fractions of the population suffer from malnutrition or are overweight. Mathematical models may provide insight into the design of a food system that improves food security for all.

Natural resources and natural environments often provide benefit flows beyond the net revenues generated by their harvest, extraction, or visits by the users. Chapter 11 argues that dynamic optimization is the tool of choice to determine the value of natural capital and ecosystem services. A model of a renewable resource—

the stock of oysters in northern Chesapeake Bay—illustrates the case. The oyster population provides not only net revenue to watermen (harvesters) but also an ecosystem service in the form of improved water quality through the removal of nutrients.

Chapter 12 focuses on the recovery of critical infrastructure systems from large-scale disruptive events. This problem is especially difficult when different infrastructure systems depend on each other. (For example, you cannot pump gas without power.) Decision-makers need tools to determine optimal orders for restoration of such interdependent infrastructure systems and guide the restoration process. This chapter shows how an infrastructure system can be modeled as a network flow problem and how optimization can help decision-makers to assess the impact of the disruption on the services provided by the system.

Chapter 13, the final chapter, takes us to the future, when it becomes feasible to perform online estimation, optimization, and control by leveraging the data collected through the Internet of Things. The chapter gives two particular applications that are important to effectively manage a city: transportation and municipal water services.

Contents

Contributors

Ash Abebe Department of Mathematics and Statistics, Auburn University, Auburn, AL, USA

Ernesto Arandia IBM Research, Dublin, Ireland

Sebastien Blandin IBM Research, Singapore, Singapore

Kevin R. Burgio Department of Ecology & Evolutionary Biology, University of Connecticut, Storrs, CT, USA

Colin J. Carlson Department of Environmental Science, Policy and Management, University of California Berkeley, Berkeley, CA, USA

Jon M. Conrad Dyson School of Applied Economics and Management, Cornell University, Ithaca, NY, USA

Tad A. Dallas Department of Environmental Science and Policy, University of California Davis, Davis, CA, USA

Bradley J. Eck IBM Research, Dublin, Ireland

Christina Edholm University of Tennessee, Knoxville, TN, USA

Department of Mathematics, Scripps College, Claremont, CA, USA

Steffen E. Eikenberry School of Mathematical and Statistical Sciences, Arizona State University, Tempe, AZ, USA

Maria Emelianenko Department of Mathematical Sciences, George Mason University, Fairfax, VA, USA

Hans Engler Mathematics and Statistics, Georgetown University, Washington, DC, USA

Wayne M. Getz Department of Environmental Science, Policy and Management, University of California Berkeley, Berkeley, CA, USA

Abba B. Gumel School of Mathematical and Statistical Sciences, Arizona State University, Tempe, AZ, USA

Hans G. Kaper Mathematics and Statistics, Georgetown University, Washington, DC, USA

Tasso J. Kaper Department of Mathematics and Statistics, Boston University, Boston, MA, USA

Scott D. King Department of Geosciences, Virginia Tech, Blacksburg, VA, USA

Scott Le Fevre Norwich University, Northfield, VT, USA

Suzanne Lenhart University of Tennessee, Knoxville, TN, USA

Department of Mathematics, Scripps College, Claremont, CA, USA

Benjamin Levy Department of Mathematics, Fitchburg State University, Fitchburg, MA, USA

Viviana Maggioni Sid and Reva Dewberry Department of Civil, Environmental, and Infrastructure Engineering, George Mason University, Fairfax, VA, USA

Theresia Marijani Department of Mathematics, University of Dar es Salaam, Dar es Salaam, Tanzania

Yair Mau Department of Soil and Water Sciences, Robert H. Smith Faculty of Agriculture, Food and Environment, The Hebrew University of Jerusalem, Rehovot, Israel

Sean A. McKenna IBM Research, Dublin, Ireland

Ehud Meron Blaustein Institutes for Desert Research and Physics Department, Ben-Gurion University of the Negev, Beersheba, Israel

Farai Nyabadza Department of Mathematics, University of Stellenbosch, Stellenbosch, South Africa

Sarah G. Nurre Pinkley Department of Industrial Engineering, University of Arkansas, Fayetteville, AR, USA

Fred S. Roberts DIMACS Center, Rutgers University, Piscataway, NJ, USA

Christiane Rousseau Département de mathématiques et de statistique. Université de Montréal, Montréal, QC, Canada

Thomas C. Sharkey Department of Industrial and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY, USA

Theodore Vo School of Mathematics, Monash University, Clayton, Victoria, Australia

Laura Wynter IBM Research, Singapore, Singapore

Abdul-Aziz Yakubu Department of Mathematics, Howard University, Washington, DC, USA

Antonios Zagaris Department of Bacteriology and Epidemiology, Wageningen Bioveterinary Research, Wageningen University and Research, Lelystad, The Netherlands

Yuval R. Zelnik Centre for Biodiversity Theory and Modelling, Theoretical and Experimental Ecology Station, CNRS, Moulis, France

Part I
Geo- and Physical Sciences

Chapter 1
Modeling the Dynamics of Glacial Cycles

Hans Engler, Hans G. Kaper, Tasso J. Kaper, and Theodore Vo

Abstract This chapter is concerned with the dynamics of glacial cycles observed in the geological record of the Pleistocene Epoch. It focuses on a conceptual model proposed by Maasch and Saltzman (J Geophys Res 95(D2):1955–1963, 1990), which is based on physical arguments and emphasizes the role of atmospheric CO_2 in the generation and persistence of periodic orbits (limit cycles). The model consists of three ordinary differential equations with four parameters for the anomalies of the total global ice mass, the atmospheric CO_2 concentration, and the volume of the North Atlantic Deep Water. In this chapter, it is shown that a simplified two-dimensional symmetric version displays many of the essential features of the full model, including equilibrium states, limit cycles, their basic bifurcations, and a Bogdanov–Takens point that serves as an organizing center for the local and global dynamics. Also, symmetry breaking splits the Bogdanov–Takens point into two, with different local dynamics in their neighborhoods.

Keywords Bifurcation analysis · Bogdanov–Takens unfolding · Conceptual model · Glacial cycles · Maasch–Saltzman model · Pleistocene climate

1.1 Introduction

Earth's climate during the *Pleistocene Epoch*—the geological period from approximately 2.6 million years before present (2.6 Myr BP) until approximately 11.7 thousand years before present (11.7 Kyr BP)—is of great interest in the

H. Engler · H. G. Kaper (✉)
Mathematics and Statistics, Georgetown University, Washington, DC, USA
e-mail: engler@georgetown.edu; hans.kaper@georgetown.edu

T. J. Kaper
Department of Mathematics and Statistics, Boston University, Boston, MA, USA
e-mail: tasso@math.bu.edu

T. Vo
School of Mathematics, Monash University, Clayton, Victoria, Australia
e-mail: theodore.vo@monash.edu

H. G. Kaper, F. S. Roberts (eds.), *Mathematics of Planet Earth*, Mathematics of Planet Earth 5, https://doi.org/10.1007/978-3-030-22044-0_1

geosciences community. The geological record in the Northern Hemisphere gives evidence of cycles of advancing and retreating continental glaciers and ice sheets, mostly at high latitudes and high altitudes.

To reconstruct the Pleistocene climate, geoscientists rely on geological proxies, particularly a dimensionless quantity denoted by δ^{18}O, which is measured in parts per mille. This quantity measures the deviation of the ratio ^{18}O/^{16}O of the stable oxygen isotopes ^{18}O and ^{16}O in a given marine sediment sample from the same ratio in a universally accepted standard sample. The relative amount of the isotope ^{18}O in ocean water is known to be higher at tropical latitudes than near the poles, since water with the heavier oxygen isotope is slightly less likely to evaporate and more likely to precipitate first. Similarly, water with the lighter isotope ^{16}O is more likely to be found in ice sheets and in rain water at high latitudes, since it is favored in atmospheric transport across latitudes. The global distribution of δ^{18}O in ocean water therefore varies in a known way between glacial and interglacial periods. A record of these variations is preserved in the calcium carbonate shells of foraminifera, a class of common single cell marine organisms. These fossil records may be sampled from deep sea sediment cores, and their age and δ^{18}O may be determined. Details are described in [34].

Figure 1.1 shows the LR04 time series of δ^{18}O over the past 5.3 million years, reconstructed from sediment core data collected at 57 geographically distributed sites around the globe [34]. As the observed isotope variations are similar in shape to the temperature variations reconstructed from ice core data for the past 420 Kyr at the Vostok Station in Antarctica, the values of δ^{18}O (right scale) have been aligned with the reported temperature variations from the Vostok ice core (left scale) [43]. The graph shows a relatively stable temperature during the period preceding the Pleistocene and increasing variability during the Pleistocene.

The typical pattern throughout most of the Pleistocene resembles that of a sawtooth wave, where a slow glaciation is followed by a rapid deglaciation. In the early Pleistocene, until approximately 1.2 Myr BP, the period of a glacial cycle averages 41 Kyr; after the mid-Pleistocene transition, which occurred from

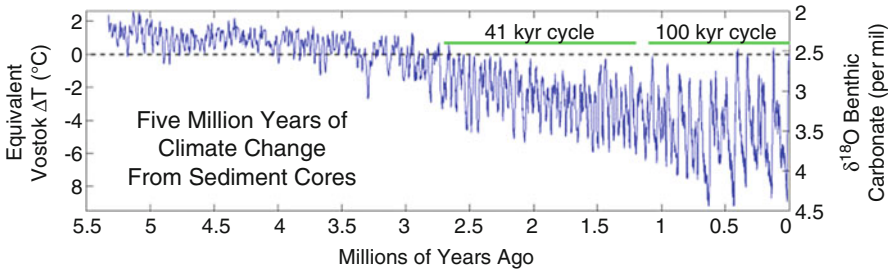


Fig. 1.1 Time series of the isotope ratio $\delta^{18}\text{O} = {}^{18}\text{O}/{}^{16}\text{O}$ (scale on the right, in parts per mille) for the past 5.3 million years [34]. The scale on the left gives the equivalent temperature anomaly (in degrees Celsius). [Source: Wikipedia, Marine Isotope Stage]

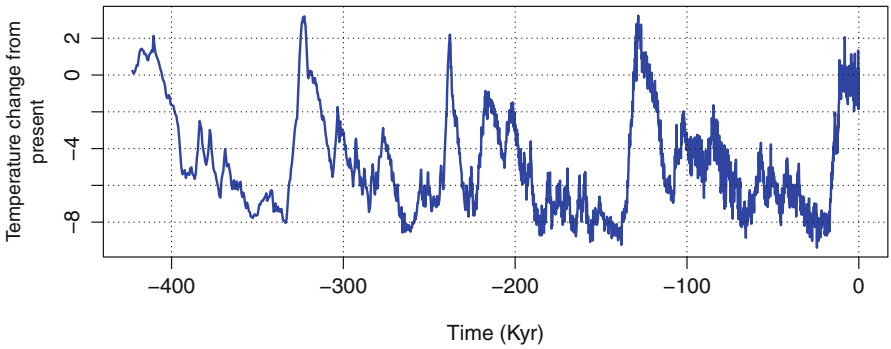


Fig. 1.2 Time series of the global mean temperature for the past 420,000 years. [Data from Carbon Dioxide Information Analysis Center (CDIAC) at Oak Ridge National Laboratory (http://cdiac.ornl.gov)]

approximately 1.2 Myr BP until approximately 0.8 Myr BP, the glacial cycles of the late Pleistocene have a noticeably greater amplitude, and their period averages 100 Kyr. Figure 1.2 shows the global mean temperature for the past 420 Kyr, reconstructed from Vostok ice core data. The 100 Kyr cycle and the sawtooth pattern are clearly visible.

These observations suggest a number of questions for climate science. What caused the glacial oscillations during the Pleistocene? Why were the periods of the glacial cycles during the early and late Pleistocene different? What could possibly have caused the transition from 41 Kyr cycles to 100 Kyr cycles during the mid-Pleistocene?

In this chapter, we discuss a conceptual model of the Pleistocene climate proposed by Maasch and Saltzman in [35] to explain the phenomenon of glacial cycles. The model is conceptual, in the sense that it describes the state of the climate in a few variables, ignoring most of the processes that go into a complete climate model, but still captures the essence of the phenomenon. It is based on sound physical principles and, as we will see, makes for an interesting application of dynamical systems theory.

The numerical continuation results for the bifurcation curves reported in this chapter were obtained using the software package AUTO [17]; see also [15, 16]. Some recent texts on issues of climate dynamics are [10, 14, 28, 36, 52].

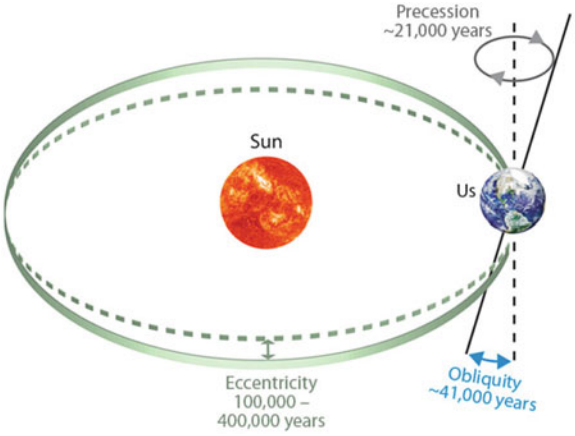
Outline of the Chapter In Sect. 1.2, we present background information to motivate the particular choices underlying the Maasch–Saltzman model. In Sect. 1.3, we derive the Maasch–Saltzman model from physical principles and formulate it as a dynamical system in a three-dimensional state space with four parameters. In Sect. 1.4, we introduce two simplifications that render the Maasch–Saltzman model symmetric and reduce it to a two-dimensional dynamical system with two parameters that can be analyzed rigorously and completely. In Sect. 1.5, we introduce asymmetry into the simplified two-dimensional model and show the effects of symmetry breaking. In the final Sect. 1.6, we summarize our results.

1.2 Background

There is general agreement that the periodicity of the glacial cycles is related to variations in the Earth's orbital parameters [33]. To first order, Earth's climate is driven by the Sun. The Earth receives energy from the Sun in the form of ultraviolet (short wavelength) radiation. This energy is redistributed around the globe and eventually reemitted into space in the form of infrared (long wavelength) outgoing radiation. The amount of energy reaching the top of the atmosphere per unit area and per unit time is known as the *insolation* (*in*cident *sol*ar radi*ation*), which is measured in watts per square meter.

1.2.1 Orbital Forcing

The insolation varies with the distance from the Earth to the Sun and thus depends on Earth's orbit around the Sun. This is the basis of the Milankovitch theory of *orbital forcing* [28, 38].



The Earth moves around the Sun in an elliptical orbit; its *eccentricity* varies with time but has dominant frequencies at approximately 100 Kyr and 400 Kyr. As the Earth moves around the Sun, it rotates around its axis. The axis is tilted with respect to the normal to the orbital plane; the tilt, known as *obliquity*, is also close to periodic, with a dominant frequency of approximately 41 Kyr. (This tilt is the main cause of the seasonal variation of our climate.)

In addition, the Earth is like a spinning top wobbling around its axis of rotation. This component of the Earth's orbit is called *precession*; its period varies from 19 to 23 Kyr.

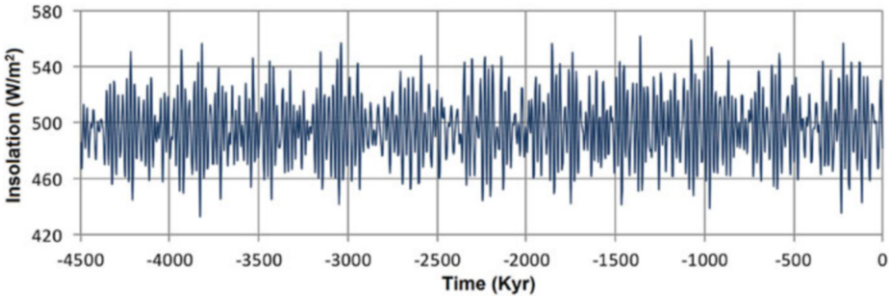


Fig. 1.3 Time series of Q^{65} during the month of July for the past 4.5 million years. [Data from [32]]

Given the three orbital parameters (eccentricity, obliquity, precession), one can compute the insolation at any latitude and at any time of the year. An example is given in Fig. 1.3, which shows the time series of Q^{65}—the average insolation at 65° North—during the month of July for the past 4.5 million years; other months show a similar behavior. A cycle with a period of approximately 400 Kyr is clearly visible. A spectral analysis reveals a dominant frequency around 21 Kyr coming from two clustered spikes in the power spectrum and another, smaller frequency component at approximately 41 Kyr.

1.2.2 Atmospheric Carbon Dioxide

The Pleistocene climate and, in particular, the mid-Pleistocene transition are topics of great interest in the geosciences community. The 41 Kyr glacial cycles of the early Pleistocene are commonly attributed to the 41 Kyr cycle of Earth's obliquity; see, for example [24, 45]. In contrast, there is less agreement on the origin of the 100 Kyr cycles of the late Pleistocene.

Some authors [20, 22, 27] attribute the 100 Kyr cycles to the eccentricity of Earth's orbit. However, simple energy balance considerations imply that variations in eccentricity are too weak to explain the surface temperature variations that are observed in the paleoclimate record. The Earth's eccentricity varies between approximately 0.01 and 0.05. At times of maximum eccentricity, the semi-major and semi-minor axes of the Earth's orbit therefore never differ by more than approximately 0.1% from the corresponding values at times of minimum eccentricity. Then the minimum solar constant at times of high eccentricity does not differ by more than 0.2% from the solar constant when eccentricity is low, implying an equilibrium temperature variation by approximately 0.05%, or less than 0.2 K [28, Chapter 2].

A possible way for orbital effects to influence the Earth's surface temperature is suggested by the greenhouse effect and the almost perfect correlation between fluctuations in the atmospheric CO_2 concentration and the surface temperature that is observed, for example, in the Vostok ice core data [43]; see Fig. 1.4.

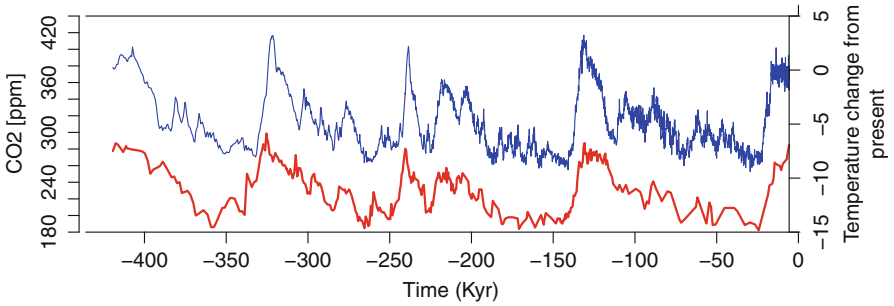


Fig. 1.4 Correlation of global mean temperature (blue) and atmospheric CO_2 concentration (red) for the past 420 Kyr. [Data from Carbon Dioxide Information Analysis Center (CDIAC) at Oak Ridge National Laboratory (http://cdiac.ornl.gov)]

Orbital variations have only a very weak effect on the composition of a planet's atmosphere. Moreover, they would affect not only CO_2, but other atmospheric components as well, and their effect would be negligible on the time scale of the glacial cycles. Hence, it is unlikely that orbital variations alone are the direct cause of the fluctuations in atmospheric CO_2 concentrations. On the other hand, a feedback mechanism connecting and reinforcing the influence of orbital forcing on surface temperatures and changes in atmospheric CO_2 appears to be plausible [47]. Changes in the carbon cycle and the climate system would then amplify each other to produce the glacial cycles, and atmospheric CO_2 would have to play a central role in such a feedback mechanism.

Saltzman was the first to propose a conceptual climate model that highlighted the role of atmospheric CO_2 in the dynamics of glacial cycles [48]. The model was further developed in joint work with Maasch in a series of chapters [35, 49, 50]. In this chapter, we focus on the model proposed by Maasch and Saltzman in [35].

1.2.3 Other Models

There certainly is no unique way to explain the phenomenon of glacial cycles and the Pleistocene climate in a comprehensive manner. Other conceptual models can be found, for example, in [3, 4, 9, 20, 26, 39–42, 46, 51, 54]. An interesting case was made by Huybers [26], who argued that the reconstruction of the temperature record from proxy data presented in Fig. 1.1 relies on orbital assumptions and is therefore subject to bias. Huybers developed an unbiased age model which does not rely on orbital assumptions and showed that the late Pleistocene glacial terminations are paced by changes in Earth's obliquity [25]. This theory would imply that the entire Pleistocene climate regime can be explained by obliquity alone. We refer the reader to [11] for a summary of the current state of the art.

As a final note, we caution that any model is a mathematical construct, and any phenomenon that results from its analysis is merely a manifestation of the

assumptions underlying the model. The question whether the model reflects the true cause(s) of the glacial cycles lies outside the domain of mathematics.

1.3 The Maasch and Saltzman Model

The Pleistocene climate model proposed by Maasch and Saltzman [35] involves five *state variables*. They are, with their associated units in square brackets:

- the total global ice mass, I [kg],
- the atmospheric CO_2 concentration, μ [ppm],
- the volume of the North Atlantic Deep Water (NADW), N [m^3],
- the global mean sea surface temperature (SST), τ [K], and
- the mean volume of permanent (summer) sea ice, η [m^3].

The volume of the NADW is a measure of the strength of the global oceanic circulation (*thermohaline circulation*, THC). We can think of N as a measure of the strength of the oceanic CO_2 pump, since the oceanic CO_2 pump is an integral part of the THC. The other variables are self-explanatory.

The state variables vary with time, albeit on rather different time scales. The total global ice mass, atmospheric CO_2 concentration, and NADW vary on the order of thousands of years, while the SST and summer sea ice vary on the order of decades or centuries. Here the focus is on the slow time scale, where we assume that the fast variables equilibrate essentially instantaneously. That is, the long-term dynamics of the climate system are described in terms of I, μ, and N (the *prognostic variables*); τ and η are *diagnostic variables*, which follow the prognostic variables in time.

1.3.1 Model Formulation

The climate model is formulated in terms of *anomalies*—deviations from long-term averages, which are indicated with a prime. The governing equations follow from plausible physical principles, which are detailed in [49, §2].

The global mean SST (τ) and the mean volume of permanent sea ice (η) vary with the total global ice mass (I) and the atmospheric CO_2 concentration (μ) but are independent of the NADW (N); in particular, τ decreases as I increases or μ decreases, while η increases as I increases or μ decreases. To leading order, the dependences are linear, so

$$\begin{aligned} \tau' &= -\alpha I' + \beta \mu', \\ \eta' &= e_I I' - e_\mu \mu'. \end{aligned} \tag{1.3.1}$$

In the absence of external forces, the governing equations for I', μ', and N' are

$$\begin{aligned}
\frac{dI'}{dt'} &= -s_1\tau' - s_2\mu' + s_3\eta' - s_4 I', \\
\frac{d\mu'}{dt'} &= r_1\tau' - r_2\eta' - (r_3 - b_3 N')N' - (r_4 + b_4 N'^2)\mu' - r_5 I', \\
\frac{dN'}{dt'} &= -c_0 I' - c_2 N'.
\end{aligned} \tag{1.3.2}$$

Time t' is measured in units of 1 year [yr]. The coefficients in these equations are positive (or zero). Maasch and Saltzman also included an external forcing term related to Q^{65} in the equation for I', but since we are interested in the internal dynamics of the system, we don't include external forcing.

The physical assumptions underlying these equations are:

- The prognostic variables relax to their respective long-term averages, so their anomalies tend to zero as time increases; in particular, I' and N' decay at a constant rate, while the decay rate of μ' increases quadratically with N' [49, §2.V].
- If the SST exceeds its mean value ($\tau' > 0$), the total global ice mass decreases and the atmospheric CO_2 concentration increases (due to outgassing); if the SST is less than its mean value ($\tau' < 0$), the opposite happens. The coupling is linear to leading order.
- Since CO_2 is a greenhouse gas, an increase in the atmospheric CO_2 concentration leads to a warmer climate and thus a decrease in the total global ice mass.
- If the volume of permanent sea ice exceeds its mean value ($\eta' > 0$), the total global ice mass increases and the atmospheric CO_2 concentration decreases; if the volume of permanent sea ice is less than its mean value ($\eta' < 0$), the opposite effect happens. The coupling is linear to leading order.
- A greater-than-average total global ice mass ($I' > 0$) negatively affects both the atmospheric CO_2 concentration and the strength of the North Atlantic overturning circulation; a less-than-average total global ice mass ($I' < 0$) has the opposite effect. The coupling is linear to leading order.
- The atmospheric CO_2 concentration decreases as the strength of the North Atlantic overturning circulation increases, but the coupling weakens (strengthens) as the strength of the NADW is above (below) average [49, §2.III a,b].

Upon substitution of the expressions (1.3.1), the governing equations (1.3.2) become

$$\begin{aligned}
\frac{dI'}{dt'} &= -a_0 I' - a_1\mu', \\
\frac{d\mu'}{dt'} &= -b_0 I' + (b_1 - b_4 N'^2)\mu' - (b_2 - b_3 N')N', \\
\frac{dN'}{dt'} &= -c_0 I' - c_2 N',
\end{aligned} \tag{1.3.3}$$

where

$$a_0 = s_4 - (\alpha s_1 + e_I s_3), \quad a_1 = s_2 + \beta s_1 + e_\mu s_3,$$
$$b_0 = r_5 + \alpha r_1 + e_I r_2, \quad b_1 = \beta r_1 + e_\mu r_2 - r_4, \quad b_2 = r_3.$$

Following [35, 49], we take $b_0 = 0$ and assume that the remaining coefficients are all positive. Note that a_0 and b_1 involve positive as well as negative contributions, so the implicit assumption is that the positive contributions dominate.

1.3.2 Dimensionless Form

Next, we reformulate the system of Eq. (1.3.3) by rescaling time,

$$t = a_0 t', \tag{1.3.4}$$

and introduce dimensionless variables,

$$X = \frac{I'}{\hat{I}}, \quad Y = \frac{\mu'}{\hat{\mu}}, \quad Z = \frac{N'}{\hat{N}}, \tag{1.3.5}$$

where $\hat{I}$, $\hat{\mu}$, and $\hat{N}$ are reference values of I, μ, and N, respectively. Since $a_0 \approx 1.00 \cdot 10^{-4}\,\mathrm{yr}^{-1}$, a unit of t corresponds to (approximately) 10 Kyr.

The governing equations for X, Y, and Z are

$$\begin{aligned} \dot{X} &= -X - \hat{a}_1 Y, \\ \dot{Y} &= (\hat{b}_1 - \hat{b}_4 Z^2) Y - (\hat{b}_2 - \hat{b}_3 Z) Z, \\ \dot{Z} &= -\hat{c}_0 X - \hat{c}_2 Z, \end{aligned} \tag{1.3.6}$$

where the dot $\dot{}$ indicates differentiation with respect to t. Recall that we have set $b_0 = 0$. The remaining coefficients are dimensionless combinations of the physical parameters in the system of Eq. (1.3.3),

$$\hat{a}_1 = \frac{a_1 \hat{\mu}}{a_0 \hat{I}}, \ \hat{b}_1 = \frac{b_1}{a_0}, \ \hat{b}_2 = \frac{b_2 \hat{N}}{a_0 \hat{\mu}}, \ \hat{b}_3 = \frac{b_3 \hat{N}^2}{a_0 \hat{\mu}}, \ \hat{b}_4 = \frac{b_4 \hat{N}^2}{a_0}, \ \hat{c}_0 = \frac{c_0 \hat{I}}{a_0 \hat{N}}, \ \hat{c}_2 = \frac{c_2}{a_0}.$$

A rescaling of the variables X, Y, and Z,

$$x = \left((\hat{c}_0/\hat{c}_2)\sqrt{\hat{b}_4}\right) X, \quad y = \left(\hat{a}_1 (\hat{c}_0/\hat{c}_2)\sqrt{\hat{b}_4}\right) Y, \quad z = \left(\sqrt{\hat{b}_4}\right) Z, \tag{1.3.7}$$

leads to the following dynamical system for the triple (x, y, z):

$$\begin{aligned} \dot{x} &= -x - y, \\ \dot{y} &= ry - pz + sz^2 - yz^2, \\ \dot{z} &= -qx - qz. \end{aligned} \tag{1.3.8}$$

The coefficients p, q, r, and s are combinations of the physical parameters,

$$p = \frac{\hat{a}_1 \hat{b}_2 \hat{c}_0}{\hat{c}_2}, \quad q = \hat{c}_2, \quad r = \hat{b}_1, \quad s = \frac{\hat{a}_1 \hat{b}_3 \hat{c}_0}{\hat{c}_2 \sqrt{\hat{b}_4}}. \tag{1.3.9}$$

The coefficients are assumed to be positive, with $q > 1$. The system of Eq. (1.3.8) is the model proposed by Maasch and Saltzman in [35, Eqs. (4)–(6)]. Note that the model considered here does not include external forcing, so it describes the *internal dynamics* of the climate system.

1.3.3 Discussion

The system of Eq. (1.3.8) is what is known as a *conceptual model*. Its derivation involves physical arguments, but there is no guarantee that it corresponds to what actually happened in the climate system during the Pleistocene. Its sole purpose is to describe a possible mechanism that explains the observed behavior of the glacial cycles.

Loosely speaking, we identify x, y, and z with the anomalies of the total amount of ice, the atmospheric CO_2 concentration, and the volume of the NADW (the strength of the oceanic CO_2 pump), respectively. Time is normalized and expressed in units of the characteristic time of the total global ice mass, typically of the order of 10 Kyr.

Because of the various transformations needed to get from the physical system (1.3.3) to the dynamical system (1.3.8), it is difficult to relate the parameters to actual physical processes. The best we can do is look at their effect on the possible solutions. For example, a nonzero value of the parameter s renders the problem asymmetric, so s is introduced to achieve the observed asymmetry of the glacial cycles. The coefficient q is the characteristic time of NADW (expressed in units of the characteristic time of the total global ice mass). The assumption $q > 1$ implies that NADW changes on a faster time scale than the total global ice mass, and as q increases, this change occurs on an increasingly faster time scale. If we rewrite the second equation as $\dot{y} = (r - z^2)y - pz - sz^2$, we see that the growth rate r of the atmospheric CO_2 concentration is balanced by the anomaly of NADW. Lastly, the coefficient p expresses the sensitivity of the atmospheric CO_2 concentration to NADW.

Conceptually, the following sequence of events hints at the possible existence of periodic solutions: (1) As the amount of CO_2 in the atmosphere increases and

y becomes positive, the total amount of ice decreases and x becomes negative (first equation); (2) As x becomes negative, the volume of NADW increases and z becomes positive (third equation); (3) As z becomes positive, the amount of atmospheric CO_2 decreases and y becomes negative (second equation). This is the first part of the cycle.

In the second part of the cycle, once y is negative, the opposite effects happen. (4) The total global ice mass starts to increase again and x becomes positive (first equation); (5) As a result, the volume of NADW decreases and eventually z becomes negative (third equation); (6) Once z is negative, y starts to increase again (second equation). This completes the full cycle and sets the stage for the next cycle. Of course, these arguments do not guarantee the existence of a periodic cycle and do not say anything about its period. The particulars will depend critically on the parameter values.

1.3.4 Computational Results

Maasch and Saltzman found computationally that the system (1.3.8) generates a limit cycle with a 100 Kyr period at the parameter values $p = 1.0$, $q = 1.2$, $r = 0.8$, and $s = 0.8$. The limit cycle is shown in Fig. 1.5. The three curves represent the total ice mass (black), the atmospheric CO_2 concentration (red), and the volume of NADW (blue) in arbitrary units. Each cycle is clearly asymmetric: a rapid deglaciation is followed by a slow glaciation. Also, the three variables are properly correlated: as the concentration of atmospheric CO_2 (a greenhouse gas) increases, the climate gets warmer and the total ice mass decreases; as the volume of NADW increases, the strength of the North Atlantic overturning circulation increases, more atmospheric CO_2 is absorbed by the ocean and, consequently, the atmospheric CO_2 concentration decreases.

More detailed numerical calculations show that the Maasch–Saltzman model possesses limit cycles in large portions of parameter space. We integrated the

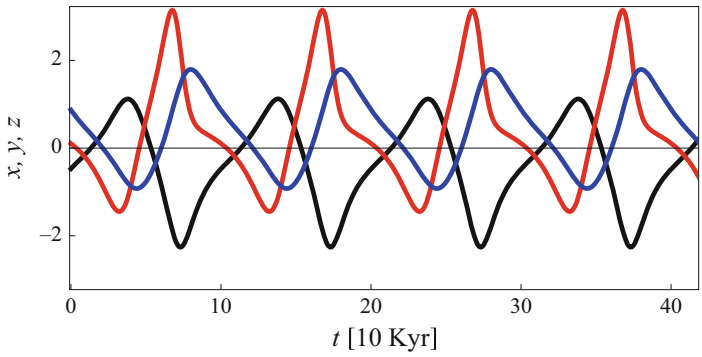


Fig. 1.5 Limit cycle of (1.3.8) at $p = 1.0$, $q = 1.2$, $r = 0.8$, $s = 0.8$

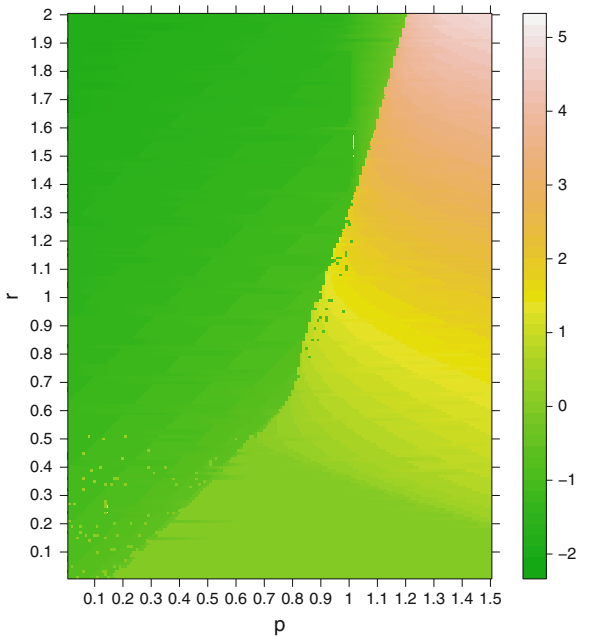


Fig. 1.6 Color map of $\overline{x}(p, r)$ for the system (1.3.8) at $q = 1.2$, $s = 0.8$, indicating convergence to an equilibrium state or a limit cycle

system (1.3.8) forward in time for a range of values of (p, r), keeping q and s fixed at the values $q = 1.2$ and $s = 0.8$, starting from a randomly chosen initial point for each (p, r), until there was a clear indication of either a limit cycle or a limit point. We then determined the quantity $\overline{x}$ for each pair (p, r),

$$\overline{x} = \lim_{t \to \infty} \sup\, x(t). \tag{1.3.10}$$

Figure 1.6 shows the function $(p, r) \mapsto \overline{x}(p, r)$ as a color map. A limiting value 0 (light green) indicates convergence to the trivial state, a nonzero negative value (dark green) convergence to a nontrivial equilibrium state, and a nonzero positive value (orange or pink) convergence to a limit cycle with a finite amplitude. Limit cycles were observed in the entire orange-colored region of the (p, r) plane. These findings, as well as other results reported by Maasch and Saltzman in [35], especially when the effects of orbital forcing are included, suggest that the conceptual model (1.3.8) may indeed provide an explanation for the Pleistocene climate record.

1.4 Simplifying the Maasch–Saltzman Model

The system (1.3.8) has four positive parameters, p, q, r, and s, where $q > 1$. As noted in Sect. 1.3.3, the parameter s introduces asymmetry into the model. If $s = 0$, the equations are invariant under reflection: if (x, y, z) is a solution, then so is

$(-x, -y, -z)$. Note furthermore that, as $q \to \infty$, the differential equation for z reduces, at least formally, to the identity $z = -x$, so the system (1.3.8) becomes two-dimensional. These observations suggest that it may be helpful to analyze the dynamics of the Maasch–Saltzman model (1.3.8) in stages where we first focus on the special case $q = \infty$ and $s = 0$ and then consider the effects of finite values of q and positive values of s.

If we set $q = \infty$ and $s = 0$, the system (1.3.8) reduces formally to a two-dimensional system with $\mathbb{Z}_2$ symmetry,

$$\begin{aligned} \dot{x} &= -x - y, \\ \dot{y} &= ry + px - x^2 y. \end{aligned} \tag{1.4.1}$$

The state variables are x and y, the state space is $\mathbb{R}^2$, p and r are parameters, and the parameter space is $\mathbb{R}^2_+$. Its dynamics can be analyzed rigorously and completely.

1.4.1 Equilibrium States and Their Stability

The origin $P_0 = (0, 0)$ is an equilibrium state of the system (1.4.1) for all values of p and r. If $r > p$, there are two additional equilibrium states, $P_1 = (x_1^*, -x_1^*)$ with $x_1^* = \sqrt{r - p}$, and $P_2 = (x_2^*, -x_2^*)$ with $x_2^* = -\sqrt{r - p}$. A linear stability analysis shows that P_0 is stable if $0 < r < \min(p, 1)$, unstable otherwise; P_1 and P_2 are stable if $0 < p < \min(r, 1)$, unstable otherwise. Thus, the parameter space is partitioned into four regions,

$$\begin{aligned} \mathrm{O} &= \{(p, r) \in \mathbb{R}^2_+ : r < p \text{ for } p < 1,\ r < 1 \text{ for } p > 1\}, \\ \mathrm{I} &= \{(p, r) \in \mathbb{R}^2_+ : 1 < r < p \text{ for } p > 1\}, \\ \mathrm{II} &= \{(p, r) \in \mathbb{R}^2_+ : 1 < p < r \text{ for } r > 1\}, \\ \mathrm{III} &= \{(p, r) \in \mathbb{R}^2_+ : p < r \text{ for } r < 1,\ p < 1 \text{ for } r > 1\}. \end{aligned} \tag{1.4.2}$$

The regions are shown in Fig. 1.7, together with representative trajectories in regions O, I, and II. The diagonal $r = p$ is the locus of pitchfork bifurcations, where P_1 and P_2 are created as stable nodes as (p, r) crosses the diagonal from region O into region III or as unstable nodes as (p, r) crosses the diagonal from region I into region II. The parabolic curves $C_1 = \{p = \frac{1}{4}(r + 1)^2\}$ and $C_2 = \{r = \frac{1}{4}(p + 1)^2\}$ (dashed curves shown in purple), which are tangent to the diagonal $r = p$ at the point $(1, 1)$, mark the boundaries between spirals and nodes.

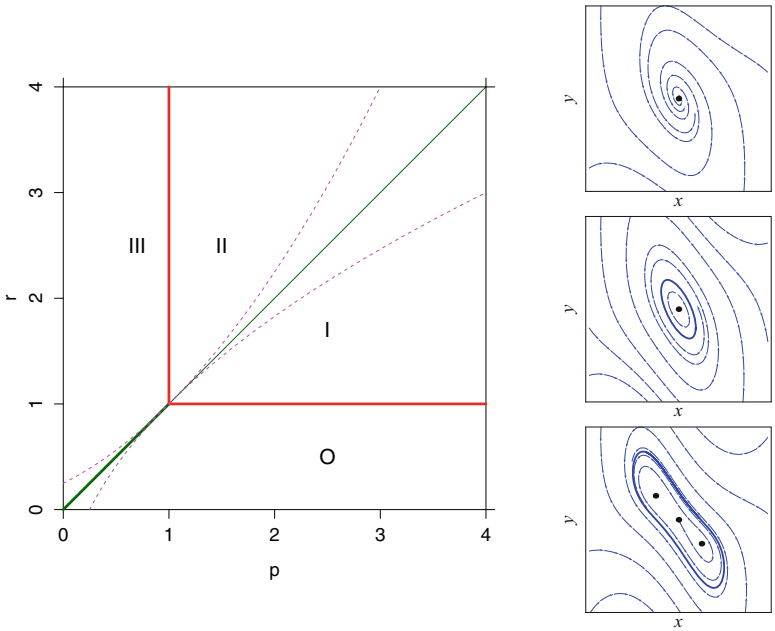


Fig. 1.7 (Left) Stability regions of the equilibrium states P_0, P_1, and P_2 of (1.4.1). (Right) Representative trajectories in region O (top), I (middle), and II (bottom)

1.4.2 Hopf Bifurcations

The system (1.4.1) is equivalent with the Liénard equation

$$\ddot{x} + g(x)\dot{x} + f(x) = 0, \tag{1.4.3}$$

where f and g are polynomial functions,

$$f(x) = x^3 - (r - p)x, \quad g(x) = x^2 - (r - 1). \tag{1.4.4}$$

A Hopf bifurcation occurs when $f(x^*) = 0$, $f'(x^*) > 0$, and x^* is a simple zero of g. The natural frequency of oscillations is $\omega^* = \sqrt{f'(x^*)}$, and the first Lyapunov coefficient at x^* is

$$\ell^* = -\frac{\omega^*}{8} \frac{d}{dx} \frac{g'(x)}{f'(x)}\bigg|_{x=x^*}, \tag{1.4.5}$$

see [31, §3.5]. If $\ell^* < 0$, the Hopf bifurcation is supercritical; if $\ell^* > 0$, it is subcritical. The sign of ℓ^* is the same as that of $f''(x^*)g'(x^*) - f'(x^*)g''(x^*)$,

which in the case of the polynomial functions f and g given in (1.4.4) is the same as that of $3(x^*)^2 + r - p$.

The red line segments in Fig. 1.7 are loci of Hopf bifurcations. On the horizontal segment at $r = 1$, which is associated with P_0, we have $x^* = 0$. The sign of ℓ^* is the same as that of $1 - p$, which is negative, so the Hopf bifurcation is supercritical. The natural frequency is $\omega^* = \sqrt{p-1}$. On the vertical segment at $p = 1$, which is associated with P_1 and P_2, we have $(x^*)^2 = r - p$. The sign of ℓ^* is the same as that of $4(r-1)$, which is positive, so the Hopf bifurcation is subcritical. The natural frequency is $\omega^* = \sqrt{2(r-1)}$.

1.4.3 Organizing Center

The point $(p, r) = (1, 1)$ plays a pivotal role in understanding the complete dynamics of the system (1.4.1). To see why, rotate the coordinate system by the transformation $(x, -(x+y)) \mapsto (x, y)$. In the new coordinates, the system (1.4.1) is

$$\begin{aligned} \dot{x} &= y, \\ \dot{y} &= (r-p)x + (r-1)y - x^2 y - x^3, \end{aligned} \tag{1.4.6}$$

and the equilibrium points are $P_0 = (0, 0)$, $P_1 = (x_1^*, 0)$, and $P_2 = (x_2^*, 0)$.

Let $P = (x^*, 0)$ be any of the equilibrium points, with $x^* = 0$, x_1^*, or x_2^*. The Jacobian of the vector field at P is

$$\begin{pmatrix} 0 & 1 \\ r - p - 3(x^*)^2 & r - 1 - (x^*)^2 \end{pmatrix}. \tag{1.4.7}$$

The matrix has a double-zero eigenvalue at the point $(p, r) = (1, 1)$, so the system (1.4.6) undergoes a Bogdanov–Takens (BT) bifurcation. Holmes and Rand [23] refer to such a point as an *organizing center*. Specifically, given the $\mathbb{Z}_2$ symmetry of the system (1.4.6), the point $(p, r) = (1, 1)$ is a BT point with $\mathbb{Z}_2$ symmetry; examples of such points are discussed in [7, Ch. 4.2] and [31, § 8.4].

The system (1.4.6) can be analyzed in the neighborhood of the organizing center by the unfolding procedure outlined in the original papers by Bogdanov [5] and Takens [53, pp. 23–30] (reprinted in [6, Chapter 1]) and described in the textbooks of Guckenheimer and Holmes [21, §7.3] and Kuznetsov [31, §8.4]. Here, we summarize the results; the details are given in Sect. 1.4.5 below.

Near the point $(p, r) = (1, 1)$, region III of Fig. 1.7 decomposes into three subregions; see Fig. 1.8. In region IIIa, there is one stable limit cycle, with a pair of unstable limit cycles in its interior, one around each of the equilibrium states P_1 and P_2. As (p, r) transits from region IIIa into region IIIb, the two unstable periodic solutions merge to become a pair of unstable homoclinic orbits to the saddle P_0. This homoclinic bifurcation curve (shown in blue) is tangent to the

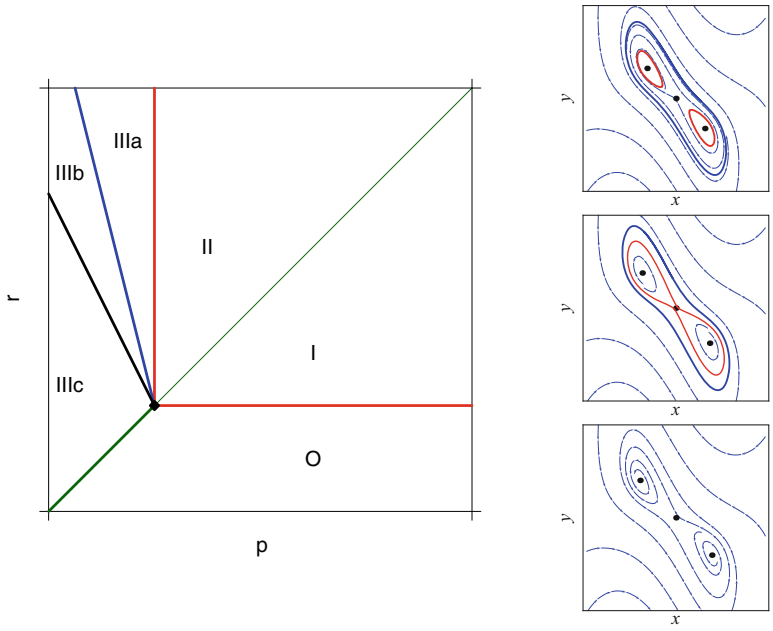


Fig. 1.8 (Left) A sketch of the bifurcation curves of (1.4.1) near the organizing center (the actual curves are shown in Fig. 1.9). (Right) Trajectories in region IIIa (top), IIIb (middle), and IIIc (bottom)

line $r - 1 = -4(p - 1)$ at the organizing center $(1, 1)$. In region IIIb, there is one stable limit cycle with an unstable limit cycle in its interior. As (p, r) transits from region IIIb into region IIIc, there is a curve of saddle-node bifurcations of limit cycles, along which the stable and unstable limit cycles disappear. This curve (shown in black) is tangent to the line $r - 1 \approx -3.03(p - 1)$ at the organizing center $(1, 1)$. In region IIIc, only the three equilibrium states remain, P_0 as an unstable saddle, P_1 and P_2 as stable spirals or nodes.

1.4.4 Computational Results

To complement the analysis, we performed an integration of the system (1.4.1) forward in time for a range of values of (p, r), following the procedure described in Sect. 1.3.4 for Fig. 1.6, to determine the quantity $\overline{x} = \limsup_{t\to\infty} x(t)$, as in (1.3.10). Figure 1.9 shows the function $(p, r) \mapsto \overline{x}(p, r)$ as a color map, together with the bifurcation curves obtained with AUTO. A limiting value 0 (light green) indicates convergence to the trivial state, a nonzero negative value (dark green) convergence to a nontrivial equilibrium state, and a nonzero positive value (orange or pink) convergence to a limit cycle with a finite amplitude.

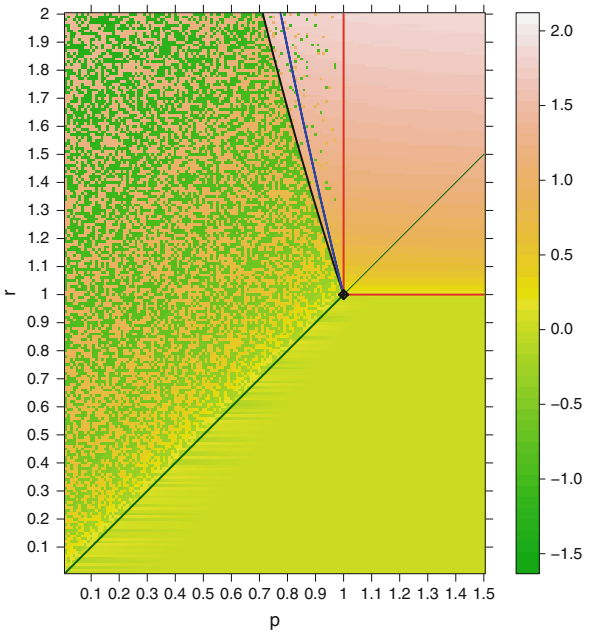


Fig. 1.9 Color map of $\overline{x}(p, r)$ and bifurcation curves for the system (1.4.1)

The stability region O of the trivial state P_0 is clearly visible, as is its Hopf bifurcation curve. As one crosses the Hopf bifurcation curve in the direction of increasing r, the color changes slowly toward increasing values of $\overline{x}$, indicating a supercritical Hopf bifurcation and periodic orbits with amplitudes $\mathcal{O}(\sqrt{r-1})$.

Throughout regions I and II, the color map changes from orange to pink as r increases, indicating that the solutions all approach the stable limit cycle that surrounds P_0; cf. Fig. 1.7.

In region IIIa, there is a similar shift to pink as r increases. One also sees some green and orange patches in region IIIa, indicating that some of the randomly chosen initial conditions lie in the basins of attraction of the stable equilibrium states P_1 and P_2. Next, in region IIIb, the color map has largely the same characteristics as in region IIIa, corresponding to the fact that solutions with initial conditions that lie inside the large unstable limit cycle approach one of the stable equilibria (green or orange), and those with initial conditions outside the unstable limit cycle approach the large stable limit cycle (pink). Finally, in region IIIc, the color map consists entirely of green and orange, indicating that all of the solutions are attracted either to P_1 or to P_2, as expected since there are no stable limit cycles in region IIIc.

1.4.5 Bogdanov–Takens Unfolding

In Sect. 1.4.3, we presented the results of a bifurcation analysis of the system (1.4.6) in a neighborhood of the organizing center at $(p, r) = (1, 1)$. In this section, we present the details of the Bogdanov–Takens unfolding procedure used to establish these results. The section is somewhat technical, but since it is self-contained, it can be skipped at first reading.

The unfolding is achieved by introducing a small positive parameter η (not to be confused with the mean volume of permanent sea ice η used in Sect. 1.3) and rescaling the dependent and independent variables,

$$x(t) = \eta u(\tilde{t}), \quad y(t) = \eta^2 v(\tilde{t}), \quad \tilde{t} = \eta t. \tag{1.4.8}$$

If (x, y) is a solution of the system (1.4.6), then (u, v) must satisfy the system

$$\begin{aligned} \dot{u} &= v, \\ \dot{v} &= \mu u - u^3 + \eta(\lambda - u^2)v. \end{aligned} \tag{1.4.9}$$

Here, the dot $\dot{}$ stands for differentiation with respect to the variable $\tilde{t}$, and λ and μ are parameters, which are defined in terms of p and r,

$$\lambda = \frac{r-1}{\eta^2}, \quad \mu = \frac{r-p}{\eta^2}. \tag{1.4.10}$$

Note that μ is negative in region I and positive in regions II and III (Fig. 1.7). Henceforth, we omit the tilde and write t, instead of $\tilde{t}$.

Remark The definition (1.4.10) of λ and μ generates a linear relation between p and r,

$$(\lambda - \mu)(r-1) = \lambda(p-1). \tag{1.4.11}$$

This is the equation of a pencil through the organizing center $(1, 1)$ parameterized by λ. Referring to the regions labeled I, II, and III in Fig. 1.7, we note that λ increases from 0 to infinity as one rotates counterclockwise from the horizontal line $\{p > 1, r = 1\}$ through region I, then decreases as one continues to rotate through region II, until $\lambda = 1$ at the vertical line $\{p = 1, r > 1\}$, and decreases further as one rotates through region III, until $\lambda = 0$ at the horizontal line $\{0 < p < 1, r = 1\}$.

The results of the local analysis of Sects. 1.4.1 and 1.4.2 may be recovered directly from the system (1.4.9), as follows. The origin $(0, 0)$ is an equilibrium state of (1.4.9) for all λ and μ, and if $\mu > 0$, there are two additional equilibrium states, $(\pm\sqrt{\mu}, 0)$. A linearization of (1.4.9) with $\mu < 0$ about $(0, 0)$ shows that the real parts of the two eigenvalues pass through zero at $\lambda = 0$, which corresponds to the line of supercritical Hopf bifurcations $\{p > 1, r = 1\}$. Similarly, a linearization

of (1.4.9) with $\mu > 0$ about $(\pm\sqrt{\mu}, 0)$ shows that the real parts of the two eigenvalues pass through zero at $\lambda = \mu$, which corresponds to the line of subcritical Hopf bifurcations $\{p = 1, r > 1\}$.

1.4.5.1 Hamiltonian Structures

The equilibrium states of the system (1.4.9) are independent of η, so they persist as $\eta \to 0$. In the limit, (1.4.9) reduces to the Hamiltonian system

$$\begin{aligned} \dot{u} &= v, \\ \dot{v} &= \mu u - u^3. \end{aligned} \tag{1.4.12}$$

Closed orbits of this system are level curves of the Hamiltonian, $H(u, v) = \frac{1}{2}v^2 - \frac{1}{2}\mu u^2 + \frac{1}{4}u^4$. If $\mu < 0$, $H(u, v)$ reaches its minimum value 0 at the origin, so the closed orbits are nested and surround the origin. The more interesting case is $\mu > 0$, where $H(u, v)$ reaches its minimum value at $(\pm\sqrt{\mu}, 0)$, and $H(u, v) = 0$ at the saddle point at the origin. We will analyze the case $\mu > 0$ in detail and return to the case $\mu < 0$ in Sect. 1.4.5.5. Recall that $\mu > 0$ implies that $r > p$, so the following analysis applies to the regions II and III in Fig. 1.7.

Figure 1.10 shows the phase portrait of the Hamiltonian system (1.4.12) with $\mu = 1$. (The phase portrait for other positive values of μ is similar.) We see that there are several types of closed orbits. There is a pair of homoclinic orbits to the origin, there are periodic orbits in the interior of each of the homoclinic orbits surrounding the equilibrium states $(\pm 1, 0)$, and there are large-amplitude periodic orbits external to the homoclinic orbits. The question is, which of these closed orbits persist as the Hamiltonian system (1.4.12) is perturbed to the system (1.4.9). Because of the $\mathbb{Z}_2$ symmetry, it suffices to consider the homoclinic orbit in the right half of the (u, v) plane and the periodic orbits in its interior; the results for the closed orbits in the left

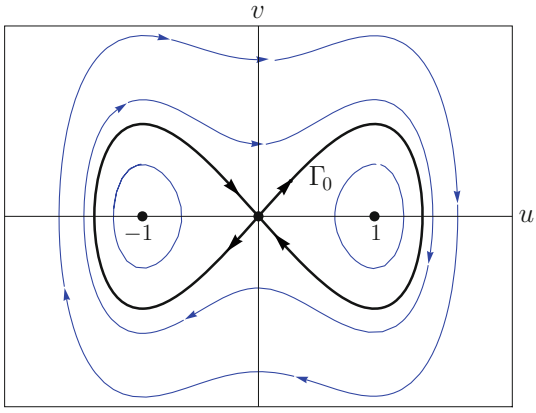


Fig. 1.10 Phase portrait of the Hamiltonian system (1.4.12) with $\mu = 1$

half of the (u, v) plane follow by reflection. Of course, we also need to consider the large-amplitude periodic orbits that are external to the homoclinic orbits. Notice the clockwise orientation of all these orbits.

1.4.5.2 Melnikov Function

The persistence of closed (homoclinic or periodic) orbits of Hamiltonian systems under perturbations can be analyzed by means of the Melnikov function. This function dates back at least to Poincaré [44]; it features in chapters by Melnikov [37] and Arnold [2] and in the book by Andronov et al. [1]. A definitive discussion can be found in the book of Guckenheimer and Holmes [21, §4.5]. The Melnikov function and the associated theory apply to a range of different systems, but for our purposes it suffices to summarize the results for the general system

$$\begin{aligned} \dot{u} &= v, \\ \dot{v} &= f(u) + \eta g(u, v). \end{aligned} \tag{1.4.13}$$

(The functions f and g are not to be confused with those in Sect. 1.4.2.) In the limit as $\eta \to 0$, this system reduces to the Hamiltonian system

$$\begin{aligned} \dot{u} &= v, \\ \dot{v} &= f(u). \end{aligned} \tag{1.4.14}$$

Let $\Gamma_0 = \{t \mapsto (u(t), v(t)), t \in I\}$ be any closed orbit of (1.4.14). The *Melnikov function* associated with Γ_0 is the integral $\int_I g(u(t), v(t))v(t)\,dt$. Thus, the Melnikov function measures the cumulative effect of the projection of the perturbed component of the vector field, $[0\ g]^{\mathrm{t}}$, on the normal vector, $[-f\ v]^{\mathrm{t}}$, of the unperturbed vector field along Γ_0. If the Melnikov function vanishes on Γ_0, then there exists—under suitable nondegeneracy conditions—a family of closed orbits Γ_η of the perturbed system (1.4.13) which are $\mathcal{O}(\eta)$ close to Γ_0 as $\eta \to 0$. Moreover, if the Melnikov function vanishes on Γ_0 and is positive (negative) on nearby orbits that are to the right as Γ_0 is traversed, then the Γ_η are locally stable (unstable).

1.4.5.3 Dynamics in Regions II and III

We apply the general results of the previous section to the closed orbits of (1.4.9) identified at the end of Sect. 1.4.5.1: either the homoclinic orbit in the right half of the (u, v) plane, or one of the periodic orbits in its interior, or one of the large-amplitude periodic orbits external to the homoclinic orbits (Fig. 1.10). We assume, without loss of generality, that $\mu = 1$, so $f(u) = u - u^3$ and $g(u, v) = (\lambda - u^2)v$.

Let $\gamma = \{(u(t), v(t)) : t \in \mathbb{R}\}$ be any of these closed orbits. To indicate a particular orbit, we label γ by the maximum value of its first coordinate, $u(t)$, on its

trajectory. We consider this label as a variable and denote it by x (not to be confused with the dependent variable x in the Maasch–Saltzman model). Thus, the function $\gamma : x \mapsto \gamma(x)$ is defined for all $x \in (1, \infty)$. Specifically, $\gamma(x)$ is the homoclinic orbit in the right half plane if $x = \sqrt{2}$, a periodic orbit inside this homoclinic if $x \in (1, \sqrt{2})$, and a large-amplitude periodic orbit that is external to the double homoclinic if $x > \sqrt{2}$.

Remark There are several ways to choose an identifier for γ. For example, we could equally well have chosen the level-set value $h = H(\gamma)$, as was done in [8].

Consider the closed orbit $\gamma(x) = \{(u_x(t), v_x(t)) : t \in I(x)\}$ for any $x > 1$, where $I(x) = \mathbb{R}$ if $x = \sqrt{2}$, and $I(x)$ is a period interval otherwise. The Melnikov function associated with $\gamma(x)$ is

$$M(\lambda, x) = \int_{I(x)} (\lambda - u^2(t))v^2(t)\, dt = \oint_{\gamma(x)} (\lambda - u^2)\, v(u)\, du = \lambda I_0(x) - I_2(x), \tag{1.4.15}$$

where I_0 and I_2 are defined by

$$I_0(x) = \oint_{\gamma(x)} v(u)\, du, \quad I_2(x) = \oint_{\gamma(x)} u^2\, v(u)\, du. \tag{1.4.16}$$

Here, we have used the relation $v = \dot{u}$ to convert the time integral to a contour integral on the closed orbit $\gamma(x)$.

Recall that $\gamma(x)$ is oriented clockwise; hence, Green's theorem yields the identity $I_0(x) = \iint_{D(x)} du\, dv$, where $D(x)$ is the domain enclosed by $\gamma(x)$. Consequently, $I_0(x) > 0$, so the condition $M(\lambda, x) = 0$ is satisfied if and only if

$$\lambda = R(x), \quad \text{where } R(x) = \frac{I_2(x)}{I_0(x)}. \tag{1.4.17}$$

If, given λ, $M(\lambda, x) = 0$ at $x = \tilde{x}$, then $M(\lambda, x) > 0$ for nearby orbits that are to the right of $\gamma(\tilde{x})$ if $R'(\tilde{x}) < 0$, and $M(\lambda, x) < 0$ for such nearby orbits if $R'(\tilde{x}) > 0$. Hence, the local stability of closed orbits is determined by the sign of $R'(\tilde{x})$.

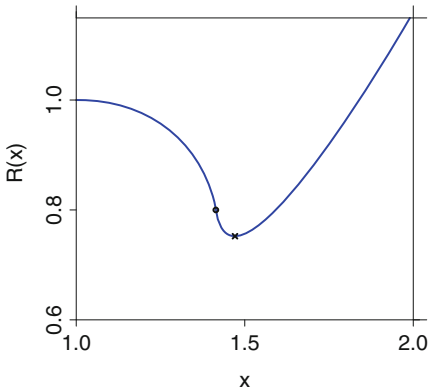
Since the components of the homoclinic and periodic orbits are known in terms of hyperbolic and elliptic functions, respectively, $R(x)$ can be evaluated explicitly. The computations were first done by Carr [7] and subsequently refined by Cushman and Sanders [12]. For example, for the homoclinic orbit,

$$\gamma(\sqrt{2}) = \{(\sqrt{2}\,\mathrm{sech}\, t, -\sqrt{2}\,\mathrm{sech}\, t \tanh\, t),\ t \in \mathbb{R}\}, \tag{1.4.18}$$

and $M(\lambda, \sqrt{2}) = \frac{4}{3}\lambda - \frac{16}{15}$. It follows that $M(\lambda, \sqrt{2}) = 0$ if and only if $\lambda = \frac{4}{5}$. Moreover, $\lambda = \frac{4}{5}$ is a simple zero. Therefore, for all sufficiently small η there exists a $\lambda(\eta) = \frac{4}{5} + \mathcal{O}(\eta)$ and a homoclinic orbit near $\gamma(\sqrt{2})$, with a symmetric result in the left half of the (u, v) plane. In the (p, r) plane, the homoclinic bifurcation curve

is, to leading order, tangent to the line $r - 1 = -4(p - 1)$ at the organizing center $(1, 1)$; see (1.4.11).

The figure below shows the graph of $R : x \mapsto R(x)$ for $1 < x < 2$.



Starting from the values $R(1) = 1$ and $R'(1) = 0$, $R(x)$ and $R'(x)$ decrease monotonically as x increases. At the homoclinic orbit (marked by a black dot), $x = \sqrt{2}$, $R(\sqrt{2}) = \frac{4}{5}$, and $\lim_{x \to \sqrt{2}} R'(x) = -\infty$. Beyond the homoclinic orbit, $R(x)$ decreases further, while $R'(x)$ increases until $R'(x) = 0$; at that point (marked by a black cross), $x = x^* \approx 1.471$ and $R''(x^*) > 0$, so $R(x)$ reaches its minimum value $R_{\min}(x^*) = \lambda^* \approx 0.752$. Beyond this point, $R(x)$ increases monotonically; $R(x) \sim x^2$ as $x \to \infty$.

Proofs of these statements, which do not use elliptic functions, can be found, for example, in [8]. It follows that closed orbits exist only for $\lambda > \lambda^*$, and they are locally stable only for $x > x^*$.

1.4.5.4 Limit Cycles in Regions II and III

The properties of the Melnikov function listed in the previous section lead to the following results for the dynamics of the perturbed system (1.4.9) with $\mu = 1$. In all statements, it is assumed that η is sufficiently small positive.

- For $\lambda > 1$ (region II), there are only stable large-amplitude limit cycles which encircle the origin and pass through points $(x, 0)$ with $R(x) > 1$;
- For $\frac{4}{5} < \lambda < 1$ (region IIIa), there are stable large-amplitude limit cycles which encircle the origin and unstable limit cycles in their interior which encircle the equilibrium points $(\pm 1, 0)$;
- For $\lambda = \frac{4}{5} + \mathcal{O}(\eta)$, there is a symmetric pair of unstable homoclinic orbits, one in each half plane;
- For $0.752\ldots < \lambda < \frac{4}{5}$ (region IIIb), there is a stable large-amplitude limit cycle and an unstable large-amplitude limit cycle in its interior, both encircling the origin;

- At $\lambda = 0.752\ldots + \mathcal{O}(\eta)$, the stable and unstable large-amplitude limit cycles join in a saddle-node bifurcation;
- For $\lambda < 0.752\ldots$ (region IIIc), there are no limit cycles.

Thus, in addition to the homoclinic bifurcation curve found earlier, which is tangent to the line $r - 1 = -4(p - 1)$ at the organizing center, there is a curve of saddle-node bifurcations of limit cycles, which, to leading order, is tangent to the line $r - 1 \approx -3.03(p - 1)$ at the organizing center. This follows from (1.4.11), with $\mu = 1$ and $\lambda^* \approx 0.752$. The two bifurcation curves partition the region III of Fig. 1.7 into the three regions IIIa, IIIb, and IIIc, as sketched in Fig. 1.8 and superimposed on the color map of Fig. 1.9.

1.4.5.5 Limit Cycles in Region I

It remains to investigate the dynamics of the system (1.4.9) for $\mu < 0$ ($r < p$, region I in Fig. 1.7). This case is considerably simpler than the case $\mu > 0$. Without loss of generality, we may assume that $\mu = -1$. The Hamiltonian is $H(u, v) = \frac{1}{2}v^2 + \frac{1}{2}u^2 + \frac{1}{4}u^4$. The closed orbits can again be identified by the maximum value, x, of its first coordinate $u(t)$, which in this case ranges over all positive values, $x > 0$. The Melnikov function is given by the same expression (1.4.15) and vanishes if $\lambda = R(x)$, as in (1.4.17). In this case, both $R(x)$ and $R'(x)$ increase as x increases, so the Melnikov theory establishes that, for each $x > 0$, there exists a value of λ (given by the simple zero of the Melnikov function) such that, for some $\lambda(\eta)$ that is $\mathcal{O}(\eta)$ close to this value, the perturbed system has a unique limit cycle.

1.5 The Asymmetric Two-Dimensional Model

Having a complete understanding of the dynamics of the symmetric two-dimensional model (1.4.1), we are in a position to study the effects of symmetry breaking. The asymmetric two-dimensional model is derived formally from the Maasch–Saltzman model (1.3.8) by setting $q = \infty$ ($z = -x$),

$$\begin{aligned} \dot{x} &= -x - y, \\ \dot{y} &= ry + px + sx^2 - x^2 y. \end{aligned} \tag{1.5.1}$$

We will see that this system has two nondegenerate Bogdanov–Takens points, which act as organizing centers in the (p, r) parameter space. The geometry of these organizing centers and the bifurcation curves emanating from them may be understood naturally as a result of the breaking of the lone $\mathbb{Z}_2$-symmetric Bogdanov–Takens point studied in Sect. 1.4.3.

1.5.1 Equilibrium States and Their Stability

The origin $P_0 = (0, 0)$ is an equilibrium state of (1.5.1) for all (positive) values of p, r, and s. If $r > p - \frac{1}{4}s^2$, there are two additional equilibrium states, namely $P_1 = (x_1^*, -x_1^*)$ and $P_2 = (x_2^*, -x_2^*)$, where

$$x_1^* = \tfrac{1}{2}[-s + \sqrt{s^2 + 4(r-p)}], \quad x_2^* = \tfrac{1}{2}[-s - \sqrt{s^2 + 4(r-p)}]. \tag{1.5.2}$$

We refer to the line $r = p - \frac{1}{4}s^2$ as the *shifted diagonal* (marked "sd" in Fig. 1.11). Note that $x_2^* < x_1^* < 0$ if $p - \frac{1}{4}s^2 < r < p$, and $x_2^* < 0 < x_1^*$ if $r > p$.

Let $P = (x^*, -x^*)$ be any of the equilibrium states, with $x^* = 0$, x_1^*, or x_2^*. The Jacobian of the vector field at P is

$$\begin{pmatrix} -1 & -1 \\ p + 2sx^* + 2(x^*)^2 & r - (x^*)^2 \end{pmatrix}. \tag{1.5.3}$$

The system is linearly stable at P if the trace is negative, $-1 + r - (x^*)^2 < 0$, and the determinant is positive, $p - r + 3(x^*)^2 + 2sx^* > 0$.

The stability results are illustrated in Fig. 1.11. We see that the parameter space $\mathbb{R}^2_+$ is partitioned into six regions, which depend on s. Referring to the labels in Fig. 1.11, these regions are

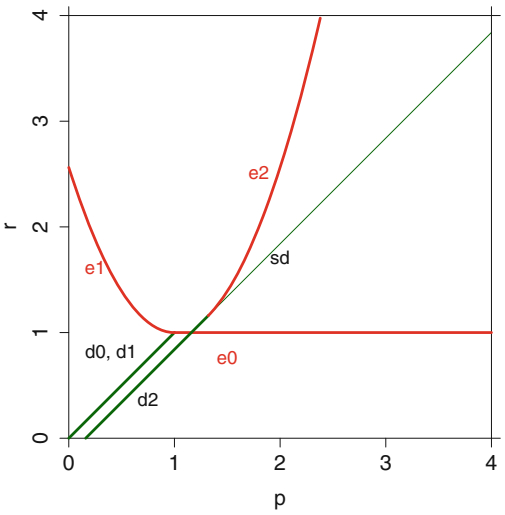


Fig. 1.11 Stability regions of P_0, P_1, and P_2 for (1.5.1) with $s = 0.8$

$$\begin{aligned}
\text{Oa} &= \left\{(p,r) \in \mathbb{R}^2_+ : \text{ between “d0” and “d2”, below “e0”}\right\},\\
\text{Ob} &= \left\{(p,r) \in \mathbb{R}^2_+ : \text{ right of “d2”, below “e0”}\right\},\\
\text{I} &= \left\{(p,r) \in \mathbb{R}^2_+ : \text{ right of “sd”, above “e0”}\right\},\\
\text{IIa} &= \left\{(p,r) \in \mathbb{R}^2_+ : \text{ between “e2” and “sd”}\right\},\\
\text{III} &= \left\{(p,r) \in \mathbb{R}^2_+ : \text{ left of “d0” and “e1”}\right\},\\
\text{IIIo} &= \left\{(p,r) \in \mathbb{R}^2_+ : \text{ between “e1” and “e2”, above “e0” and “d2”}\right\}.
\end{aligned} \tag{1.5.4}$$

Summarizing the results of the stability analysis, we find that

- P_0 is stable in regions Oa and Ob, undergoes a supercritical Hopf bifurcation on “e0” with natural frequency $\omega_0^* = \sqrt{p-1}$;
- P_1 is stable in region III, undergoes a subcritical Hopf bifurcation on “e1” with natural frequency $\omega_1^* = \sqrt{2(r-1) + s\sqrt{r-1}}$; and
- P_2 is stable in regions Oa, III, and IIIo, undergoes a subcritical Hopf bifurcation on “e2” with natural frequency $\omega_2^* = \sqrt{2(r-1) - s\sqrt{r-1}}$.

The introduction of asymmetry ($s > 0$) results in two changes. The vertical line $\{p = 1, r > 1\}$, which is the locus of Hopf bifurcations for P_1 and P_2 in the symmetric case (Fig. 1.7), unfolds into a parabola. The vertex of this parabola is at the point $(1, 1)$, and the parabola is tangent to the shifted diagonal $r = p - \frac{1}{4}s^2$ at the point $(1 + \frac{1}{2}s^2, 1 + \frac{1}{4}s^2)$. As we will see in Sect. 1.5.2, both these points are organizing centers. For convenience, we label them

$$Q_1 = (1,1), \quad Q_2 = (1 + \tfrac{1}{2}s^2, 1 + \tfrac{1}{4}s^2). \tag{1.5.5}$$

As (p, r) moves across the shifted diagonal in the direction of decreasing p, the equilibrium states P_1 and P_2 emerge in a saddle-node bifurcation. If the crossing occurs above Q_2, P_1 and P_2 are both unstable; if it occurs below Q_2, P_1 is unstable while P_2 is stable. In the region Oa, P_0 and P_2 co-exist as stable equilibria, while P_1 is an unstable equilibrium. On the diagonal for $p < 1$, P_0 and P_1 exchange stability in a transcritical bifurcation.

1.5.2 Organizing Centers

We make the change of variables $(x, -(x + y)) \mapsto (x, y)$ as in Sect. 1.4.3. In the new coordinate system, (1.5.1) becomes

$$\begin{aligned}
\dot{x} &= y,\\
\dot{y} &= (r-p)x + (r-1)y - (s+y)x^2 - x^3.
\end{aligned} \tag{1.5.6}$$

This is a special case of general four-parameter planar vector fields which arise as part of the unfolding of vector fields $\dot{x} = y$, $\dot{y} = -x^3 - x^2 y$; they have been studied extensively in [13, 29] (codimension-two singularities) and [18] (codimension-three singularities).

The equilibrium points of the system (1.5.6) are $P_0 = (0, 0)$, $P_1 = (x_1^*, 0)$, and $P_2 = (x_2^*, 0)$, where x_1^* and x_2^* are again given by (1.5.2).

Let $P = (x^*, 0)$ be any of the equilibrium points. The Jacobian of the vector field at P is

$$\begin{pmatrix} 0 & 1 \\ r - p - 2sx^* - 3(x^*)^2 & r - 1 - (x^*)^2 \end{pmatrix}. \tag{1.5.7}$$

If $x^* = 0$, the Jacobian has a double-zero eigenvalue at Q_1 for any s, and if $x^* = x_1^*$ or $x^* = x_2^*$, it has a double-zero eigenvalue at Q_2. Hence, the introduction of asymmetry causes the organizing center to unfold into a center at Q_1 associated with P_0 and a center at Q_2 associated with P_1 and P_2.

Figure 1.12 shows the bifurcation curves emanating from the two organizing centers for $s = 0.8$ (the value chosen by Maasch and Saltzman). They were computed with the AUTO continuation package.

There are three Hopf bifurcation curves (shown in red), two emanating from Q_1 and one emanating from Q_2: (1) a parabolic curve to the left of Q_1, where P_1 undergoes a subcritical Hopf bifurcation; (2) a horizontal line $\{r = 1, p > 1\}$ to the right of Q_1, where P_0 undergoes a supercritical Hopf bifurcation; and (3) a parabolic curve to the right of Q_2, along which P_2 undergoes a subcritical Hopf bifurcation. These three curves are the same as the ones identified in the local analysis, cf. Sect. 1.5.1. In addition, there are three homoclinic bifurcation curves (shown in blue), two emanating from Q_1 (solid blue) and one emanating from Q_2 (dashed

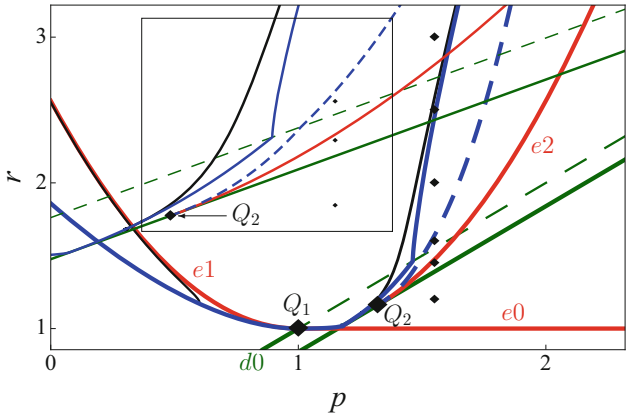


Fig. 1.12 Stability boundaries and bifurcation curves for the system (1.5.1) with $s = 0.8$. The inset shows a neighborhood of Q_2

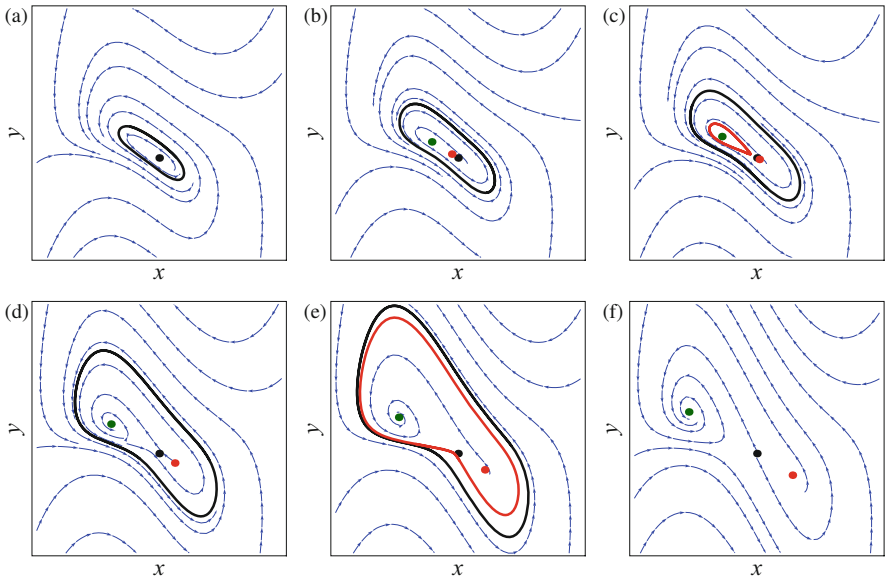


Fig. 1.13 Phase planes of the system (1.5.1) at the six small, black diamond markers in Fig. 1.12. The markers all lie on the vertical line $p = 1.55$; (**a**) $r = 1.2$, (**b**) $r = 1.45$, (**c**) $r = 1.6$, (**d**) $r = 2.0$, (**e**) $r = 2.5$, (**f**) $r = 3.0$

blue). These curves are identified by an unfolding procedure, as in Sect. 1.4.5. Since the points Q_1 and Q_2 are both nondegenerate Bogdanov–Takens points, there are no other bifurcation curves besides the Hopf and homoclinic bifurcation curves emanating from them.

Figure 1.13 shows the phase portraits at the six small black diamond markers along the vertical line $p = 1.55$ in Fig. 1.12. The color scheme is as follows: The flow of (1.5.1) is shown as blue streamlines. The black and red curves correspond to the stable and unstable limit cycles, respectively. The equilibrium states P_0, P_1, and P_2 are indicated by black, red, and green markers, respectively.

- Frame (a), $r = 1.2$: a stable limit cycle surrounds the unstable equilibrium point P_0.
- Frame (b), $r = 1.45$: the equilibrium points P_1 and P_2 exist but are unstable, a stable limit cycle surrounds P_1 and P_2.
- Frame (c), $r = 1.6$: the equilibrium points P_1 and P_0 have switched positions, P_2 has become stable, an unstable limit cycle surrounds P_2.
- Frame (d), $r = 2.0$: the unstable limit cycle has disappeared in the homoclinic bifurcation.
- Frame (e), $r = 2.5$: a large-amplitude unstable limit cycle exists inside the large-amplitude stable limit cycle.
- Frame (f), $r = 3.0$: the large-amplitude stable and unstable limit cycles have disappeared in a saddle-node bifurcation; P_2 is the only attractor.

1.6 Summary

The purpose of this chapter is to show an interesting application of dynamical systems theory to a problem of climate science. The object of investigation is a model of the Pleistocene climate, proposed by Maasch and Saltzman in 1990. The model is a *conceptual* model designed specifically to explain the persistence of glacial cycles during the Pleistocene Epoch. The Milankovitch theory of orbital forcing establishes a correlation between glacial cycles and periodic oscillations in the Earth's orbit around the Sun, but orbital forcing by itself is insufficient to explain the observed temperature changes.

The Maasch–Saltzman model incorporates a feedback mechanism that is driven by greenhouse gases, in particular atmospheric CO_2. The model is based on plausible physical arguments, and preliminary computational experiments indicate that it reproduces several salient features of the Pleistocene temperature record. In this chapter, the emphasis is on the internal dynamics of the model. We focused on the prevalence and bifurcation properties of limit cycles in the various parameter regimes in the absence of external forcing.

The Maasch–Saltzman model is formulated in terms of the anomalies of the total global ice mass, the atmospheric CO_2 concentration, and the volume of the North Atlantic Deep Water (NADW, a measure of the strength of the North Atlantic overturning circulation). It consists of three differential equations with four parameters and is difficult to analyze directly. Our results indicate how one can obtain fundamental insight into its complex dynamics by first considering a highly simplified two-dimensional version. The dimension reduction is achieved by (formally) letting one of the parameters—representing the rate of change of the volume of NADW relative to that of the total global ice mass—tend to infinity. The approximation is justified by the observation that the NADW changes on a much faster time scale than the total global ice mass.

The two-dimensional model has two primary parameters, p and r, which are both positive, and one secondary parameter s, which reflects the asymmetry between the glaciation and deglaciation phases of the glacial cycles. By first ignoring the asymmetry ($s = 0$), we obtained a complete understanding of the dynamics and the persistence of limit cycles.

Figure 1.8, which is a sketch of the various bifurcation curves in the (p, r) parameter space, summarizes the main results. The origin is an equilibrium state P_0 for all (p, r); P_0 is stable in region O. In addition, there are two equilibrium states P_1 and P_2 in regions II and IIIa-c; they are generated in a pitchfork bifurcation along the diagonal $r = p$ and are stable in region IIIa-c. Stable limit cycles exist in regions O, I, II, IIIa, and IIIb. They are created in supercritical Hopf bifurcations along the boundary between regions O and I. In region I, they surround the unstable equilibrium state P_0, and in the other regions they surround all three equilibrium states. Along the boundary between regions II and IIIa, a pair of unstable limit cycles, each surrounding one of the two stable equilibrium states P_1 and P_2, are created in a subcritical Hopf bifurcation. These newborn limit cycles grow in

amplitude until they become homoclinic orbits at the boundary between regions IIIa and IIIb; in region IIIb, they have merged into one large-amplitude unstable limit cycle, which surrounds the three equilibrium states and sits just inside the large-amplitude stable limit cycle. Finally, the stable and unstable large-amplitude limit cycles merge and disappear in a saddle-node bifurcation as (p, r) crosses the lower boundary of region IIIb into region IIIc. All these results have been confirmed computationally; see Fig. 1.9.

Having thus obtained a complete understanding of the dynamics of the symmetric simplified model, we then re-introduced asymmetry ($s > 0$). A comparison of Fig. 1.12 with Fig. 1.8 shows the effects of symmetry breaking. The main change is that the single organizing center, which governed the dynamics in the symmetric case, splits into two organizing centers. Also, the curves of homoclinic bifurcations and saddle-node bifurcations of limit cycles become more complex.

The complexity of the bifurcation diagram of Fig. 1.12 gives an indication why it is difficult to analyze the dynamics of the Maasch–Saltzman model directly. It also justifies our approach of first analyzing the highly simplified model and then gradually relaxing the constraints that were imposed to derive the simplified model [19]. In that paper, we briefly examine the effects of slowly varying the parameters p and r in the full model and identify this as a slow passage through a Hopf bifurcation curve, with a resulting delayed loss of stability. This is the main mechanism for the mid-Pleistocene transition in the Maasch–Saltzman model. The effects of orbital forcing on the model are shown to be important also for this mechanism, since they may advance or delay the loss of stability further.

Acknowledgement The work of T.J. Kaper and Th. Vo was supported in part by NSF grant DMS-1616064.

References

1. Andronov, A.A., Leontovich, E.A., Gordon, I.I., et al.: Theory of Bifurcations of Dynamic Systems on a Plane. Israel Program of Scientific Translations, Jerusalem (1971)
2. Arnold, V.: Instability of dynamical systems with several degrees of freedom. Soviet Math. Dokl. **5**, 581–585 (1964)
3. Ashkenazy, Y., Tziperman, E.: Are the 41 kyr glacial oscillations a linear response to Milankovitch forcing? Quat. Sci. Rev. **23**, 1879–1890 (2004)
4. Ashwin, P., Ditlevsen, P.: The Mid-Pleistocene transition as a generic bifurcation on a slow manifold. Clim. Dyn. **45**(9–10), 2683–2695 (2015)
5. Bogdanov, R.I.: Versal deformations of a singular point of a vector field on the plane in the case of zero eigenvalues. Funct. Anal. Appl. **9**(2), 144–145 (1975)
6. Broer, H.W., Krauskopf, B., Vegter, G.: Global Analysis of Dynamical Systems. Institute of Physics Publishing, London (2001)
7. Carr, J.: Applications of Centre Manifold Theory. Springer, New York (1981)
8. Chow, Y.K.: Melnikov's method with applications. Tech. rep., MA thesis, The University of British Columbia (2001)
9. Clark, P., Alley, R., Pollard, D.: Northern hemisphere ice-sheet influences on global climate change. Science **5442**, 1104–1111 (1999)

10. Cook, K.H.: Climate Dynamics. Princeton University Press, Princeton (2013)
11. Crucifix, M.: Oscillators and relaxation phenomena in Pleistocene climate theory. Phil. Trans. R. Soc. A **370**(1962), 1140–1165 (2012). https://doi.org/10.1098/rsta.2011.0315
12. Cushman, R., Sanders, J.A.: A codimension two bifurcation with a third order Picard–Fuchs equation. J. Differ. Equ. **59**, 243–256 (1985)
13. Dangelmayr, G., Guckenheimer, J.: On a four-parameter family of planar vector fields. Arch. Ration. Mech. Anal. **97**, 321–352 (1987)
14. Dijkstra, H.A.: Nonlinear Climate Dynamics. Cambridge University Press, Berlin, Heidelberg (2013)
15. Doedel, E.J.: AUTO: a program for the automatic bifurcation analysis of autonomous systems. Congr. Numer. **30**, 265–284 (1981)
16. Doedel, E.J., Keller, H.B., Kernevez, J.P.: Numerical analysis and control of bifurcation problems (i): Bifurcation in finite dimensions. Int. J. Bifurcat. Chaos pp. 493–520 (1991)
17. Doedel, E.J., Champneys, A.R., Fairgrieve, T.F., et al.: AUTO-07P: continuation and bifurcation software for ordinary differential equations. Tech. rep., Concordia University, Montreal, Canada (2007). http://cmvl.cs.concordia.ca/
18. Dumortier, F., Roussarie, R., Sotomayor, J.: Generic 3-Parameter Families of Planar Vector Fields, Unfoldings of Saddle, Focus and Elliptic Singularities With Nilpotent Linear Parts, vol. 1480, pp. 1–164. Springer Lecture Notes in Mathematics. Springer, Berlin (1991)
19. Engler, H., Kaper, H.G., Kaper, T.J., Vo, T.: Dynamical systems analysis of the Maasch–Saltzman model of glacial cycles. Phys. D **359**, 1–20 (2017)
20. Ghil, M.: Cryothermodynamics: the chaotic dynamics of paleoclimate. Phys. D **77**, 130–159 (1994)
21. Guckenheimer, J., Holmes, P.: Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields, 3rd printing, revised and corrected edn. Springer, New York (1985)
22. Hayes, J.D., Imbrie, J., Shackleton, N.J.: Variations in the Earth's orbit: pacemaker of the ice ages. Science **194**, 1121–1132 (1976)
23. Holmes, P., Rand, D.: Phase portraits and bifurcations of the non-linear oscillator $\ddot{x} + (\alpha + \gamma x)\dot{x} + \beta x + \delta x^3 = 0$. Int. J. Nonlinear Mech. **15**, 449–458 (1980)
24. Huybers, P.: Pleistocene glacial variability and the integrated insolation forcing. Science **313**, 508–511 (2006)
25. Huybers, P.: Glacial variability over the last two million years: an extended depth-derived agemodel, continuous obliquity pacing, and the Pleistocene progression. Quat. Sci. Rev. **26**, 37–55 (2007)
26. Huybers, P., Wunsch, C.: Obliquity pacing of the late-Pleistocene glacial cycles. Nature **434**, 491–494 (2005)
27. Imbrie, J., Raymo, M.E., Shackleton, N.J., et al.: On the structure and origin of major glaciation cycles. 1. Linear responses to Milankovitch forcing. Paleoceanography **6**, 205–226 (1992)
28. Kaper, H.G., Engler, H.: Mathematics & Climate. OT131. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (2013)
29. Khibnik, A., Krauskopf, B., Rousseau, C.: Global study of a family of cubic Liénard equations. Nonlinearity **11**, 1505–1519 (1998)
30. Kuznetsov, Y.A.: Practical computation of normal forms on center manifolds at degenerate Bogdanov–Takens bifurcations. Int. J. Bifurcation Chaos **15**(11), 3535–3546 (2005)
31. Kuznetsov, Y.A.: Elements of Applied Bifurcation Theory, vol. 112. Springer, Berlin (2013)
32. Laskar, J., Fienga, A., Gastineau, M., et al.: La2010: a new orbital solution for the long-term motion of the earth. Astron. Astrophys. **532**, A89 (2011)
33. Le Treut, H., Ghil, M.: Orbital forcing, climatic interactions, and glaciation cycles. J. Geophys. Res. Oceans **88**(C9), 5167–5190 (1983). http://dx.doi.org/10.1029/JC088iC09p05167
34. Lisiecki, L.E., Raymo, M.E.: A Pliocene-Pleistocene stack of 57 globally distributed benthic δ^{18}O records. Paleoceanography **20**(1) (2005). http://dx.doi.org/10.1029/2004PA001071. PA1003
35. Maasch, K.A., Saltzman, B.: A low-order dynamic model of global climate variability over the full Pleistocene. J. Geophys. Res. **95**(D2), 1955–1963 (1990)

36. Marshall, J., Plumb, R.A.: Atmosphere, Ocean and Climate Dynamics: An Introductory Text. Academic Press, London (2007)
37. Melnikov, V.: On the stability of the center for time-periodic perturbations. Trans. Moscow Math. Soc. **12**, 1–57 (1963)
38. Milankovič, M.: Kanon der erdbestrahlung und seine anwendung auf das eiszeitenproblem. Tech. rep., University of Belgrade (1941)
39. Paillard, D.: Modèles simplifiés pour l'étude de la variabilité de la circulation thermohaline au cours des cycles glaciaire-interglaciaire. Ph.D. thesis, Univ. Paris-Sud (1995)
40. Paillard, D.: The timing of Pleistocene glaciations from a simple multiple-state climate model. Nature **391**, 378–391 (1998)
41. Paillard, D.: Glacial cycles: toward a new paradigm. Rev. Geophys. **39**, 325–346 (2001)
42. Paillard, D., Parrenin, F.: The Antarctic ice sheet and the triggering of deglaciation. Earth Planet. Sci. Lett. **227**, 263–271 (2004)
43. Petit, J.R., Davis, M., Delaygue, G., et al.: Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. Nature **399**, 429–436 (1999)
44. Poincaré, H.: Sur les équations de la dynamique et le problème des trois corps. Acta Math. **13**, 1–270 (1890)
45. Raymo, M.E., Nisancioglu, K.H.: The 41 kyr world: Milankovitch's other unsolved mystery. Paleoceanography **18**(1) (2003). http://dx.doi.org/10.1029/2002PA000791. PA1011
46. Raymo, M., Oppo, D., Curry, W.: The mid-Pleistocene climate transition: a deep sea carbon isotopic perspective. Paleoceanography **12**, 546–559 (1997)
47. Rutherford, S., D'hondt, S.: Early onset and tropical forcing of 100,000-year Pleistocene glacial cycles. Nature **408**(6808), 72–75 (2000)
48. Saltzman, B.: Carbon dioxide and the δ^{18}O record of late Quaternary climate change: a global model. Clim. Dyn. **1**, 77–85 (1987)
49. Saltzman, B., Maasch, K.A.: Carbon cycle instability as a cause of the late Pleistocene ice age oscillations: modeling the asymmetric response. Global Biogeochem. Cycles **2**, 177–185 (1988)
50. Saltzman, B., Maasch, K.A.: A first-order global model of late Cenozoic climatic change II. Further analysis based on a simplification of CO_2 dynamics. Clim. Dyn. **5**, 201–210 (1991)
51. Shackleton, N.J.: The 100,000-year ice-age cycle identified and found to lag temperature, carbon dioxide and orbital eccentricity. Science **289**, 1897–1902 (2000)
52. Sun, D.Z., Bryan, F.: Climate Dynamics: Why Does Climate Vary? Wiley, Wiley (2013)
53. Takens, F.: Forced oscillations and bifurcations. Tech. Rep. 3, Mathematics Institute, Rijksuniversiteit Utrecht, the Netherlands (1974). Reprinted in Chapter 1: Broer, H.W., Krauskopf, B., Vegter, G.: Global Analysis of Dynamical Systems. Institute of Physics Publishing, London (2001)
54. Tziperman, E., Gildor, H.: On the mid-Pleistocene transition to 100-kyr glacial cycles and the asymmetry between glaciation and deglaciation times. Paleoceanography **18**(1) (2003). http://dx.doi.org/10.1029/2001pa000627. PA1001

Chapter 2
Mathematics of the *Not-So-Solid* Solid Earth

Scott D. King

Abstract As a result of climatic variations over the past 700,000 years, large ice sheets in high-latitude regions of the Earth formed and subsequently melted, loading and unloading the surface of the Earth. This chapter introduces the mathematical analysis of the vertical motion of the solid Earth in response to this time-varying surface loading. This chapter focuses on two conceptual models: the first, proposed by Haskell [Physics, **6**, 265–269 (1935)], describes the return to equilibrium of a viscous half-space after the removal of an applied surface load; the second, proposed by Farrell and Clark [Geophys. J. Royal Astr. Soc., **46**, 647–667 (1976)], illustrates the changes in sea level that occur when ice and water are rearranged on the surface of the Earth. The sea level equation proposed by Farrell and Clark accounts for the fact that sea level represents the interface between two dynamic surfaces: the sea surface and the solid Earth, both of which are changing with time.

Keywords Gravitational potential · Sea level · Stokes equation · Viscous relaxation

2.1 Ice Ages and Glacial Isostatic Adjustment

For the past 700,000 years, the Earth's climate has alternated between glacial and interglacial conditions, with a periodicity on the order of 100,000 years. A conceptual model emphasizing the roles of orbital variations and atmospheric CO_2 concentration is explored in Chap. 1 of this volume. During glacial periods, lower temperatures result in the growth of large ice sheets at higher latitudes, removing water from the ocean basins and lowering sea levels. During interglacial periods, these large ice sheets melt, returning water stored on land to the oceans, resulting in a relative rise in sea levels. The movement of water in both liquid and

S. D. King (✉)
Department of Geosciences, Virginia Tech, Blacksburg, VA, USA
e-mail: sdk@vt.edu

H. G. Kaper, F. S. Roberts (eds.), *Mathematics of Planet Earth*, Mathematics of Planet Earth 5, https://doi.org/10.1007/978-3-030-22044-0_2

solid form between continental land masses and ocean basins during the glacial–interglacial cycle creates a time-varying mass load on the surface of the Earth on a time scale that is short compared with the response time of the Earth's surface. The mass of these ice sheets is sufficient to deform the solid Earth, causing subsidence and then, upon subsequent melting of the ice sheet, rebound of the surface. The response of the solid Earth to the time-varying surface load brought about by the waxing and waning of large-scale ice sheets is called *Glacial Isostatic Adjustment* (GIA). Isostasy (or isostatic) is a term used by Earth scientists to describe the Archimedean principle that any object, wholly or partially immersed in a stationary fluid, is buoyed up by a force equal to the weight of the fluid displaced by the object.

During the last great ice age, Scandinavia and North America were covered with thick sheets of ice up to 5 km thick (Fig. 2.1). In northern Europe, the northward extent of the ice sheet covered Svalbard and Franz Josef Land, and the southern boundary passed through Germany and Poland. In North America, the ice covered most of Canada, extending as far south as the Missouri and Ohio Rivers, and eastward to Manhattan. When the ice sheets melted, the surface of the Earth began to return to its equilibrium elevation (rebound), a process that continues to the present day [21, 22, 39, 42]. The water stored in these ice sheets lowered the sea level globally by 115–135 m relative to present-day sea levels [23], and present-day sea levels are at or near the maximum level in the glacial–interglacial cycle.

While water is also present both in the atmosphere and stored as groundwater within the near surface of the Earth, the volumes of water involved in the atmospheric water cycle (e.g., precipitation, evaporation, and transpiration) and stored as groundwater do not vary significantly over the glacial/interglacial time scale. Changes in these water reservoirs have a much smaller effect on the solid Earth in comparison with the changing mass load of the ice sheets.

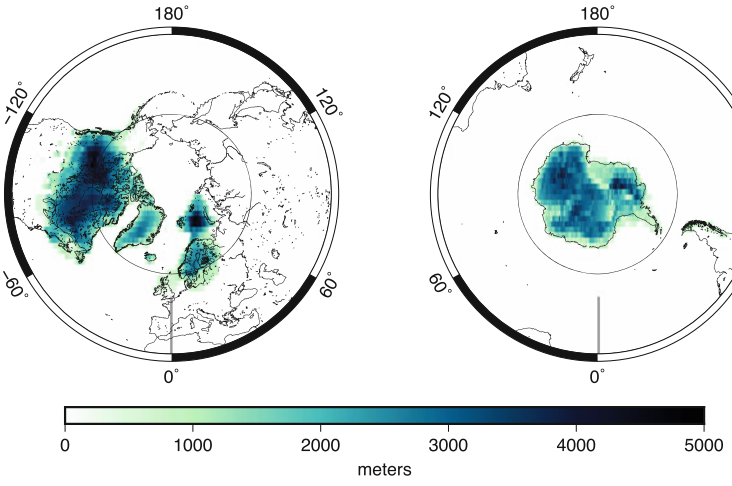


Fig. 2.1 Ice thickness 26,500 years before present based on the ICE-6G model [34]

2.1.1 Sea Level Changes

Tidal gauges were originally designed to measure the daily and monthly changes in water level due to tides in shallow harbors. A typical tidal gauge would consist of a mechanical float enclosed in a cylindrical well to isolate the float from wind waves. After removing the daily and monthly tidal signals, it is possible to derive a record of the mean sea level from these historical tidal gauge measurements; at some locations, the records are continuous, covering several centuries. Of course, these records are spatially limited to coastal regions, for obvious reasons. It is also necessary to reference a tidal gauge to a local geodetic benchmark, to ensure that the local land surface is stable and that the recorded measurement reflects a change in sea level and not local subsidence or uplift of the land. While still used in some locations, these early gauges have been superseded by pressure, acoustic/ultrasonic, or radar gauges. (Here and throughout the remainder of this chapter, we use the term "sea level" to refer to the level of a hypothetical ocean surface in the absence of wind waves.)

Figure 2.2 shows a time series of annual mean sea level anomalies for Amsterdam [43] and Stockholm [7]. (The *sea level anomaly* is the deviation of the actual sea level from some reference level.) The point of this figure is to illustrate the trend as a function of time, and this is not dependent on the specifics of the reference baseline. In Amsterdam, the sea level increased nearly 200 mm during the period 1700–1925, while the sea level in Stockholm decreased almost 1000 mm between 1770 and 1980. While sea level observations at these locations continue to be recorded, the changes in instrumentation and analysis techniques mean that matching modern tide-gauge measurements with these historical records is a nontrivial exercise. For the purpose here, these historical records are sufficient to illustrate the longer-term trends in the sea-level observations.

The traditional approach to estimating the impact of ice sheets on the sea level assumed that measuring the age of submerged beaches (e.g., by radiocarbon dating) in stable areas was sufficient to determine the historical changes in sea level. The assumption was that sea level rise is a global phenomenon and that, just as increasing the water level in a bathtub would increase the water level everywhere, sea level rise at one point would inform the global trend. This allowed researchers to extend the

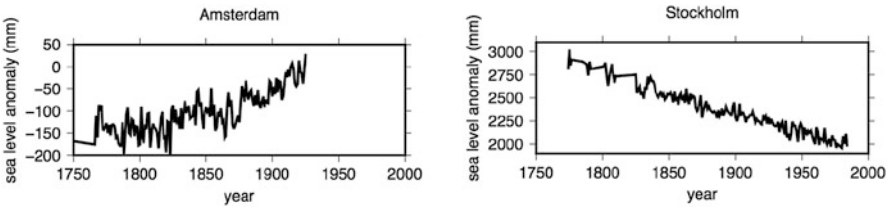


Fig. 2.2 Annual mean sea level anomalies for Amsterdam [43] and Stockholm [7]. Data obtained from http://www.psmsl.org/data/longrecords/

sea-level record further back in time, although with a greater degree of uncertainty. The time series in Fig. 2.2 present a problem for this traditional approach. Why does the sea level appear to be rising at one location and falling at another? The model described by Farrell and Clark [9], which we will discuss in Sect. 2.3, illustrates that the distribution of ice and the shape of the surface of the solid Earth play an important role in the analysis of sea-level change.

In any given area, the local sea level is the intersection of two dynamic surfaces: the sea surface and the irregular solid surface, both of which are changing with time. Globally, the sea level is not at a uniform radial distance from the center of the Earth (or some other suitable reference frame) but varies spatially because it follows a surface of equal gravitational potential. While water is removed and added from the ocean basin with the growth and melting of the ice sheets, the gravitational attraction of the ice sheets deforms the ocean surface, thus changing the gravitational potential in the region around the sheet. Also, the change in the mass of both the ocean and the ice sheet as water moves from one to the other creates a time-varying load that deforms the surface of the Earth. As the surface deforms, matter within the Earth is redistributed, the gravitational attraction changes, and the sea level responds in turn. This is the topic that will be analyzed in detail in the remainder of the chapter.

Outline of the Chapter Following is an outline of the chapter. Section 2.2 contains a review of the classical problem studied by Haskell [11, 12] of a mass load on the surface of a viscous half-space. The analysis yields an estimate of the viscosity of the interior of the Earth, which justifies the assumption that the effects of momentum and rotation on the slow creeping flow in the interior of the Earth can be neglected. Haskell's analysis predicts a uniform rise in sea level everywhere. Section 2.3 focuses on spatial variations of the gravitational potential, following the classical work of Farrell and Clark [9]. The analysis assumes a rigid Earth but allows for a nonuniform sea level rise. The subsequent Sect. 2.4 builds on this work by adding elastic deformations to obtain a more realistic model of a solid Earth. Since purely elastic behavior is not consistent with the GIA observations, some degree of viscous behavior is required. The most common model of a viscoelastic medium is the Maxwell rheology model, which is discussed in Sect. 2.5. The final section, Sect. 2.6, summarizes the main points of the chapter with references to more in-depth reviews and describes various open problems.

2.2 The Haskell Problem: Viscous Relaxation of the Solid Earth

It may seem extraordinary that on time scales longer than we can perceive the *terra firma* upon which we go about our daily lives actually behaves like a fluid, albeit a highly viscous fluid. Yet, the study of Earth's tectonic plates shows that the surface of the Earth moves with velocities on the order of tens of millimeters per year [26, 38, 45]—that is, roughly the rate at which human finger nails grow [3]. In

addition to the horizontal motions of the Earth's surface, the surface deforms vertically as a result of both imposed surface loads and stresses from within the Earth [4].

The first mathematical formulation of the rebound of the surface of the Earth after the melting of an ice sheet was given by Haskell [11, 12]. Haskell calculated the flow within a semi-infinite, incompressible, viscous half-space, subject to an initial periodic surface displacement given by

$$w_m = w_{m0} \cos \frac{2\pi x}{\lambda}. \tag{2.2.1}$$

Here, λ is the wavelength of the initial load. The amplitude of the deformation of the surface is assumed to be much smaller than its wavelength, $w_m \ll \lambda$. The load-induced displacement generates a hydrostatic pressure gradient, which acts to restore the surface of the Earth to the undeformed equilibrium state ($w = 0$).

The equation of *mass conservation* for an incompressible fluid is

$$\nabla \cdot U = 0, \tag{2.2.2}$$

where U is the velocity vector describing the fluid motion. The equation of *momentum conservation* is obtained by applying Newton's second law to the fluid motion and using the assumption that the stress in the fluid is the sum of a pressure term and a viscous term that is proportional to the gradient of the velocity. The resulting equation is the *Navier–Stokes equation*, which describes the dynamics of fluids in many areas of engineering and science,

$$\rho \left[\frac{\partial U}{\partial t} + U \cdot \nabla U \right] = -\nabla p + \eta \nabla^2 U. \tag{2.2.3}$$

Here, ρ is the density of the fluid, p the pressure, and η the viscosity; $\frac{\partial \cdot}{\partial t}$ is the partial derivative with respect to time.

Viscosityis a measure of the resistance of a fluid to gradual deformation by shear stress. Honey is more resistant to flow than water, so honey has a larger viscosity than water. In the SI system of units, viscosity is measured in Pascal-seconds (Pa s). The viscosities of some common fluids are listed in the adjacent table.

Fluid	Density (kg/m^3)	Viscosity (Pa s)
Air	1.3	10^{-5}
Water	1000	10^{-3}
Olive oil	916	0.1
Honey	1450	10
Glacial ice	800–900	10^{15}

While it may not be obvious that the solid interior of the Earth deforms, consider the movement of glaciers. Glaciers are sometimes called "rivers of ice," and they actually flow in response to gravity acting on their own mass. The rate of glacial motion ranges from less than a meter per year to as much as 30 m per day when the base of the glacier is decoupled from the underlying bedrock by soft sediments and meltwater.

For specific problems, it is often possible to simplify the Navier–Stokes equation because one or more terms in the equation are significantly smaller than the others. To show this, it is helpful to rewrite the equation in terms of dimensionless variables, which are of order 1, multiplied by a dimensional scaling constant. For example, length can be written as $x = x'L$, where x' is of order 1 and L represents the characteristic length scale of the problem. In the problem under consideration, the length scale is the wavelength of the applied load, λ. Similarly, the depth can be written as $y = y'L$, using the same length scale of the problem, L. The velocity can be rewritten as $U = U'U_0$, where U_0 is the characteristic velocity of the problem. Here, the velocity of Earth's tectonic plates serves as a reasonable estimate of the characteristic velocity, $U_0 = 0.01$ m/yr, or $U_0 \approx 3.16 \times 10^{-9}$ m/s. (Even though the second is the unit of time in the SI system, geoscientists think of plate velocities in millimeters per year. There are approximately $\pi \times 10^7$ s in a year.) The characteristic time can now be defined in terms of L and U_0, $t = t'L/U_0$. A logical choice for pressure scaling is $p = p'\eta U_0/L$, which results in units of Pascals, the SI unit of pressure.

Substituting the above relationships into Eq. (2.2.3), we obtain the Navier–Stokes equation in dimensionless form,

$$Re\left[\frac{\partial U'}{\partial t'} + U' \cdot \nabla' U'\right] = -\nabla' p' + (\nabla')^2 U', \tag{2.2.4}$$

where the scaling constants and properties of the fluid have been grouped into a single term, the *Reynolds number*,

$$Re = \frac{\rho U_0 L}{\eta}. \tag{2.2.5}$$

The units in the Reynolds number cancel, so Re is a dimensionless quantity—one of several that arise in the study of fluid mechanics. It is also noteworthy that the primed Eq. (2.2.4) is dimensionless. This is useful for a variety of reasons; for example, if the physical properties of different problems result in the same Reynolds number, their solutions will be identical. Hence, if the characteristic length is increased by a factor of 10, the dimensionless solution will be the same if the viscosity is also increased by a factor of 10. The dimensional solution can be recovered by multiplying the dimensionless solution by the scaling constants.

While nothing has been said yet about the viscosity of the interior of the Earth, it is not hard to imagine that it is large, at least as large as the viscosity of glacial ice; hence, the Reynolds number will be very small, and the terms on the right-hand side of the Navier–Stokes equation can be ignored. This assumption will be checked after

the final solution has been obtained. Following the same procedure, it is easy to show that the terms representing the effects of the Earth's rotation are similar in magnitude to the momentum terms on the left-hand side of the Navier–Stokes equation. Thus, if our analysis shows that momentum can be ignored, so can rotation.

Consider a 2-dimensional domain in a vertical plane. Choose a Cartesian coordinate system in the plane, with horizontal coordinate x and vertical coordinate y, with y increasing downward. The surface of the Earth is represented by a function $y = w(x)$. Let u and v denote the x and y components, respectively, of the velocity vector U.

The components of Eq. (2.2.4) in the x and y direction are

$$-\frac{\partial p}{\partial x} + \eta\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) = 0, \tag{2.2.6}$$

$$-\frac{\partial p}{\partial y} + \eta\left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2}\right) = 0. \tag{2.2.7}$$

This set of equations can be solved using the stream-function formulation [2]; a step-by-step solution can be found, for example, in [14]. In 2-dimensions, the stream function, $\psi(x, y, t)$, is a scalar whose partial derivatives are related to the components of the velocity,

$$u = \frac{\partial \psi}{\partial y}, \quad v = -\frac{\partial \psi}{\partial x}. \tag{2.2.8}$$

Note that, by construction, $U = (u, v)$ satisfies the equation of mass conservation (2.2.2).

Because the initial displacement of the surface varies in x with a functional form $\cos(2\pi x/\lambda)$, the stream function will vary with a functional form $\sin(2\pi x/\lambda)$. By taking the derivative of Eq. (2.2.6) with respect to y and the derivative of Eq. (2.2.7) with respect to x and subtracting the two resulting equations, we eliminates the pressure. Then, upon substitution of the expressions (2.2.8) we obtain a single biharmonic equation for the scalar ψ. Separating ψ into a function that varies only in x (i.e., $\sin(2\pi x/\lambda)$) and a function that varies only in $Y(y)$, we find that the stream function must have the form

$$\psi = \sin\left(\frac{2\pi x}{\lambda}\right)\left[(A + By)e^{-2\pi y/\lambda} + (C + Dy)e^{2\pi y/\lambda}\right], \tag{2.2.9}$$

where A, B, C, and D are constants, to be determined by the boundary conditions. The constants C and D must be zero because the components of the velocity field must remain finite as the depth of the half-space (y) goes to infinity. Differentiating ψ, the components of velocity are

$$u = \sin\left(\frac{2\pi x}{\lambda}\right)\left[\frac{2\pi}{\lambda}(A + By) - B\right]e^{-2\pi y/\lambda}, \tag{2.2.10}$$

$$v = \cos\left(\frac{2\pi x}{\lambda}\right)\frac{2\pi}{\lambda}(A + By)\, e^{-2\pi y/\lambda}. \tag{2.2.11}$$

The shear tractions exerted on the solid Earth by the atmosphere are negligible, so the horizontal velocity of the flow in the solid Earth is zero at the deformed surface—that is, at $y = w(x)$. Since the displacements of the solid Earth (at most hundreds of meters) are small compared to the size of a typical ice sheet (a thousand kilometers or more), we may assume that $w \ll \lambda$ and apply the boundary conditions at the equilibrium surface—that is, at $y = 0$—instead. Johnson and Fletcher [14] show how to apply the boundary conditions to the deformed surface when w is not small; their solution reduces to the one presented here in the case of small deformations. Setting $u = 0$ at $y = 0$ yields $B = 2\pi A/\lambda$.

To find A, the hydrostatic pressure resulting from the topography ($-\rho g w$, where g is the acceleration due to gravity) is set equal to the normal stress generated by the flow at the surface ($p - 2\mu\frac{\partial v}{\partial y}$), where ρ is the density of the mantle,

$$-\rho g w = p - 2\mu\frac{\partial v}{\partial y}. \tag{2.2.12}$$

The pressure at $y = 0$ can be found by substituting Eqs. (2.2.10) and (2.2.11) into Eq. (2.2.6) and integrating,

$$p|_{y=0} = 2\mu A\left(\frac{2\pi}{\lambda}\right)^2 \cos\left(\frac{2\pi x}{\lambda}\right). \tag{2.2.13}$$

Because $\frac{\partial v}{\partial y}|_{y=0} = 0$, Eq. (2.2.12) reduces to

$$w = -\frac{2\mu A}{\rho g}\left(\frac{2\pi}{\lambda}\right)^2 \cos\left(\frac{2\pi x}{\lambda}\right). \tag{2.2.14}$$

The key step is to substitute Eq. (2.2.11) (with $B = 2\pi A/\lambda$) into Eq. (2.2.14) and recognize that the vertical velocity at the surface is the derivative of the displacement with time, $v = dw/dt$ (at $y = w$). Once again, because the displacements are small compared with the size of a typical ice sheet, we apply this condition at $y = 0$,

$$v|_{y=0} = \left.\frac{dw}{dt}\right|_{y=0} = A\frac{2\pi}{\lambda}\cos\left(\frac{2\pi x}{\lambda}\right) = -w\frac{\lambda g\rho}{4\pi\mu}. \tag{2.2.15}$$

Upon integration, we obtain the expression for w,

$$w(t) = w_{m0}\exp\left(-\frac{\lambda g\rho}{4\pi\mu}t\right). \tag{2.2.16}$$

Hence, the surface of the Earth decays to an equilibrium position as the "fluid" mantle flows from regions of elevated topography to regions of low topography. The grouping of constants $\frac{4\pi\mu}{\lambda g\rho}$ has units of time and is the characteristic time scale of the Glacial Isostatic Adjustment (GIA)—that is, the time it takes the topography to decay by $1/e$. Using reasonable values for the density of the solid Earth, $\rho = 3300$ kg m^{-3}, the acceleration due to gravity, $g = 10$ m s^{-2}, and the spatial scale of the ice sheet, $\lambda = 1000$ km, matching the time scale of GIA from the tide-gauge and beach data requires the viscosity of the mantle to be on the order of $\eta \sim 10^{21}$ Pa s. This is sometimes referred to as the *Haskell value of mantle viscosity* and is an average or effective value, as it assumes that the Earth is a homogeneous fluid. Using the same values of $\rho$and λ, taking $u = 0.01$ m yr^{-1} and the Haskell value of viscosity, $\eta = 10^{21}$ Pa s, we obtain a Reynolds number on the order of 10^{-21}, which justifies the initial assumption that the inertial terms in the Navier–Stokes equation can be ignored.

While the Haskell problem is simplified, the characteristic time of GIA, $\frac{4\pi\mu}{\lambda g\rho}$, is approximately 12,000 years. The last glacial maximum (i.e., the time when the ice sheets were at their largest spatial extent) occurred approximately 26,500 years ago and the North American and European ice sheets began to retreat about 20,000 years ago. The characteristic time predicts that vertical rebound of the Earth's surface should still be continuing—a prediction that has been validated with high-precision GPS observations [13, 21, 22, 27, 30, 39, 42]. Other independent geophysical constraints on mantle viscosity are broadly consistent with the Haskell result [16]. Until recently, observations of vertical uplift were measured almost exclusively along coast lines via sea- and lake-level changes, requiring climatic, hydrographic, and tectonic corrections, and horizontal motions could not be accurately observed at all. This state of affairs changed with the development of high-precision GPS.

2.3 Gravitational Potential: The Spatial Variability of Sea Level

One might assume that estimating the change in sea level is as simple as estimating the mass of ice sheets at their maximum extent, converting this mass to an equivalent volume of water, and adding that volume of water to the ocean. This approach predicts that sea level should have risen by an equal amount everywhere, which is inconsistent with the observations [22, 35], as shown, for example, in the time-series of Fig. 2.2. Two effects are missing: First, there is a gravitational attraction between the ocean and ice sheets, and second, both the ocean and the ice sheets deform the Earth's surface. Sea level is the intersection of these two dynamic surfaces (the sea surface and the solid Earth surface), both of which are changing with time.

To illustrate the role of gravitational attraction on sea level, consider the simplified problem of a rigid sphere that is initially covered by a thin ocean of uniform depth. This problem is discussed in Farrell and Clark [9], and the text below

follows their derivation. We will simplify the problem further by assuming that the ocean has zero density, yet is at the same time in gravitational equilibrium (which assumes that the ocean has a nonzero density). While these assumptions are clearly inconsistent, they allow for an analytic solution of the problem.

For a spherically symmetric Earth, where r is the distance between the observer and the Earth's center of mass, the gravitational potential is

$$V(r) = \frac{GM_E}{r}, \tag{2.3.1}$$

where G is Newton's gravitational constant and M_E is the total mass of the Earth (which includes the solid Earth and the ocean). Equation (2.3.1) is valid for $r \geq a$, where a is the radius of the Earth. A direct result of the spherical symmetry of the problem is that the sea level of this uniform-depth ocean will be equal to a everywhere, because the gravitational potential at the surface, $V(a)$, is a constant. Now suppose that an ice sheet of mass M_I is extracted from the surface at $r = a$ and placed at a single point on the Earth's surface. Let θ measure the angular distance between the point mass (ice sheet) and an observer. Then the new gravitational potential field is

$$V^I(r,\theta) = \frac{G(M_E - M_I)}{r} + \frac{GM_I}{\sqrt{r^2 + a^2 - 2ar\cos\theta}}. \tag{2.3.2}$$

The superscript ^{I} denotes the combined potential of the Earth plus ice sheet. Note that $M_E - M_I$, and therefore the first term in Eq. (2.3.2), is still spherically symmetric; however, $V^I(a,\theta)$ is a function of θ and therefore $r = a$ is no longer the sea level because $V^I(a,\theta)$ is not constant. Defining a new surface at $r = a + \varepsilon$, where $V^I(a+\varepsilon,\theta) = V^I(a)$ and assuming that $M_I \ll M_E$, it follows that $\varepsilon \ll a$ and a first-order Taylor expansion can be used to approximate $V^I(a+\varepsilon,\theta)$,

$$V^I(a+\varepsilon,\theta) = V^I(a,\theta) + \varepsilon\frac{\partial V^I(a,\theta)}{\partial r}. \tag{2.3.3}$$

It is sufficiently accurate to use the approximation $\frac{\partial V^R(a,\theta)}{\partial r} = -g$, where g is the acceleration due to gravity at the Earth's surface. Rearranging Eq. (2.3.3) and substituting $g = \frac{GM}{a^2}$, we obtain

$$\varepsilon(\theta) = \frac{M_I a}{M_E}\left(\frac{1}{2\sin(\theta/2)} - 1\right). \tag{2.3.4}$$

At this point, the analysis has yet to account for the volume of water that has been lost from the ocean; therefore, $V^I(a+\varepsilon(\theta),\theta)$ is constant but is not the sea level. To account for the reduced ocean volume, recall that if $a + \varepsilon(\theta)$ is an equipotential surface, then for any constant $c \ll a$, $a+\varepsilon(\theta)+c = a+\varepsilon_2(\theta)$ is also an equipotential surface. The trick is to choose a value of c so as to conserve the total mass of the

system. Farrell and Clark suggest that the result thus obtained is an accurate estimate of sea level.

To calculate c, note that the volume between the surfaces a and $a + \varepsilon$ integrated over the sphere is zero, so in order to conserve mass, it must be the case that

$$\int_0^{\pi} 2\pi\rho_w c a^2 \sin\theta \, d\theta + M_I = 0, \tag{2.3.5}$$

where ρ_w is the density of sea water. Solving Eq. (2.3.5) for c and using $M_E = (4/3)\pi a^3 \rho_E$, where ρ_E is the mean density of the Earth, we obtain

$$\varepsilon_2(\theta) = \frac{M_I a}{M_E}\left(\frac{1}{2\sin(\theta/2)} - 1 - \frac{\rho_E}{3\rho_w}\right). \tag{2.3.6}$$

The first two terms on the right-hand side represent the distortion of sea level due to the gravitational attraction of the ice; the third term is the uniform fall in sea level due to the removal of a volume of water equivalent to the ice mass M_I from the oceans. Figure 2.3 shows the change in sea level (normalized by the predicted uniform sea-level drop) due to the removal of an amount of water equivalent to M_I as a function of the angular distance from the ice mass.

At a point 60° from the ice mass, the predicted sea-level drop is the same as predicted from the uniform sea-level decrease. Beyond 60°, the sea-level drop is greater than the uniform prediction, while within 20° of the ice load the sea level actually rises due to the gravitational attraction indexgravitational attractionof the ice acting on the ocean. This result provides a qualitative explanation for the historical sea-level trends observed at Amsterdam and Stockholm shown in Fig. 2.2: Stockholm is closer to the center of the Fennoscandian ice sheet than Amsterdam. However, a word of caution is appropriate here, because the assumptions made in this section may limit the applicability of the results. Nonetheless, it is instructive as an illustration of the role of the gravitational attraction of the ice sheet on sea level.

The problem described here considers the shoreline to be spatially fixed with time as the sea level rises and falls during a glacial cycle. Or to help the reader visualize,

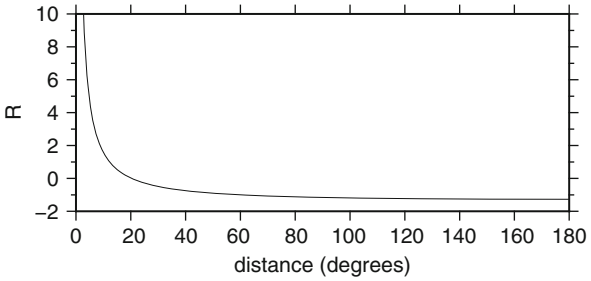


Fig. 2.3 The normalized change in sea level as a function of distance from the ice mass for a rigid Earth, including the effect of the gravitational attraction of the ice sheet

it is equivalent to assuming that the edges of the ocean basins are characterized by steep cliffs that prevent the water from moving either landward or oceanward. More accurate and complex shoreline calculations are described and compared in Mitrovica and Milne [24].

2.3.1 Extending the Solution to an Irregular Ice Distribution

To extend the point-mass problem to any arbitrary volume of ice, Eq. (2.3.6) is convolved with a function that represents the variation in ice thickness, $I(\theta', \phi')$. The change in sea level, $S(\theta, \phi)$, due to a change in ice mass is given by the integral

$$S(\theta,\phi) = \iint_{ice} \left[\frac{a}{M_E} \left(\frac{1}{2\sin(\alpha/2)} - 1 - \frac{\rho_E}{3\rho_w} \right) \right] \rho_I \, I(\theta', \phi') \, a \sin(\theta') \, d\theta' \, d\phi', \tag{2.3.7}$$

where α is the arc length between a point (θ, ϕ) on the ocean and the location (θ', ϕ') of the ice. The term in brackets, which is identical to (2.3.6) with unit ice mass (M_I), is the Green's function for this problem.

To account for the gravitational attraction of the mass of the ocean on sea level, one can similarly convolve the Green's function with sea level. An iterative solution strategy for this problem is discussed in Farrell and Clark [9].

2.4 Deformation of the Solid Earth: The Elastic Earth

While the analysis in Sect. 2.3 is instructive and provides a possible qualitative explanation for the trends observed in the sea-level curves shown in Fig. 2.2, the results of Sect. 2.3 are inconsistent with the Haskell problem in Sect. 2.2 because the Earth was assumed to be rigid. In this section we will outline how the sea-level equation, Eq. (2.3.7), can be extended to include the deformation of the solid Earth. First, it is necessary to briefly review the possible ways in which the solid Earth might deform.

When placed under a load, the surface of the Earth exhibits both elastic and viscous behavior. A material is said to behave elastically when it deforms instantaneously in response to an applied force and returns to its original state immediately after the force is removed. A spring is often used as the classic example of elastic behavior. On the other hand, a viscous material undergoes transient, permanent deformation when a force is applied. Honey is often used as the classic example of a viscous fluid. A material that behaves both elastically and viscously is called a *viscoelastic material*. A viscoelastic material will experience both instantaneous and transient deformation upon the application of a force. When the original force is removed, the transient deformation is reversed; however, unlike

the elastic material, the viscoelastic material does not return to it's initial state and some permanent deformation is retained. The child's toy Silly-Putty is often used as an example of a viscoelastic material. When Silly-Putty is dropped, it bounces like a rubber ball, exhibiting elastic behavior in response to the short-timescale force of the Silly-Putty and accelerating due to the force of gravity as it falls, until it is stopped by the immovable floor. If a similar force is applied over a longer time period, the Silly-Putty yields and stretches into a long thin strand, much like taffy.

The Earth also exhibits this dual deformation behavior depending on the time scale of the forcing function. Over the time period of several million years, Earth's surface deforms viscously when subjected to an applied load such as the mass of a volcano or the volume of water in the ocean basin [8]. This is the mode of deformation that Haskell assumed to be appropriate in the problem described in Sect. 2.2. On a time scale of seconds to hours, the Earth behaves elastically in response to seismic waves. Since the time scale of the growth and decay of ice sheets is on the order of 100,000 years, elastic deformation cannot be ignored [22, 46]. The derivation of the elastic response of an incompressible spherically symmetric Earth is given in [1, 5, 9]. Here, we focus on how the sea-level equation (2.3.7) can be modified to account for the deformation of the Earth.

The solution of the elastic deformation problem requires solving the linear momentum equation (similar to the Navier–Stokes equation for viscous flow) for the displacement (instead of the velocity as in the viscous flow problem), coupled with the solution of a Poisson equation for the gravitational potential of a spherically symmetric body with material properties that are functions only of the radius, subject to a disk-shaped surface boundary load. The elastic deformation of a spherical body subject to a disc point load can be represented in terms of three *Love numbers*, h_l, k_l, and l_l, which depend only upon the radius, r, and the degree of the spherical harmonic, l [20].

A short digression here is necessary to introduce spherical harmonics [29, Chapter 14.30]. Spherical harmonics are often used to represent functions on a sphere. They play the same role on a sphere as sines and cosines on a line; as such they appear frequently in geophysical problems. With the proper normalization, spherical harmonics can be written in terms of Legendra polynomials, $P_{lm}(\cos\theta)$, multiplied by cosines and sines in the azimuth ϕ,

$$Y_{lm}(\theta,\phi) = P_{lm}(\cos\theta)(C\cos m\phi + S\sin m\phi), \tag{2.4.1}$$

where l is the spherical harmonic degree and m is the spherical harmonic order. The spherical harmonic functions form a set of basis functions on a sphere, so they can be used to represent any function on a sphere with an infinite set of coefficients, with many similarities between spherical harmonics and Fourier series analysis in terms of solution techniques. For example, if the topography of a planet is given by topo(θ,ϕ), then

$$\text{topo}(\theta,\phi) = \sum_{l=0}^{\inf}\sum_{m=-l}^{l} T_{lm}Y_{lm}(\theta,\phi), \tag{2.4.2}$$

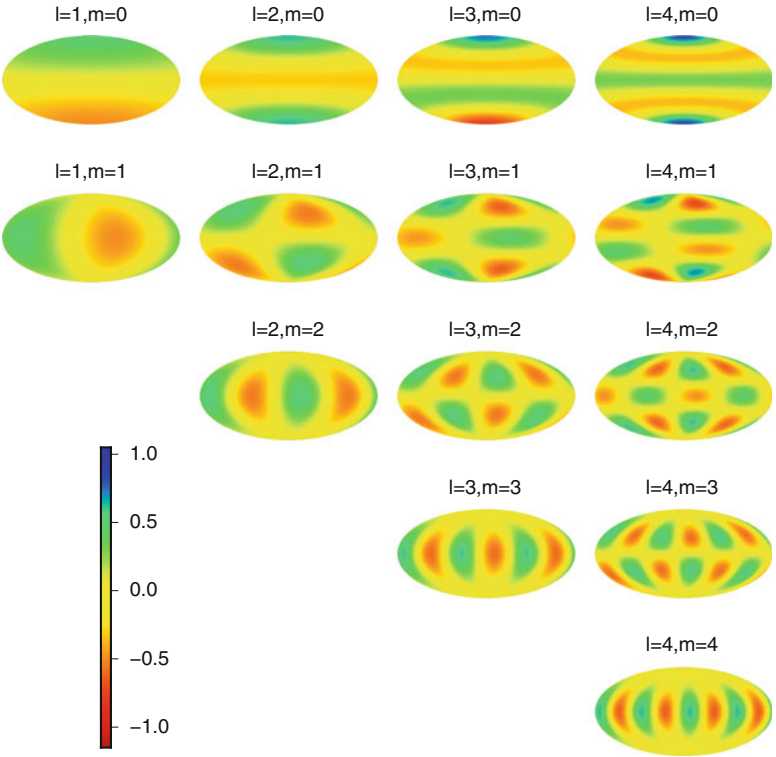


Fig. 2.4 Plots of normalized spherical harmonics for degree and orders 1 through 4

where the coefficients T_{lm} are independent of θ and ϕ. The spherical harmonic functions, normalized so that their integral over the sphere is one, for degrees 1 through 4 are plotted in Fig. 2.4. Spherical harmonics of order zero ($m = 0$) are only functions of latitude; they are often called zonal harmonics. Spherical harmonics with equal degree and order ($l = m$) are only functions of longitude; they are often called sectoral harmonics, because their pattern resembles the sections of an orange. The other harmonics are simply called mixed harmonics.

Two practical rules of thumb, which are useful when thinking about spherical harmonics: $P_{lm}(\cos\theta)$ has $(l-m)$ zero crossings between the North and South pole, while $\cos(m\phi)$ has $2m$ zero crossings between $0 \leq \phi \leq 2\pi$. Thus, when thinking about scaling properties represented in spherical harmonics on the surface of the Earth (6371 km), $\frac{2\pi \cdot 6371}{l-m} \approx \frac{40{,}000}{l-m}$ km, and for longitude, $\frac{2\pi \cdot 6371}{m} \approx \frac{40{,}000}{m}$ km.

Returning to the elastic response of a spherical body subject to a disc point load, we note that each set of boundary conditions defines a distinct Green's function and, thus, a different triplet of Love numbers. While the determination of the Love numbers is beyond the scope of this chapter (see [1, 5, 9, 20] for details), the Love numbers have a straight forward interpretation. If V_l is a single term in the spherical harmonic expansion of the gravitational potential V with a perturbation of degree l,

then $k_l V_l$ is the gravitational potential due to the elastic deformation within the Earth. In the spherically symmetric elastic deformation problem, the solution is comprised of only the zonal spherical harmonics ($m = 0$). Thus, the perturbation in the gravitational potential of degree l on the surface is the sum of the perturbation due to the applied mass, V_l, and the perturbation due to the new arrangement of matter within the Earth, $k_l V_l$. The quantity $h_l V_l / g$ is the radial displacement of the solid surface away from the reference spherical surface, $r = a$. When $h_l V_l / g$ is positive, the radius of the Earth's solid surface after the deformation is greater than the original radius a, and when $h_l V_l / g$ is negative, the radius the Earth's solid surface after the deformation is smaller than the original radius a. The Love number l_l, is related to tangential displacements and is not relevant to the vertical load problem.

To apply the Love numbers to the sea-level equation, starting with Eq. (2.3.2), the gravitational potential is expanded in terms of Legendre polynomials,

$$V(r, \theta) = \frac{a\, g}{M_E} \sum_{l=0}^{\infty} \left(\frac{a}{r}\right)^{l+1} P_l(\cos\theta), \tag{2.4.3}$$

where the P_l is the Legendre polynomial of order l. At the surface ($r = a$), the infinite series has the finite sum,

$$\sum_{l=0}^{\infty} P_l(\cos\theta) = \frac{1}{2\sin(\theta/2)}, \tag{2.4.4}$$

which implies the equivalence of Eqs. (2.3.2) and (2.4.3). While Eq. (2.4.4) is at first not obvious, it follows from the Legendra polynomial generating function,

$$\sum_{l=0}^{\infty} x^l P_l(\mu) = \frac{1}{(1 - 2\mu x + x^2)^{1/2}}, \tag{2.4.5}$$

with $x = 1$, $\mu = \cos\theta$, and the trigonometric identity $\sin(\theta/2) = \sqrt{(1 - \cos\theta)/2}$.

For each spherical harmonic degree l, $(1 + k_l)V_l$ is the perturbation potential on the spherical surface $r = a$ and $h_l V_l / g$ is the displacement of the solid boundary with respect to the reference surface, $r = a$. It follows that the perturbed gravitational potential on the displaced boundary of the solid Earth is $V_l^E = (1 + k_l + h_l)V_l$, because $-g(h_l V_l / g)$ is the change in the gravitational potential that occurs when moving from the reference surface ($r = a$) to the newly deformed boundary. The Green's function for the elastic problem can therefore be represented by

$$V_l^E = \frac{ag}{M_E} \sum_{l=0}^{\infty} (1 + k_l + h_l)\, P_l(\cos\theta), \tag{2.4.6}$$

and the solution for the sea level proceeds following the approach described for the rigid Earth in Sect. 2.3 [9].

2.5 Deformation of the Solid Earth: The Maxwell Rheology

While Sect. 2.4 illustrates the solution to the sea-level equation for a purely elastic Earth, elastic deformation is not consistent with the GIA observations. When the deforming force is removed, an elastic material instantaneously returns to the equilibrium state; therefore, the elastic deformation model predicts that the Earth's surface would have returned to the equilibrium state as the ice sheets melted and today no deformation due to GIA should be expected. Hence, some degree of viscous behavior is required to explain the GIA observations.

While there are several possible models for viscoelastic materials, the *Maxwell rheology model* is the one that is predominantly used in GIA studies. In the context of Maxwell rheology, the sea-level equation becomes time dependent.

Figure 2.5 shows a simple representation of a Maxwell solid as a purely viscous damper connected in series with a purely elastic spring. Under an applied axial stress, the total stress, σ_{Total}, and the total strain, $\varepsilon_{\text{Total}}$, are defined as follows:

$$\sigma_{\text{Total}} = \sigma_D = \sigma_S, \tag{2.5.1}$$

$$\varepsilon_{\text{Total}} = \varepsilon_D + \varepsilon_S. \tag{2.5.2}$$

The subscript D refers to the damper (viscous deformation), the subscript S to the spring (elastic deformation). Taking the derivative of strain with respect to time, we obtain

$$\frac{d\varepsilon_{\text{Total}}}{dt} = \frac{d\varepsilon_D}{dt} + \frac{d\varepsilon_S}{dt} = \frac{\sigma}{\eta} + \frac{1}{E}\frac{d\sigma}{dt}, \tag{2.5.3}$$

where E is the elastic modulus and η the viscosity.

Calculating sea-level changes on a viscoelastic Earth requires a Green's function for the perturbation to the gravitational potential, which depends on both the distance from a point mass and the time that has elapsed since the mass was applied to the Earth's surface. The Green's function contains all the necessary information relating to the rheological structure of the Earth. Green's functions for a range of Maxwell Earth models were determined by Peltier [33], who used the correspondence principle in conjunction with classical elastodynamics.

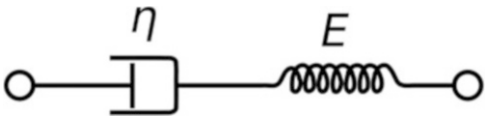


Fig. 2.5 The Maxwell rheology model, a linear spring and viscous dashpot in series

2.6 Discussion

While there has been significant progress in understanding the response of the solid Earth and sea level due to changes in the distribution of ice and water over the surface of the Earth, challenges remain. A comprehensive overview of the processes that affect sea level can be found in a recent article by Cazenave and Nerem [6]. Other factors that could affect sea-level changes can be grouped into four broad categories: (1) changes in the volume of the ocean basin as the result of a change in the topography of the ocean floor; (2) changes in the shape of the gravitational potential; (3) local subsidence or uplift of the land-sea interface; and (4) changes in the total volume of water in the ocean basin. We briefly discuss some details of each category.

Changes in the Volume of the Ocean Basin A change in the volume of the ocean basin will occur when tectonic plates reorganize or when there is a change in the velocity of an oceanic plate. This is because the ocean-floor topography, which geoscientists call *bathymetry*, is controlled by the conductive cooling of the oceanic plates [31]. A half-space that cools due to conduction has a square-root of time functional form. Parsons and Sclater [31] show that ocean bathymetry should increase with the square root of age of the ocean flow, where the new crust that forms at a mid-ocean ridge defines time zero. If the ocean plate moves away from a mid-ocean ridge faster, then the bathymetry will be shallower at a fixed distance from the ridge. If the ocean plate moves slower, then the bathymetry will be deeper at the same distance. Thus, if the velocity of an oceanic plate increases, the volume of the ocean will decrease (over time) and sea level will increase (all other factors being equal), while if the velocity of an oceanic plate decreases, the volume of the ocean will increase and sea level will fall. This change in shape of the sea floor changes on a time scale of millions to tens of millions of years, significantly longer than the time scale of GIA, and does not impact current estimates of sea level.

Changes in the Shape of the Gravitational Potential A change in the shape of the gravitational potential is controlled by the time scale of the redistribution of mass. The distribution of ice/water on the surface of the Earth (Sect. 2.3) is already included in the GIA analysis. On the other hand, the erosion of rock by ice sheets, which generates material that is then incorporated into the ice and subsequently deposited as the ice melts, is not accounted for in current GIA models. Ice sheets transport eroded material away from the center of glaciation to the continental shelf edge, where it is deposited in a series of fans. Nygård et al. [28] estimate that 32,000 km^3 of sediment have been deposited on the North Sea Fan off the west coast of Norway over the last 450 Myr. This time span is significantly longer than the glacial–interglacial cycle, so the accumulation of sediment from a single glacial–interglacial cycle is probably significantly less. Even so, the total mass of sediment is small compared to the estimate of 5×10^6 km^3 of ice that was loaded onto (and then removed from) the Fennoscandian platform in a period of roughly 100,000 years [18]. The effects of other changes such as anthropogenic groundwater

pumping or climate-related processes have been shown to be small compared with the GIA signal [19, 36].

Local Subsidence or Uplift of the Land-Sea Interface Local subsidence and uplift of the land surface that are not due to GIA are monitored by carefully tying the tide-gauge measurements into a global geodetic reference frame. In this regard, the development of high-precision GPS measurements has significantly reduced one of the major sources of uncertainty in the GIA observations. In addition, because GPS can measure uplift and horizontal motions over land, this has significantly expanded the range of GIA observations [13, 21, 22, 27, 30, 39, 42].

Changes in the Total Volume of Water in the Ocean Basin One of the effects of anthropogenic climate change is that the large ice sheets over Greenland and Antarctic are melting at increasingly faster rates. The resulting changes in sea level have been shown to possess a unique pattern indicating where the most active melting is currently taking place [25, 41]. This melting will change the volume of water in the ocean basin and will change the shape of the gravitational potential because of the redistribution of the ice/water, following the same logic as described in Sect. 2.3. In principle, this can be modeled with the same analysis used to study the longer-time scale processes associated with the melting of glacial–interglacial ice sheets. The challenge is unraveling the ongoing effect of GIA from the last glacial cycle with the present-day changes in ice/water due to current melting of the Greenland and Antarctic ice sheets.

Two significant areas of uncertainty in the analysis of GIA observations are the distribution of ice at the last glacial maximum [34] and the rheology of the mantle [16]. Mantle rheology impacts not only the response of the surface to imposed surface loads such as ice and water, but also controls the motion of the tectonic plates [10] and regulates the heat flow from within the Earth by controlling the vigor of convection within the solid Earth [37]. While the viscosity of minerals that make up the mantle is strongly dependent on temperature and pressure and could vary with grain size and strain rate, depending on the deformation mechanism [16], many GIA studies focus on depth-dependent viscosity profiles. While viscosity models other than depth-dependent models have been considered [32, 40, 44, 46], the primary control on vertical surface motion is from depth-dependent rheology, which is by far the rheology that has been given the most attention. Lateral variations in rheology are likely to be most prevalent at the boundary between continents and oceans [15, 17], which is where the sea-level observations are made.

The ongoing deformation of the solid Earth and the associated change in sea level in response to the glacial cycle are challenging interdisciplinary problems with a strong historical connection to the mathematical community. Advancing our understanding would benefit from new data assimilation and modeling strategies, improvements in viscoelastic modeling, a better understanding of mantle rheology, and a more complete understanding of present-day changes in ice load.

Acknowledgements The author acknowledges support from NSF Grant EAR-1250988.

References

1. Backus, G.E.: Converting vector and tensor equations to scalar equations in spherical coordinates. Geophys. J. **13**, 71–79 (1967)
2. Batchelor, G.K.: An Introduction to Fluid Dynamics. Cambridge University Press, Cambridge (1967)
3. Bean, W.B.: Nail growth. Thirty-five years of observation. Arch. Intern. Med. **140**, 73–76 (1980)
4. Braun, J.: The many surface expressions of mantle dynamics. Nature Geosc. **3**, 825–833 (2010)
5. Cathles, L.M.: The Viscosity of the Earth's Mantle. Princeton Univ. Press, Princeton (1975)
6. Cazenave, A., Nerem, R.S.: Present-day sea level change: observations and causes. Rev. Geophys. **42**(RG3001) (2004)
7. Ekman, M.: The world's longest continued series of sea-level observations. Pure Appl. Geophys. **127**, 73–77 (1988)
8. England, P.C., Houseman, G.A.: Finite strain calculations of continental deformation II. Comparison with the India-Asia collision zone. J. Geophys. Res. **91**, 3664–3676 (1986)
9. Farrell, W.E., Clark, J.A.: On postglacial sea level. Geophys. J. R. Astron. Soc. **46**(3), 647–667 (1976)
10. Forte, A.M., Peltier, W.R., Dziewoński, A.M.: Inferences of mantle viscosity from tectonic plate velocities. Geophys. Res. Lett. **18**, 1747–1750 (1991)
11. Haskell, N.A.: The motion of a viscous fluid under a surface load. Physics **6**, 265–269 (1935)
12. Haskell, N.A.: The motion of a viscous fluid under a surface load, part 2. Physics **7**, 56–61 (1936)
13. Johansson, J.M., Davis, J.L., Scherneck, H.G., et al.: Continuous GPS measurements of postglacial adjustment in fennoscandia – 1. Geodetic results. J. Geophys. Res. **107**, 2157 (2002)
14. Johnson, A.M., Fletcher, R.C.: Folding of Viscous Layers. Columbia University Press, New York (1994)
15. King, S.D.: Archean cratons and mantle dynamics. Earth Planet. Sci. Lett. **234**, 1–14 (2005)
16. King, S.D.: Reconciling laboratory and observational models of mantle rheology in geodynamic modeling. J. Geodyn. **100**, 33–50 (2016)
17. King, S.D., Anderson, D.L.: Edge driven convection. Earth Planet. Sci. Lett. **160**, 289–296 (1998)
18. Lambeck, K., Yokoyama, Y., Johnston, P., et al.: Global ice volumes at the Last Glacial Maximum and early lateglacial. Earth Planet. Sci. Lett. **181**, 513–527 (2000)
19. Landerer, F.W., Swenson, S.C.: Accuracy of scaled GRACE terrestrial water storage estimates. Water Resource Res. **48**(W04531) (2012). https://doi.org/10.1029/2011WR011453
20. Love, A.E.H.: The stress produced in a semi-infinite solid by pressure on part of the boundary. Phil. Tran. Roy. Soc. London, Ser. A **228**, 377–379 (1929)
21. Mazzotti, S., Lambert, A., Henton, J., et al.: Absolute gravity calibration of GPS velocities and glacial isostatic adjustment in mid-continent North America. Geophys. Res. Lett. **38**, L24311 (2011). https://doi.org/10.1029/2011GL049846
22. Milne, G.A., Davis, J.L., Mitrovica, J.X., et al.: Space-geodetic constraints on glacial isostatic adjustment inFennoscandia. Science **291**, 2381–2385 (2001)
23. Milne, G.A., Mitrovica, J.X., Scherneck, H.G.: Estimating past continental ice volume from sea-level data. Quat. Sci. Rev. **21**, 361–376 (2002)
24. Mitrovica, J.X., Milne, G.A.: On post-glacial sea level: I. general theory. Geophys. J. Int. **154**, 253–267 (2003)
25. Mitrovica, J.X., Tamisiea, M.E., Davis, J.L., et al.: Recent mass balance of polar ice sheets inferred from patterns of global sea-level change. Nature **409**, 1026–1029 (2001)
26. Morgan, W.J.: Rises, trenches, great faults, and crustal blocks. J. Geophys. Res. **73**, 1959–1982 (1968)
27. Nocquet, J.M., Calais, E., Parsons, B.: Geodetic constraints on glacial isostatic adjustment in Europe. Geophys. Res. Lett. **32**, L06308 (2005)

28. Nygård, A., Sejrup, H.P., Haflidason, H., et al.: The glacial North Sea fan, southern Norwegian Margin: architecture and evolution from the upper continental slope to the deep-sea basin. Mar. Pet. Geol. **22**, 71–84 (2005)
29. Olver, F.W.J., Olde Daalhuis, A.B., Lozier, D.W., et al. (eds.): NIST Digital Library of Mathematical Functions. http://dlmf.nist.gov/. Release 1.0.19 of 2018-06-22
30. Park, K.D., Nerem, R.S., Davis, J.L., et al.: Investigation of glacial isostatic adjustment in the northeast US using GPS measurements. Geophys. Res. Lett. **29**, 1509–1512 (2002)
31. Parsons, B., Sclater, J.G.: An analysis of the variation of ocean floor bathymetry and heat flow with age. J. Geophys. Res. **82**, 803–827 (1977)
32. Paulson, A., Zhong, S., Wahr, J.: Modelling post-glacial rebound with lateral viscosity variations. Geophys. J. Int. **163**, 357–371 (2005)
33. Peltier, W.: Impulse response of a Maxwell Earth. Rev. Geophys. **12**, 649–669 (1974)
34. Peltier, W.R., Argus, D.F., Drummond, R.: Space geodesy constrains ice-age terminal deglaciation: The global ICE-6G_C (VM5a) model. J. Geophys. Res. **120**, 450–487 (2015)
35. Peltier, W.R., Tushingham, A.M.: Influence of glacial isostatic-adjustment on tide-gauge measurements of secular sea-level change. J. Geophys. Res. **96**, 6779–67,960 (1991)
36. Ramillien, G., Bouhours, S., Lombard, A., et al.: Land water storage contribution to sea level from GRACE geoid data over 2003–2006. Global Planet. Change **60**, 381–392 (2008)
37. Schubert, G., Turcotte, D.L., Olson, P.: Mantle Convection in the Earth and Planets. Cambridge University Press, Cambridge (2001)
38. Sella, G.F., Dixon, T.H., Mao, A.: REVEL: A model for recent plate velocities from space geodesy. J. Geophys. Res. **107**(B4), ETG 11-1–ETG 11-30 (2002). https://doi.org/10.1029/2000JB000033
39. Sella, G.F., Stein, S., Dixon, T.H., et al.: Observation of glacial isostatic adjustment in "stable" North America with GPS. Geophys. Res. Lett. **34**, L02306 (2007). https://doi.org/10.1029/2006GL027081
40. Spada, G., Antonioli, A., Cianetti, S., et al.: Glacial isostatic adjustment and relative sea-level changes: the role of lithospheric and upper mantle heterogeneities in a 3-d spherical earth. Geophys. J. Int. **165**, 692–702 (2006)
41. Tamisiea, M.E., Mitrovica, J.X., Milne, G.A., et al.: Global geoid and sea level changes due to present-day ice mass fluctuations. J. Geophys. Res. **106**, 30,849–30,863 (2001)
42. Tamisiea, M.E., Mitrovica, J.X., Davis, J.L.: GRACE gravity data constrain ancient ice geometries and continental dynamics over Laurentia. Science **5826**, 881–883 (2007)
43. van Veen, J.: Bestaat er een geologische bodemdaling te Amsterdam sedert 1700? Tijdschrift Koninklijk Nederlandsch Aardrijkskundig Genootschap **2: LXII** (1945)
44. Wang, H., Wu, P.: Effects of lateral variations in lithospheric thickness and mantle viscosity on glacially-induced surface motion on a spherical, self-gravitating Maxwell Earth. Earth Planet. Sci. Lett. **244**, 576–589 (2006)
45. Wilson, J.T.: Did the Atlantic close and then reopen? Nature **211**, 676–681 (1966)
46. Wu, P.: Mode coupling in a viscoelastic self-gravitating spherical earth induced by axisymmetric loads and lateral viscosity variations. Earth Planet. Sci. Lett. **197**, 1–10 (2002)

Chapter 3
Mathematical Challenges in Measuring Variability Patterns for Precipitation Analysis

Maria Emelianenko and Viviana Maggioni

Abstract This chapter addresses some of the mathematical challenges associated with current experimental and computational methods to analyze spatiotemporal precipitation patterns. After a brief overview of the various methods to measure precipitation from *in situ* observations, satellite platforms, and via model simulations, the chapter focuses on the statistical assumptions underlying the most common spatiotemporal and pattern-recognition techniques: stationarity, isotropy, and ergodicity. As the variability of Earth's climate increases and the volume of observational data keeps growing, these assumptions may no longer be satisfied, and new mathematical methodologies may be required. The chapter discusses spatiotemporal decorrelation measures, a nonstationary intensity-duration-function, and 2-dimension reduction methodologies to address these challenges.

Keywords Centroidal Voronoi tessellation · Data reduction · Decorrelation · Empirical orthogonality functions · Ergodicity · Isotropy · Precipitation patterns · Stationarity · Statistical assumptions

3.1 Introduction

Precipitation occurs when a portion of the atmosphere becomes saturated with water vapor, so that the water condenses and precipitates by gravity. Precipitation is a critical component of the water and energy cycles, providing moisture for processes such as runoff, biogeochemical cycling, evapotranspiration, groundwater recharge, carbon exchange, and heat fluxes. The main forms of precipitation include rain,

M. Emelianenko (✉)
Department of Mathematical Sciences, George Mason University, Fairfax, VA, USA
e-mail: memelian@gmu.edu

V. Maggioni
Sid and Reva Dewberry Department of Civil, Environmental, and Infrastructure Engineering, George Mason University, Fairfax, VA, USA
e-mail: vmaggion@gmu.edu

H. G. Kaper, F. S. Roberts (eds.), *Mathematics of Planet Earth*, Mathematics of Planet Earth 5, https://doi.org/10.1007/978-3-030-22044-0_3

sleet, snow, and hail, but this chapter discusses liquid precipitation only, and the term "precipitation" is used here as a synonym for "rain."

Precipitation is highly variable, both in space and time. This variability affects the dynamics of many hydrological processes at and near ground level. Information on precipitation characteristics and precipitation patterns is therefore critical for understanding these complex hydrological processes, as well as for monitoring and predicting extreme events such as floods and droughts [63]. Access to high-resolution high-quality rainfall data and information about spatiotemporal precipitation patterns can benefit applications at all levels; examples are hazard mitigation, agricultural planning, and water resources management at the regional level [33, 37, 46]; controlling stormwater runoff, managing reservoirs and detention ponds, cleaning streams and channels, and closing roads or parking lots during extreme precipitation events at the local level.

However, estimating precipitation is challenging because it involves many factors, including the natural temporal and spatial variability of precipitation, measurement errors, and sampling uncertainties, especially at fine temporal and spatial scales. The spatiotemporal variability of precipitation patterns is changing heterogeneously due to climate change, and those changes have an impact on the tools used to make decisions and optimize water management. This chapter focuses on some of the mathematical and statistical issues related to variability of precipitation patterns.

Outline of the Chapter In Sect. 3.2, we briefly discuss various methods to measure precipitation, whether *in situ*, remotely, or by using model simulations. In Sect. 3.3, we review the strengths and limitations of current methods to analyze spatiotemporal precipitation patterns. We discuss decorrelation measures in Sect. 3.4 and dimension reduction strategies in Sect. 3.5. In Sect. 3.6, we present some concluding remarks.

3.2 Estimating Precipitation

Precipitation can be estimated through three main approaches: (1) *in situ* measurements, (2) remote sensing (including weather radars and satellite sensors), and (3) model simulations [52].

3.2.1 In Situ *Measurements*

The only direct method to measure precipitation is through rain gauges (also known as *pluviometers*) which collect and measure the amount of rain over a period of time. There are several types of rain gauges; the most common one is the tipping bucket. Precipitation is collected in a funnel and channeled into a small container. After a set amount of precipitation is collected, the device tips, dumping the water,

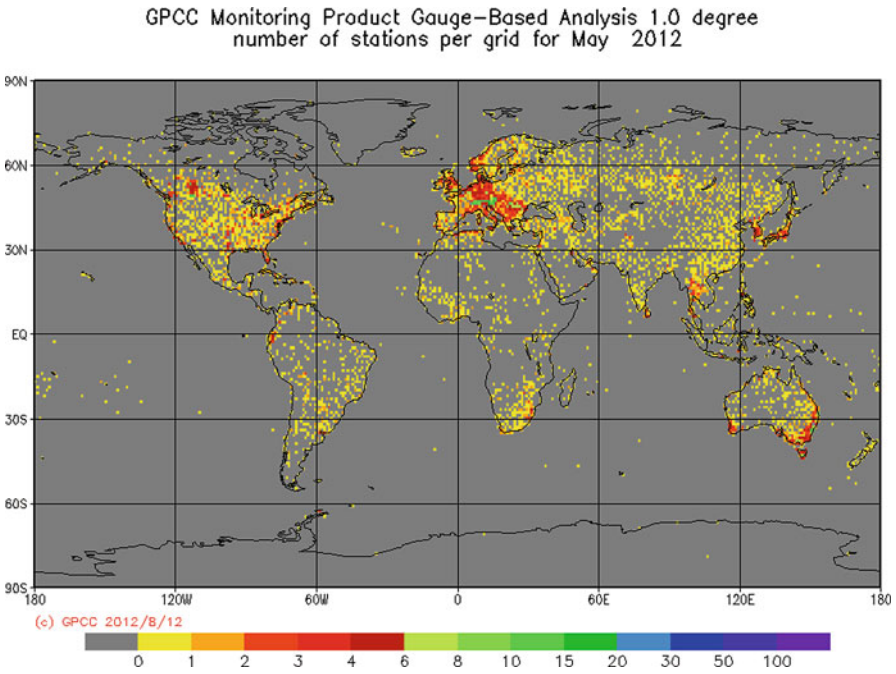


Fig. 3.1 Number of stations used by the global precipitation climatology center (GPCC) for May 2012. Figure produced with "GPCC Visualizer" [61], courtesy of National Center for Atmospheric Research Staff (Eds), last modified 29 Jun 2018. Retrieved from https://climatedataguide.ucar.edu/climate-data/gpcc-global-precipitation-climatology-centre

and sending a signal that is automatically recorded by a data logger. Rain gauges may underestimate rainfall because of wind effects and evaporation.

Rain-gauge networks can provide measurements with high temporal resolution, but obtaining a spatially representative measurement requires a sufficiently large number of samples to account for variability of terrain, microclimate, and vegetation. Moreover, *in situ* measurements are localized and limited in spatial and temporal coverage [43]. One of the main applications of ground-monitoring networks is for assessing flood risk through early warning systems [3]. However, their usefulness is limited by the spatial representativeness of local measurements and the network density, especially over important climatic regions like the tropical rain forests and mountainous areas (Fig. 3.1).

A ground-based alternative to monitor precipitation is weather radar which provides spatially distributed information on rainfall (Fig. 3.2). Weather radars send directional pulses of microwave radiation connected to a parabolic antenna. Wavelengths are of the order of a few centimeters, which is about ten times larger than the average diameter of water droplets and ice particles. These particles bounce part of the energy in each pulse back to the radar (reflectivity). As they move farther from the source, the pulses spread out, crossing a larger volume

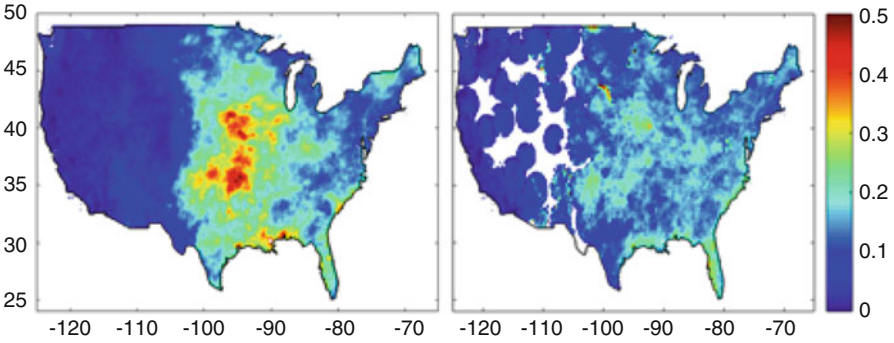


Fig. 3.2 Average precipitation maps for Summer 2015 from a satellite precipitation product that combines infrared and microwave observations (left) and ground-based weather radars (right) across the continental USA

of air, and therefore their resolution decreases with distance. Doppler radars are common and observe not only rainfall rates, but also the motion of rain droplets. However, weather radar estimates are affected by uncertainties associated with rain-path attenuation, the lack of uniqueness in the reflectivity-to-rain-rate relationship, radar calibration and contamination by ground return problems, sub-resolution precipitation variability, and complex terrain effects [10, 46, 51]. Moreover, ground-based monitoring systems, like rain gauges and weather radars, require substantial financial and technological investments to support their operation and maintenance on a continuous basis over a long period.

3.2.2 Remote Sensing

A way to overcome these issues is the use of satellite precipitation products, which are nowadays available on a global scale at increasing spatial and temporal resolution. Precipitation estimates can be derived from a range of observations from many different on-board satellite sensors. Specifically, rainfall can be inferred from visible imagery, since thick clouds, which are more likely to be associated with rainfall, tend to be brighter than the surface of the Earth. Infrared (IR) images are more suitable because they are available night and day, and heavier convective rainfall tends to be associated with larger taller clouds with colder cloud tops. Another method uses passive microwave (PMW) sensors, since emissions from rain droplets lead to an increase in PMW radiation. And scattering caused by precipitating ice particles leads to a decrease in PMW radiation.

Several techniques have been developed to exploit the synergy between IR radiances and PMW observations (Fig. 3.2). Examples include the TRMM multi-satellite precipitation analysis (TMPA) [41], the climate prediction center morphing (CMORPH) technique [42], and, most recently, the integrated multi-satellite retrievals for GPM (global precipitation measurement) (IMERG) [40], which

merges precipitation estimates from PMW and IR sensors and monthly surface rain-gauge data to provide half-hourly precipitation estimates on a 0.1° grid over the 60°N-S domain. In other cases, artificial neural networks (ANNs) are used to derive precipitation estimates by combining information from multichannel and multisensor observations, like the precipitation estimation from remotely sensed information using ANNs (PERSIANN) [39]. The availability of these products has opened new venues to support water management and hydrologic applications globally. Especially in poorly gauged regions, satellite precipitation products may be the only input data to allow flow predictions downstream with enough lead time to implement management and response actions [64].

Satellite observations can be affected by detection errors, as well as systematic and random errors. Detection errors include missed events (when satellite observes no rain, but there is rain at the ground) and false alarms (when the satellite sees rain, but it does not rain). In the case of successful detection, the estimated rain rate may still be affected by systematic and/or random errors, which depend on the accuracy of the remote sensor (retrieval error) and the lack of continuity in the coverage by low earth-orbiting satellites (sampling error, [7]). Typical sources of retrieval error are due to sub-pixel inhomogeneity in the rainfall field [48], whereas sampling errors are related to the satellite orbit, swath width, and space-time characteristics of rainfall [14]. The performance of satellite precipitation products is also influenced by factors such as seasonal precipitation patterns, storm type, and background surface [31, 33, 57, 66]. Detection, systematic, and random errors all play a pivotal role in hydrological applications (e.g., flood forecasting) and water resource management.

High-mountain regions are among the most challenging environments for precipitation measurements (whether from the ground or from satellites) due to extreme topography and large weather and climate variability. These regions are typically characterized by a lack of *in situ* data, but are also prone to flash floods whose consequences can be devastating.

3.2.3 Model Simulation

Numerical weather prediction (NWP) models provide a third option for estimating precipitation at global and regional scales. NWP models estimate the state of the atmosphere (including air density, pressure, temperature, and velocity) at a given time and location using fluid dynamics and thermodynamics equations. These models are rather accurate for large-scale organized systems. However, their performance deteriorates in the case of more localized events that are not governed by large-scale flows and whose spatial and temporal variability cannot be explicitly captured by the model resolution. NWP model forecasts can be improved by more accurate parameterizations and by constraining model analyses with moisture-, cloud-, and precipitation-related observations through data assimilation systems, such as 4D-Var and ensemble Kalman filter methods [6, 50].

3.3 Assessment of Spatial and Temporal Patterns

Changes in spatiotemporal precipitation patterns have a direct impact on the spatial and temporal distribution of water resources and the occurrence of natural hazards [69]. The hydrological community has adopted a set of geostatistical tools for measuring spatiotemporal correlations in rainfall [5, 65]. As mentioned in multiple sources [65], some of the notions come with tacit assumptions that often lead to their misuse in practice. While the complete list is beyond the scope of this chapter, we will review some of the key elements of such analyses and point out some of their strengths and limitations.

3.3.1 Definitions

Assume that rainfall corresponds to a stochastic process $\eta(\mathbf{u}, t)$, where $\mathbf{u} \equiv (x, y)$ is a vector representing the spatial coordinates in a given area, t stands for time, and $\eta(\cdot)$ is a measure of the intensity of the rainfall. In a practical setting, one typically considers an observation map in the form of a *snapshot matrix* $A = A_{i,j} \in \mathbb{R}^{N\times n}$, where $A_{i,j} = \eta(\mathbf{u}_i, t_j)$ is the rainfall observed at location i at time t_j ($i = 1, \dots, N$; $j = 1, \dots, n$). Typically, for hydrological applications, $N \gg n$. Different statistical characterizations of the process are used, depending on the purpose of the study.

Spatial Variability If the focus is on spatial correlations, time series may be integrated over time at each location. Following [5], we define the *depth*, Z, of the rainfall over a time interval of length T at the location $\mathbf{u}$, by the integral

$$Z(\mathbf{u}) = \int_t^{t+T} \eta(\mathbf{u}, \tau)\, d\tau, \tag{3.3.1}$$

and its *intensity*, X, by the integral

$$X(\mathbf{u}) = \frac{1}{T} \int_t^{t+T} \eta(\mathbf{u}, \tau)\, d\tau. \tag{3.3.2}$$

The *mean*, m, of the rainfall at $\mathbf{u}$ is

$$m(\mathbf{u}) = E[Z(\mathbf{u})], \tag{3.3.3}$$

where $E[\cdot]$ denotes the expected value over all realizations of the process—that is, over all different measurements at a certain location. After subtracting the mean, we obtain the *detrended* or *centered process*, Y,

$$Y(\mathbf{u}) = Z(\mathbf{u}) - m(\mathbf{u}). \tag{3.3.4}$$

The *covariance function* is defined in terms of the detrended process.

$$\mathrm{Cov}(\mathbf{u}_1, \mathbf{u}_2) = E[Y(\mathbf{u}_1)Y(\mathbf{u}_2)] = E[Z(\mathbf{u}_1)Z(\mathbf{u}_2)] - m(\mathbf{u}_1)m(\mathbf{u}_2). \tag{3.3.5}$$

Similarly, the *covariance matrix* is $\Sigma = E[Y^T Y] = E[Z^T Z] - m^T m$, where the (i, j)th entry represents the covariance between the depth of rainfall at the ith and jth spatial location. The *correlation function* is a normalized version of the covariance function,

$$R(\mathbf{u}_1, \mathbf{u}_2) = \frac{\mathrm{Cov}(\mathbf{u}_1, \mathbf{u}_2)}{\sigma(\mathbf{u}_1)\sigma(\mathbf{u}_2)}, \tag{3.3.6}$$

where σ is the *standard deviation*,

$$\sigma(\mathbf{u}_i) = \left\{E[Z(\mathbf{u}_i) - m(\mathbf{u}_i)]^2\right\}^{1/2} = E[Y^2(\mathbf{u}_i)]^{1/2}. \tag{3.3.7}$$

A concept that is commonly used in hydrology is that of a *semivariogram function*,

$$\Gamma(\mathbf{u}_1, \mathbf{u}_2) = \frac{1}{2}E\{[Y(\mathbf{u}_1) - Y(\mathbf{u}_2)]^2\}. \tag{3.3.8}$$

The covariance and semivariogram functions are symmetric,

$$\mathrm{Cov}(\mathbf{u}_1, \mathbf{u}_2) = \mathrm{Cov}(\mathbf{u}_2, \mathbf{u}_1), \qquad \Gamma(\mathbf{u}_1, \mathbf{u}_2) = \Gamma(\mathbf{u}_2, \mathbf{u}_1).$$

Note that the covariance is a measure of the association between the two variables $Z(\mathbf{u}_1)$ and $Z(\mathbf{u}_2)$, while the semivariogram function is a measure of their dissociation.

The above definitions of the various statistical quantities work for any time interval $[t, t + T]$. For instance, one may decide to study daily, monthly, or yearly averages, as appropriate. The longer the period over which the data are integrated, the more one may expect temporal variations to be suppressed.

Temporal Variability If temporal variability is of interest, it is important to keep as much of the original temporal information as possible when computing variograms and correlations. So, while integrated data are attractive from the processing point of view, in climate research one always defines statistical characteristics using the original map $\eta(\mathbf{u}, t)$. Thus, the *mean* is defined as a time average,

$$m(\mathbf{u}) = \langle \eta(\mathbf{u}, t) \rangle, \tag{3.3.9}$$

and the *centered data* (also called *anomalies*) are given by

$$Y = Y(\mathbf{u}, t) = \eta(\mathbf{u}, t) - m(\mathbf{u}). \tag{3.3.10}$$

The correlation function, standard deviation, and semivariogram are defined in terms of anomalies as in (3.3.6)–(3.3.8).

The statistical quantities defined above all have discrete analogs. For example, $m_i = \frac{1}{n}\sum_{j=1}^{n} A_{i,j}$ is the time average of a certain realization of the rainfall field at location i; the anomalies $y_{i,j} = a_{i,j} - m_i$ are the entries of the *anomaly matrix* $Y = y_{i,j}$, and the corresponding *covariance matrix* is $\Sigma = Y^T Y \in \mathbb{R}^{n\times n}$. The eigenvectors of this covariance matrix Σ are the *empirical orthogonal functions*, which we will discuss in Sect. 3.5. Note that, while the size of the matrix Σ is normally much smaller than that of the original detrended matrix Y, the condition of the covariance matrix is given by $\mathrm{cond}(\Sigma) = \mathrm{cond}(Y)^2$, so it is not surprising that ill-conditioning in the original data presents an issue for many geospatial applications [1].

The correlation function, standard deviation, and semivariogram function are collectively referred to as *variograms* of the process; they represent the structure of the spatial dependence of the process and variability in the reference area A.

3.3.2 Statistical Assumptions in Hydrological Analyses

The effective use of the statistical quantities defined in Sect. 3.3.1 depends critically on a number of regularity assumptions for the underlying stochastic process. In hydrological analyses, the rainfall process is commonly assumed to be second-order stationary, isotropic, and ergodic. We briefly recall the relevant definitions.

Stationarity The field $Z(\mathbf{u})$ is *first-order stationary* if

$$E[Z(\mathbf{u})] = m = \text{constant}, \quad \forall \mathbf{u} \in A, \tag{3.3.11}$$

and *second-order stationary* or *weakly stationary* if it is first-order stationary and, in addition,

$$\mathrm{Var}[Z(\mathbf{u})] = \sigma^2 = \text{constant}, \quad \forall \mathbf{u} \in A, \tag{3.3.12}$$

$$\mathrm{Cov}(\mathbf{u}_1, \mathbf{u}_2) = \mathrm{Cov}(\mathbf{u}_1 - \mathbf{u}_2), \quad \forall \mathbf{u}_2, \mathbf{u}_2 \in A. \tag{3.3.13}$$

For a second-order stationary process, $\Gamma(\mathbf{u}_1, \mathbf{u}_2) = \Gamma(\mathbf{u}_1 - \mathbf{u}_2) = \Gamma(\mathbf{h})$, where $\mathbf{h} = \mathbf{u}_1 - \mathbf{u}_2$, and $\mathrm{Cov}(\mathbf{u}_1, \mathbf{u}_2) = \mathrm{Cov}(\mathbf{h}) = E[Z(\mathbf{u}+\mathbf{h})Z(\mathbf{u})] - m^2$ for all $\mathbf{u}_2$, $\mathbf{u}_2 \in A$. Furthermore,

$$\begin{aligned}\Gamma(\mathbf{h}) &= \tfrac{1}{2}E[Z(\mathbf{u}-\mathbf{h}) - Z(\mathbf{u})]^2 \\ &= \tfrac{1}{2}E[Z(\mathbf{u}+\mathbf{h})^2] - E[Z(\mathbf{u})Z(\mathbf{u}+\mathbf{h})] + \tfrac{1}{2}E[Z(\mathbf{u})]^2 \\ &= \mathrm{Cov}(\mathbf{0}) - \mathrm{Cov}(\mathbf{h}).\end{aligned}$$

Isotropy The field $Z(\mathbf{u})$ is *isotropic* if spatial variability, measured by the covariance or semivariogram function, does not depend on the direction of the vector $\mathbf{h} = \mathbf{u}_1 - \mathbf{u}_2$,

$$\mathrm{Cov}(\mathbf{h}) = \mathrm{Cov}(|\mathbf{h}|) = \mathrm{Cov}(h), \tag{3.3.14}$$

$$\Gamma(\mathbf{h}) = \Gamma(|\mathbf{h}|) = \Gamma(h), \tag{3.3.15}$$

where $h = |\mathbf{h}|$ is the distance between two locations $\mathbf{u}_1$ and $\mathbf{u}_2$.

Ergodicity A dynamic process is said to be *ergodic* if time averages coincide with sample averages,

$$E(\eta(\mathbf{u}, t)) = \langle \eta(\mathbf{u}, t) \rangle. \tag{3.3.16}$$

In the case of an ergodic process, the estimates of the moments obtained on the basis of the available realizations converge in probability to the theoretical moments as the sample size increases. The process will tend to a limiting distribution regardless of the initial state [44]. In practice, this enables one to obtain estimates even from a single realization of the process.

3.3.3 What If the Assumptions Are Not Satisfied?

Figure 3.3 shows a realization of the precipitation process. The data (blue dots) represent the annual maximum precipitation (in inches) recorded at Beardstown in the State of Illinois (USA) during the period 1903–2000. Connecting the dots, we see that the maximum moves up and down without much regularity, but a linear regression analysis shows an overall upward trend (solid blue line). The mean (solid purple line) is approximately 2.3 in. over the first 55 years (1903–1958) and approximately 2.8 in. over the next 42 years (1958–2000), an increase of more than 20%. The variance (dotted red lines) also increases over time, albeit more slowly. The example shows that the rainfall process is clearly not stationary, so at least one of the hypotheses discussed in Sect. 3.3.2 is violated. Then the question is, what to do?

Nonstationarity The paper by Milly et al. [53] entitled *Stationarity Is Dead: Whither Water Management?*, which appeared in Science in 2008, served as a wake-up call for scientists in the field of hydrology and water resources engineering. Water management systems have been designed and operated for decades under the assumption of stationarity. However, this assumption has long been compromised by human disturbances in river basins such as dams, diversions, irrigation, land-use change, channel modifications, and drainage work. In addition, the timing and characteristics of precipitation—the most critical hydrological input—are also being modified by a changing climate, as demonstrated in Fig. 3.3. The hydrological

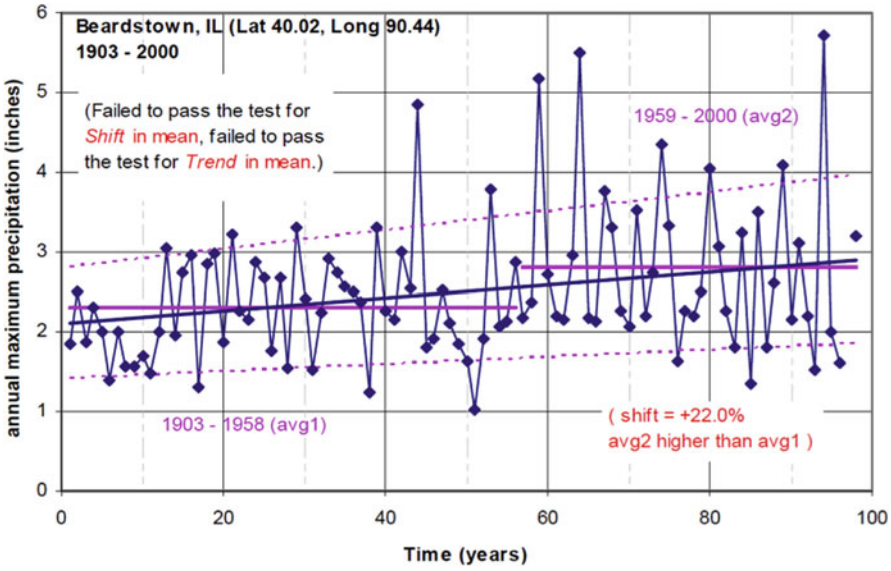


Fig. 3.3 Example of nonstationary changes in both the precipitation mean and variance [9]

literature on the analysis of long-term precipitation and runoff data is very thin [35]. Hodgkins and Dudley [36] show that most North American streams are experiencing earlier spring runoff, and DeGaetano et al. [24] show that nearly two-thirds of the trends in the 2-, 5-, and 10-years return-period rainfall amounts are positive. At the same time, the expected recurrence intervals have decreased by about 20%; for example, the 50 year storm based on 1950–1979 data is expected to occur once every 40 years based on 1950–2007 data.

Nonstationarity introduces multiple challenges for hydrological analysis, as recognized by several authors [35, 36, 49, 53]. Bonnin et al. [9] show trends in the intensity-duration-frequency (IDF) rainfall curves for the Ohio river basin. A particularly active area of research is the development of nonstationary rainfall IDFs, where theoretical advances in extreme value theory (EVT) turn out to be especially useful (see [16] and references therein). In particular, Cheng et al. [16] describes a new framework for estimating stationary and nonstationary return levels, return periods, and extreme points, which relies on Bayesian inference; the framework is implemented in NEVA software [15]. Ref. [16] offers a case study based on a global temperature dataset, comparing predictions based on stationary and nonstationary extreme value analysis. The study combines local processes (urbanization, local temperature change) and global processes (ENSO cycle, IOD, global temperature change) as time covariates for rainfall IDF, based on Hyderabad data [2]. The comparison shows that the IDF curves derived from the stationary models are underestimating the extreme events of all duration and for all return periods.

Nonisotropy Hydrological processes (soil moisture, streamflow, evapotranspiration) are extremely sensitive to small-scale temporal and spatial rainfall variability. Although ground-based weather radars have been particularly popular for forcing hydrological models that simulate a basin hydrological response, several authors have indicated that the interaction between the variability of precipitation (including spatial and temporal variations) and the resolution of a hydrological model is still poorly understood, especially when radar data are used in an urban environment [12, 22, 54]. If we assume a perfect hydrological model, and we force it with perfect rainfall input, we should expect that the accuracy of a streamflow simulation increases as the resolution of the model and the input increase. However, the finest available radar rainfall temporal resolution does not necessarily provide the best estimation of peak streamflow in a distributed hydrological model. This is the result of uncertainty and errors related to both the precipitation measurement techniques, as discussed in Sect. 3.2, and the model physics [4, 58].

The spatial resolution of precipitation data must be functionally coupled with the temporal resolution to fully reproduce the hydrological response of an urban catchment. For instance, Berne et al. [8] proposed the relation $\Delta s = \frac{3}{2}\sqrt{\Delta t}$ to couple the spatial scale (Δs, in km) with the temporal scale (Δt, in minutes) for rainfall processes in urban catchments.

More recently, Ochoa-Rodriguez et al. [55] fitted the variogram of the spatial structure of rainfall over a peak storm period with an exponential model. They concluded that the minimum required spatial resolution was one-half the characteristic length scale r_c of the storm, which they defined in terms of the variogram range r[L], $r_c = (2\pi/3)^{1/2} r$[L]. A unique relationship linking the temporal and spatial resolutions of precipitation adequate for the reproduction of the hydrological response of a catchment basin is yet to be found.

Nonergodicity Most of the literature simply assumes without evidence that precipitation and hydrological processes in general are ergodic; for example, see [27, 45, 56]. However, a recent study [67] indicates that the assumption may not be fully justified. The author proposed an approach to assess the mean ergodicity of hydrological processes based on the autocorrelation function of a dataset. The approach was tested on monthly rainfall time series at three locations, two in China and one in the State of Michigan (USA). The results showed that, at all three locations, the ergodicity assumption was met only during a few months of the year. Therefore, statistical metrics computed on the basis of data collected during those months do not meet the ergodicity assumption (sample statistics) and cannot be used as proper approximations for the population statistics. Moreover, the ergodicity assumption was met in different months at different locations, so ergodicity cannot be transferred to a different region and/or period. More work is clearly needed to establish the limits of validity of the ergodicity assumption.

Scenarios where the ergodicity assumption is not met have been studied even less frequently than scenarios where the stationarity and isotropy assumptions are not met, partially because of the difficulty of testing it in the absence of large quantities of high-quality data spanning a reasonable period of time.

In statistical mechanics, one often uses nonergodic Monte Carlo simulations to create multiple realizations for estimating statistical information on the dynamic processes over the region in question [47]. In the geospatial sciences, this approach is often infeasible.

An attempt has been made to formulate nonergodic versions of covariograms for the case of preferentially sampled data. However, as argued in [23], these measures do not offer a clear advantage over standard ergodic statistics for studying spatial dependence or making spatial predictions. Developing appropriate data transformations is considered a more promising direction.

In the mathematical literature, much attention is currently being paid to fractional diffusion processes, which typically generate nonergodic behavior. Some recent work aims to develop a metric quantifying nonergodicity [62]. This direction may also be useful for hydrological applications.

3.4 Decorrelation Measures

Correlation functions are standard tools for measuring spatial and temporal dependencies in the rainfall fields [11, 17]. Figure 3.4 shows both the temporal and spatial correlation functions for a precipitation dataset for the State of Oklahoma (USA) during the period March–October, 2011.

In the case of spatial correlations, one computes the correlation of the two time series associated with any two measurement points (for example, two rain gauges or two pixels) as a function of their distance. A common approximation is the exponential model with the so-called *nugget effect* [19, 20],

$$\rho_g(d) = c_0 \exp\left[-(d/d_0)^{s_0}\right]. \tag{3.4.1}$$

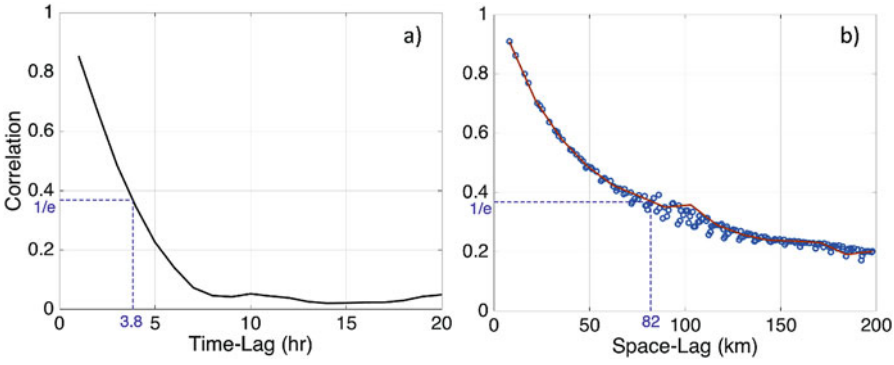


Fig. 3.4 (**a**) Temporal and (**b**) spatial correlations of CMORPH precipitation data for the State of Oklahoma (USA) at 8 km/1 h resolution during the period March–October, 2011

Here, c_0 is the nugget parameter, which corresponds to the correlation value for near-zero distances [21]; d is the separation distance, the distance between the two measurement points; d_0 is the scale parameter, which corresponds to the spatial decorrelation distance; and s_0 is the correlogram shape parameter, which controls the behavior of the model near the origin for small separation distances. The quantity $(1 - c_0)$ is the instant decorrelation due to random errors in the rainfall observations [18]. The separating distance at which the correlation is $1/e$ defines the correlation length for the (assumed) exponential variogram model.

In the case of temporal dependencies, autocorrelations are plotted as a function of the time lag. The lag-1 correlation is commonly adopted as a viable index of rainfall decorrelation in time [38, 60].

The exponential model (3.4.1) with the corresponding "$1/e$ rule" is only one of several models for fitting semivariograms; linear, spherical, and Gaussian models are possible alternatives [5]. The choice of model has to be made based on the analysis of statistical data, and one should not adopt the decorrelation definition provided above as the default option. In fact, one may suspect that for regions with slowly decaying correlations (for example, flat regions with low spatial variability), the "$1/e$ rule" might only work after a sufficient increase in the domain size. In other cases, the data might not support the exponential modeling assumption at all, and corresponding adjustments of the methodology would have to be performed. These modeling subtleties and tacit assumptions are sometimes a source of ambiguity in the literature, which may lead to erroneous conclusions.

3.5 Dimension Reduction Techniques

One of the many challenges of modeling and understanding spatiotemporal precipitation patterns is the large amount of data that needs to be processed. For example, in the relatively small-scale NASA Merra dataset, precipitation is given by monthly averages on a 50×91 grid representing a map of the contiguous USA at 50 km resolution over a period of 35 years, amounting to a total of 1,911,000 entries. However, much more detailed information at higher spatial (on the order of 100 m regionally and 1 km globally) and temporal (hourly) resolutions is required to assess the storage, movement, and quality of water at and near the land surface [68]. Higher-resolution data bring higher data volumes: for the previous example, there would be more than $3 \cdot 10^{12}$ entries for a map of the contiguous US at 1 km resolution and hourly intervals. Some form of data and dimension reduction is called for.

In a general sense, one may attempt to find a decomposition of the data (*signal*) of the form

$$\eta(\mathbf{u}, t) = \sum_{k=1}^{N} \alpha_k(t)\mathbf{p}_k(\mathbf{u}) + \text{noise}, \tag{3.5.1}$$

where the $\mathbf{p}_k$ are *characteristic patterns* used to approximate the data (also called *guess patterns* or *predictors*), and the α_k are the *amplitudes* or *principal components*

of the corresponding patterns. The patterns $\mathbf{p}_k$ are spatial structures that account for temporal variations of the rainfall data $\eta(\mathbf{u}, t)$. When plotted as functions of time, the amplitudes α_k convey information on how the patterns evolve in time.

Mathematically, finding the "best" patterns and principal components for a given dataset is a projection problem, "Find a subspace approximating a given set of data in an optimal (for example, least-squares) sense." To solve this problem, various treatments have been proposed within the geophysical community by different groups and authors [65]. Here, we attempt to place these methods in perspective against methodologies developed independently in the mathematics community. While some techniques exist in both literatures (sometimes under different names), other methods have not yet penetrated the language barrier between the two disciplines.

EOF Method The method of empirical orthogonal functions (EOF) is one of the staple tools in geostatistics, which has received much attention in the hydrological literature. As mentioned in Sect. 3.3.1, EOFs are the eigenvectors of the covariance matrix $Y^T Y$. In the mathematical and statistical literature, the EOF method is referred to as *singular value decomposition* (SVD) or *principal component analysis* (PCA) and belongs to the class of *proper orthogonal decomposition* (POD) methods. In geospatial theory, it goes by the name *Karhunen–Loève analysis*.

Let Y denote the $N \times n$ matrix of detrended observations (also called "snapshot matrix" if $n < N$), whose columns are modified snapshots of rainfall data at a given time. If $C = \frac{1}{n} Y^T Y$ is the normalized correlation matrix, then a POD basis is comprised of the vectors

$$\phi_i = \frac{1}{\sqrt{n\lambda_i}} Y \chi_i, \quad i = 1, \dots, n,$$

where χ_i is the normalized eigenvector ($|\chi_i| = 1$) corresponding to the ith largest eigenvalue λ_i of C. The POD basis vectors are the first n left singular vectors of the snapshot matrix Y obtained by using the SVD decomposition of Y, $Y = U \Sigma V^T$, so $\phi_i = \mathbf{u}_i$ for $i = 1, \dots, n$.

Let $\{\psi_i\}_{i=1}^n$ be an arbitrary orthonormal basis for the span on the modified snapshot set $\{x_j\}_{j=1}^n$. Then the projection onto the d-dimensional subspace spanned by $\{\psi_i\}_{i=1}^n$ is

$$P_{\psi,d} x_j = \sum_{i=1}^{d} (\psi_i, x_j) \phi_i. \tag{3.5.2}$$

The POD basis is optimal in the sense that the approximation error

$$\varepsilon = \sum_{j=1}^{n} |x_j - P_{\psi,d} x_j|^2 \tag{3.5.3}$$

is minimized for $\psi_i = \phi_i, i = 1, \dots, d$.

While EOFs present an attractive tool for studying spatiotemporal variability patterns in precipitation data, care should be taken when interpreting the results of such analysis, as pointed out in [26]. In short, while it is tempting to find physical relevance for each of the EOF "modes," the orthogonality condition built into this methodology often renders such interpretation useless. Rotated EOF technique is often used as a better alternative; however, a deeper analysis is normally needed to decipher the meaning of the EOF-based patterns.

CVT-Based Techniques In the mathematical community, an alternative dimension reduction technique based on *centroidal Voronoi tessellations* (CVTs) has recently gained popularity. While the list of applications is growing quickly, the method remains relatively under-explored in hydrological applications. The presentation below is based on [13].

The idea of the CVT technique is to find a fixed number of representative points ("generators") to decompose the original high-dimensional space into a finite number of subspaces with relatively small loss of accuracy. The main ingredient of this method is the "density function," usually denoted $\rho(\mathbf{x})$, which can be constant or a function of $\mathbf{x}$, depending on the application. For instance, ρ can be used to represent a variety of physical characteristics such as the local characteristic length scale [59], signal intensity [32], or the desired grid resolution [28]. In [25], ρ is used to represent spatial rainfall variability.

More precisely, given a snapshot matrix $X = \{\mathbf{x}_j\}_{j=1}^n \in W \subset \mathbb{R}^N$, the goal is to find a set of points $\{\mathbf{z}_i\}_{i=1}^k \in \mathbb{R}^N$, such that W can be decomposed in Voronoi regions, $W = \cup_{i=1}^k V_i$, with a minimum tessellation error, $\mathscr{E}[\{\mathbf{z}_i, V_i\}_{i=1}^k]$. A Voronoi region V_i is defined as

$$V_i = \{\mathbf{x} \in W : |\mathbf{x} - \mathbf{z}_i| \leq |\mathbf{x} - \mathbf{z}_j|, \ j = 1, \ldots, k, \ j \neq i\}, \tag{3.5.4}$$

and the tessellation error is given by

$$\mathscr{E}[\{\mathbf{z}_i, V_i\}_{i=1}^k] = \sum_{i=1}^{k} \sum_{\mathbf{x} \in V_i} \rho(\mathbf{x}) |\mathbf{x} - \mathbf{z}_i|^2. \tag{3.5.5}$$

It can be shown that the tessellation error is minimal if and only if $\mathbf{z}_i = \mathbf{z}_i^*$ for $i = 1, \ldots, k$, where $\mathbf{z}_i^*$ is the mass centroid of the Voronoi region V_i [30]. At the minimum,

$$\sum_{\mathbf{x} \in V(\mathbf{z}^*)} \rho(\mathbf{x}) |\mathbf{x} - \mathbf{z}^*|^2 = \inf_{\mathbf{z} \in \mathbb{R}^N} \sum_{\mathbf{x} \in V(\mathbf{z})} \rho(\mathbf{x}) |\mathbf{x} - \mathbf{z}|^2. \tag{3.5.6}$$

Figure 3.5 gives two examples of CVTs for different types of densities.

In the discussion of the EOF method, we saw that the optimal basis was comprised of the set of vectors $\{\phi_i\}_{i=1}^d$. In the CVT method, the situation seems similar: the optimal basis is the set of generators $\{\mathbf{z}_i\}_{i=1}^k$. However, there are many

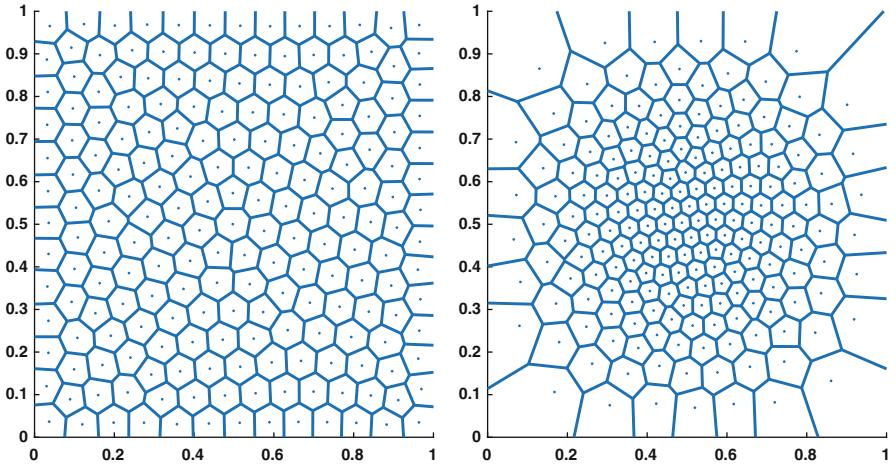


Fig. 3.5 Two CVT tessellations of the unit square, $W = [0, 1]^2$; (left) $\rho(x, y) = 1$, (right) $\rho(x, y) = \exp^{-20(x^2+y^2)}$

differences between the two approaches. POD minimizes the functional ε as in (3.5.3), while CVT minimizes the error $\mathscr{E}$ given by (3.5.5). POD requires one to solve an $n \times n$ eigenvalue problem, where n is the number of snapshots, which is not very amenable to adaptive computations. While the CVT methodology is in general cheaper than POD, there are often numerical caveats associated with CVT computations. For an overview of CVT-related numerical techniques, we refer the reader to [29, 34]. Several case studies based on rainfall data highlighting the features of the CVT and POD approaches are presented in [25].

3.6 Concluding Remarks

In this chapter, we have presented an overview of experimental and computational methodologies and reviewed some of the mathematical challenges associated with the field of precipitation analysis. In particular, we focused our attention on the statistical assumptions underlying some of the commonly used pattern-recognition techniques. Because of the instability of the current climate, the validity of these assumptions should naturally fall under scrutiny. As abundant satellite and *in situ* observation data continue to pour in, one must reconsider the long-standing notions of stationarity, homogeneity, and ergodicity and be prepared to adopt new mathematical methodologies. In this chapter, we reviewed decorrelation measures, nonstationary extensions of intensity-duration-functions, and two types of dimension reduction methodologies with associated challenges. While some of these efforts are well under way, others are still in their infancy, and rigorous mathematical analysis is needed to address these challenges.

Acknowledgements This work was instigated at the Mason Modeling Days workshop held at George Mason University, generously supported by the National Science Foundation grant DMS-1056821. The authors are grateful to Paul Houser for stimulating discussions at the initial stages of this collaboration. ME also wishes to thank Hans Engler and Hans Kaper for their encouragement over the years, and for introducing this research group to the MPE community.

References

1. Ababou, R., Bagtzoglou, A.C., Wood, E.F.: On the condition number of covariance matrices in kriging, estimation, and simulation of random fields. Math. Geol. **26**(1), 99–133 (1994). https://doi.org/10.1007/BF02065878
2. Agilan, V., Umamahesh, N.V.: What are the best covariates for developing non-stationary rainfall intensity-duration-frequency relationship? Adv. Water Resources **101**, 11–22 (2017)
3. Artan, G., Gadain, H., Smith, J.L., et al.: Adequacy of satellite derived rainfall data for streamflow modeling. Nat. Hazards **43**, 167–185 (2007)
4. Atencia, A., Mediero, L., Llasat, M.C., et al.: Effect of radar rainfall time resolution on predictive capability of a distributed hydrological model. Hydrol. Earth Syst. Sci. **15**, 3809–3827 (2011)
5. Bacchi, B., Kottegoda, N.: Identification and calibration of spatial correlation patterns of rainfall. J. Hydrol. **165**, 311–348 (1995)
6. Bauer, P., Lopez, P., Benedetti, A., et al.: Implementation of 1D + 4D-Var assimilation of precipitation-affected microwave radiances at ECMWF. I: 1D-Var. Q. J. Roy. Meteorol. Soc. **132**(620), 2277–2306 (2006)
7. Bell, T.L., Kundu, P.K.: Dependence of satellite sampling error on monthly averaged rain rates: comparison of simple models and recent studies. J. Climate **13**(2), 449–462 (2000)
8. Berne, A., Delrieu, G., Creutin, J.D., et al.: Temporal and spatial resolution of rainfall measurements required for urban hydrology. J. Hydrol. **299**, 166–179 (2004)
9. Bonnin, G.M., Maitaria, K., Yekta, M.: Trends in rainfall exceedances in the observed record in selected areas of the United States 1. J. Am. Water Resour. Assoc. **47**(6), 1173–1182 (2011)
10. Borga, M., Anagnostou, E.N., Frank, E.: On the use of real-time radar rainfall estimates for flood prediction in mountainous basins. J. Geophys. Res. **105**(D2), 2269–2280 (2000)
11. Bras, R.L., Rodriguez-Iturbe, I.: Random Functions and Hydrology. Courier Corporation, Chelmsford (1985)
12. Brown, P.E., Diggle, P.J., Lord, M.E., et al.: Space-time calibration of radar rainfall data. J. Royal Statistical Society: Series C (Applied Statistics) **50**(2), 221–241 (2001)
13. Burkardt, J., Gunzburger, M., Lee, H.C.: Centroidal Voronoi tessellation-based reduced order modeling of complex systems. SIAM J. Sci. Comput. **28**(2), 459–484 (2006)
14. Chang, A.T., Chiu, L.S.: Nonsystematic errors of monthly oceanic rainfall derived from SSM/I. Mon. Weather Rev. **127**(7), 1630–1638 (1999)
15. Cheng, L.: Nonstationary Extreme Value Analysis (NEVA) software package, version 2.0. http://amir.eng.uci.edu/neva.php (2014)
16. Cheng, L., AghaKouchak, A., Gilleland, E., et al.: Non-stationary extreme value analysis in a changing climate. Clim. Chang. **127**(2), 353–369 (2014). https://doi.org/10.1007/s10584-014-1254-5
17. Chumchean, S., Sharma, A., Seed, A.: Radar rainfall error variance and its impact on radar rainfall calibration. Phys. Chem. Earth, Parts A/B/C **28**(1–3), 27–39 (2003)
18. Ciach, G.: Local random errors in tipping-bucket rain gauge measurements. J. Atmos. Ocean. Technol. **20**(5), 752–759 (2003)
19. Ciach, G.J., Krajewski, W.F.: On the estimation of radar rainfall error variance. Adv. Water Resour. **22**(6), 585–595 (1999)

20. Ciach, G.J., Krajewski, W.F.: Analysis and modeling of spatial correlation structure in small-scale rainfall in Central Oklahoma. Adv. Water Resour. **29**(10), 1450–1463 (2006)
21. Cressie, N.A.C.: Statistics for Spatial Data. John Wiley and Sons, Hoboken (1993)
22. Cristiano, E., Ten Veldhuis, M.C., van de Giesen, N.: Spatial and temporal variability of rainfall and their effects on hydrological response in urban areas – a review. Hydrol. Earth Syst. Sci. **21**, 3859–3878 (2017)
23. Curriero, F.C., Hohn, M.E., Liebhold, A.M.: A statistical evaluation of non-ergodic variogram estimators. Environ. Ecol. Stat. **9**, 89–110 (2002)
24. DeGaetano, A.T.: Time-dependent changes in extreme-precipitation return-period amounts in the continental united states. J. Appl. Meteor. Climatol. **48**, 2086–2099 (2009)
25. Di, Z., Maggioni, V., Mei Y., Vazquez M., Houser P., Emelianenko M., 2019, arXiv, arXiv:1908.10403
26. Dommenget, D., Latif, M.: A cautionary note on the interpretation of EOFs. J. Climate **15**, 216–225 (2001)
27. Duan, J., Goldys, B.: Ergodicity of stochastically forced large scale geophysical flows. J. Math. Math. Sci. **28**, 313–320 (2001)
28. Du, Q., Gunzburger, M.: Grid generation and optimization based on centroidal Voronoi tessellations. Appl. Math. Comput. **133**, 591–607 (2002)
29. Du, Q., Faber, V., Gunzburger, M.: Centroidal Voronoi tessellations: applications and algorithms. SIAM Review **41**, 637–676 (1999)
30. Du, Q., Emelianenko, M., Ju, L.: Convergence of the Lloyd algorithm for computing centroidal Voronoi tessellations. SIAM J. Num. Anal. **44**, 102–119 (2006)
31. Ebert, E.E., Janowiak, J.E., Kidd, C.: Comparison of near-real-time precipitation estimates from satellite observations and numerical models. Bull. Amer. Meteor. Soc. **88**, 47–64 (2007)
32. Emelianenko, M.: Fast multilevel CVT-based adaptive data visualization algorithm. Numer. Math. Theor. Meth. Appl. **3**(2), 195–211 (2010)
33. Gottschalck, J., Meng, J., Rodell, M., et al.: Analysis of multiple precipitation products and preliminary assessment of their impact on global land data assimilation system land surface states. J. Hydrometeorl. **6**, 573–598 (2005)
34. Hateley, J.C., Wei, H., Chen, L.: Fast methods for computing centroidal Voronoi tessellations. J. Sci. Comput. **63**(1), 185–212 (2015)
35. Hirsch, R.M.: A perspective on nonstationarity and water management. J. Amer. Water Resources Assoc. (JAWRA) **47**(3), 436–446 (2011)
36. Hodgkins, G.A., Dudley, R.W.: Changes in the timing of winter–spring streamflows in eastern North America. Geophys. Res. Lett. **33**, 1913–2002 (2006)
37. Hossain, F., Anagnostou, E.N.: Assessment of current passive-microwave- and infrared-based satellite rainfall remote sensing for flood prediction. J. Geophys. Res. **109** (2004)
38. Hossain, F., Anagnostou, E.N.: A two-dimensional satellite rainfall error model. IEEE Trans. Geosci. Remote Sens. **44**(6), 1511–1522 (2006)
39. Hsu, K., Gao, X., Sorooshian, S., et al.: Precipitation estimation from remotely sensed information using artificial neural networks. J. Appl. Meteor. **36**, 1176–1190 (1997)
40. Huffman, G.J., Bolvin, D.T., Nelkin, E.J., et al.: The TRMM multisatellite precipitation analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. J. Hydrometeorol. **8**(1), 38–55 (2007)
41. Huffman, G.J., Bolvin, D., Braithwaite, D., et al.: Integrated Multi-satellite Retrievals for GPM (IMERG), version 4.4. NASA's Precipitation Processing Center. Accessed 31 March 2015. ftp://arthurhou.pps.eosdis.nasa.gov/gpmdata/
42. Joyce, R.J., Janowiak, J.E., Arkin, P.A., et al.: Cmorph: a method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution. J. Hydrometeorl. **5**, 487–503 (2004)
43. Kidd, C., Bauer, P., Turk, J., et al.: Intercomparison of high-resolution precipitation products over northwest Europe. J. Hydrometeorl. **13**, 67–83 (2012)
44. Kottegoda, N.T.: Stochastic Water Resources Technology. Palgrave, Macmillan (1980). https://books.google.com/books?id=3SiuCwAAQBAJ

45. Koutsoyiannis, D.: Stochastic simulation of hydrosystems. Water Encyclopedia **3**, 421–430 (2005)
46. Krajewski, W.F., Anderson, M.C., Eichinger, W.E., et al.: A remote sensing observatory for hydrologic sciences: a genesis for scaling to continental hydrology. Water Resour. Res. **42**(7), W07,301 (2006)
47. Krauth, W.: Statistical Mechanics: Algorithms and Computations. Oxford Master Series in Physics. Oxford University Press, UK (2006). https://books.google.com/books?id=B3koVucDyKUC
48. Kummerow, C.: Beamfilling errors in passive microwave rainfall retrievals. J. Appl. Meteorol. **37**(4), 356–370 (1998)
49. Lins, H.F.: A note on stationarity and non-stationarity. 14th Session of the Commission for Hydrology (2012)
50. Lorenc, A.C.: The potential of the ensemble Kalman filter for NWP—a comparison with 4D-Var. Q. J. R. Meteorol. Soc. **129**(595), 3183–3203 (2003)
51. Marzano, F.S., Picciotti, E., Vulpiani, G.: Rain field and reflectivity vertical profile reconstruction from c-band radar volumetric data. IEEE Trans. Geosci. Remote Sens. **42**(4), 1033–1046 (2004)
52. Michaelides, S., Levizzani, V., Anagnostou, E.N., et al.: Precipitation science: measurement, remote sensing, climatology and modeling. Atmos. Res. **94**, 512–533 (2009)
53. Milly, P.C.D., Betancourt, J., Fallkenmark, M., et al.: Stationarity is dead: whither water management? Science **319**, 573–574 (2008)
54. Nikolopoulos, E., Borga, M., Zoccatelli, D., et al.: Catchment scale storm velocity: quantification, scale dependence and effect on flood response. Hydrol. Sci. J. **59**, 1363–1376 (2014)
55. Ochoa-Rodriguez, S., Wang, L., Gires, A., et al.: Impact of spatial and temporal resolution of rainfall inputs on urban hydrodynamic modelling outputs: a multi-catchment investigation. J. Hydrol. **531**, 389–407 (2015)
56. Oliveira, T.F., Cunha, F.R., Bobenrieth, R.F.M.: A stochastic analysis of a nonlinear flow response. Probab. Eng. Mech. **21**, 377–383 (2006)
57. Oliveira, R., Maggioni, V., Vila, D., et al.: Characteristics and diurnal cycle of GPM rainfall estimates over the Central Amazon Region. Remote Sens. **8**(7), 544 (2016)
58. Rafieeinasab, A., Norouzi, A., Kim, S., et al.: Toward high-resolution flash flood prediction in large urban areas: analysis of sensitivity to spatiotemporal resolution of rainfall input and hydrologic modeling. J. Hydrol. **531**, 370–388 (2015)
59. Ringler, T., Ju, L., Gunzburger, M.: A multiresolution method for climate system modeling: application of spherical centroidal Voronoi tessellations. Ocean Dyn. **58**, 475–498 (2008)
60. Rodriguez-Iturbe, I., Isham, V.: Some models for rainfall based on stochastic point processes. Proc. R. Soc. Lond. A **410**(1839), 269–288 (1987)
61. Schneider, U., Fuchs, T., Meyer-Christoffer, A., et al.: Global precipitation analysis products of the GPCC. Global Precipitation Climatology Centre (GPCC), DWD, Internet Publication **112** (2008)
62. Schwarzl, M., Godec, A., Metzler, R.: Quantifying non-ergodicity of anomalous diffusion with higher order moments. Sci. Rep. **7**, 3878 (2017)
63. Scofield, R.A., Kuligowski, R.J.: Status and outlook of operational satellite precipitation algorithms for extreme-precipitation events. Weather Forecast. **18**, 1037–1051 (2003)
64. Serrat-Capdevila, A., Valdes, J.B., Stakhiv, E.: Water management applications for satellite precipitation products: synthesis and recommendations. J. Am. Water Resour. Assoc. **50**, 509–525 (2014)
65. von Storch, H., Navarra, A.: Analysis of Climate Variability Applications of Statistical Techniques. Springer, Berlin (1999)
66. Tian, Y., Peters-Lidard, C.D., Choudhury, B.J., et al.: Multitemporal analysis of TRMM-based satellite precipitation products for land data assimilation applications. J. Hydrometeorol. **8**, 1165–1183 (2007)
67. Wang, H., Wang, C., Zhao, Y., et al.: Toward a practical approach for ergodicity analysis. Nonlin. Processes Geophys. Discuss. **2**, 1425–1446 (2015)

68. Wood, E., Roundy, J.K., Troy, T.J., et al.: Hyper-resolution global land surface modeling: meeting a grand challenge for monitoring Earth's terrestrial water. Water Resour. Res. **47**, W05,301 (2011)
69. Zhang, Q., Sun, P., Singh, V.P., et al.: Spatial-temporal precipitation changes (1956–2000) and their implications for agriculture in China. Global Planet. Change **82**, 86–95 (2012)

Part II
Life Sciences

Chapter 4
Mathematics of Malaria and Climate Change

Steffen E. Eikenberry and Abba B. Gumel

Abstract This chapter is concerned with malaria and the impact of climate change on the spread of malarial diseases on the African continent. The focus is on mathematical models describing the dynamics of malaria under various climate scenarios. The models fit into the Ross–Macdonald framework, with extensions to incorporate a fuller description of the *Anopheles* mosquito life cycle and the basic physics of aquatic anopheline microhabitats. Macdonald's basic reproduction number, $\mathscr{R}_0$, is used as the primary metric for malaria potential. It is shown that the inclusion of air–water temperature differences significantly affects predicted malaria potential. The chapter includes several maps that relate the local ambient temperature to malaria potential across the continent. Under plausible global warming scenarios, western coastal Africa is likely to see a small decrease in malaria potential, while central, and especially eastern highland Africa, may see an increase in malaria potential.

Keywords Anopheles mosquito · Malaria · Ross–Macdonald framework · Basic reproduction number · Malaria potential · Africa

4.1 Introduction

Environmental conditions have always been of profound importance in shaping the epidemiology of infectious diseases. This fact is perhaps best exemplified by the ancient disease of malaria or, more precisely, the collection of closely related malarial diseases.

Caused by *plasmodium* parasites, malaria is spread via the *Anopheles* mosquito. Five species are known to cause the disease in humans, namely *P. falciparum*, *P. vivax*, *P. ovale*, *P. malariae*, and *P. knowlesi* [4]. The first of these, *P. falciparum*,

S. E. Eikenberry · A. B. Gumel (✉)
School of Mathematical and Statistical Sciences, Arizona State University, Tempe, AZ, USA
e-mail: seikenbe@asu.edu; agumel@asu.edu

H. G. Kaper, F. S. Roberts (eds.), *Mathematics of Planet Earth*, Mathematics of Planet Earth 5, https://doi.org/10.1007/978-3-030-22044-0_4

accounts for nearly all global malaria mortality, and most of these deaths occur in children under the age of five in sub-Saharan Africa [105].

The life cycles of both *Plasmodium* and *Anopheles* depend sensitively and nonlinearly on temperature. Thus, anthropogenic global warming may shift and/or expand the geographic range of malarial disease. This phenomenon has been the object of much mathematical modeling and is also the topic of interest in the current chapter. What do mathematical models tell us about the effects of climate change on the dynamics of malaria? Our geographic focus is on Africa, given the burden of disease on this continent.

Outline of the Chapter The chapter is organized as follows. In Sect. 4.2, we provide basic information about malaria: the history of the disease, biology of malaria, its immunology and epidemiology, and the influence of weather and climate on the dynamics of the disease. We then move to mathematics in Sect. 4.3.1. After an overview of the panoply of mathematical models of malaria, we introduce the Ross–Macdonald framework, which forms the basis for most modeling efforts. We discuss thermal-response functions for the various parameters in this framework, and map the resulting malaria potential as a function of temperature across the globe, with and without climate change. In Sect. 4.4, we incorporate elements of the more complex vector life cycle into the Ross–Macdonald framework. We present the hydrodynamics of an immature anopheline habitat and compare predicted anopheline abundance—obtained with the extended model, which includes rainfall, the vector life cycle, and hydrodynamics—with historical data from the WHO's Garki Project in northern Nigeria. We close this section with several maps of malaria potential over the African continent under different modeling options. We summarize our conclusions in the final Sect. 4.5.

4.2 Basic Information About Malaria

It is generally believed that a combination of local and global environmental changes caused the initial spread of *P. falciparum* malaria in proto-agricultural Africa, around 10,000 years ago, when the last ice age ended with a period of global warming, and the onset of the climatically stable Holocene Epoch created conditions favorable to agriculture [22]. Warmer temperatures and increased anopheline habitat created by the clearing of forests for crops, along with concentrated human settlements, created the conditions for both vector and parasite to thrive [83, 103]. By historical times, *P. falciparum* and *P. vivax* had likely spread to much of the inhabited world, even reaching Britain within the last 1000 years [83].

4.2.1 History of Malarial Disease

In antiquity, it was known that in certain seemingly unhealthy areas the population was prone to, among other ailments, periodic fevers (a hallmark of malarial disease),

especially in the summer and autumn. By the late eighteenth century, it was established that, while certain diseases apparently spread directly from person to person (a process termed *contagion*), others were endemic to certain parts of the world and, it seemed, contracted from the environment itself. The putative cause of this latter form of transmission was termed *miasma* (Greek for "pollution"), a kind of poisonous air thought to emanate from soil or rotting matter, and it was believed that the high temperature, humidity, and soils of tropical areas, per se, gave rise to the disease [32].

The central role of low temperatures in limiting the range of malaria was recognized by the mid 1800s by German investigators, who determined that native malaria transmission was limited to areas with average summer temperatures above 15 or 16 °C [65]. In the late nineteenth century (predating modern control programs), malarial disease was likely at its global maximum and clearly concentrated in the warm equatorial band, with its burden falling progressively toward the poles [65].

The close historical concordance between climate and malaria is illustrated visually in Fig. 4.1. The map in the left panel, reconstructed from the famous 1968 publication of Lysenko and Semashko [65], shows the approximate global burden of malarial disease at its global maximum, which happened in most areas in the late nineteenth century. Lysenko and Semashko drew upon a wide variety of sources to construct the first comprehensive map of the global distribution of malaria. Hay and colleagues more recently published a digitized version of this map [55], which has been used in multiple articles, for example, [45].

Why certain tropical areas and marshy regions seemed so prone to malarial disease was finally discerned mechanistically in the late 1800s. Charles Laveran discovered writhing protozoan parasite within the red blood cells of malaria patients in 1880 [29], while Sir Ronald Ross (1857–1932), a British physician, discerned that

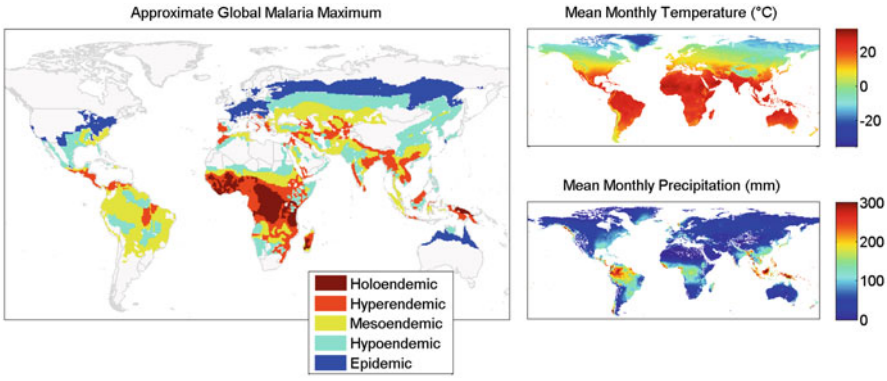


Fig. 4.1 (Left) The approximate distribution of malaria at its global maximum, based on the 1968 publication of Lysenko and Semashko [65]. The map was digitized by color-coding the textures of the original map, and then georeferenced and extracted using QGIS 2.14.3 with GRASS 7.04. (Right) Mean surface temperature and mean precipitation, based on the WorldClim 2.0 database [41]

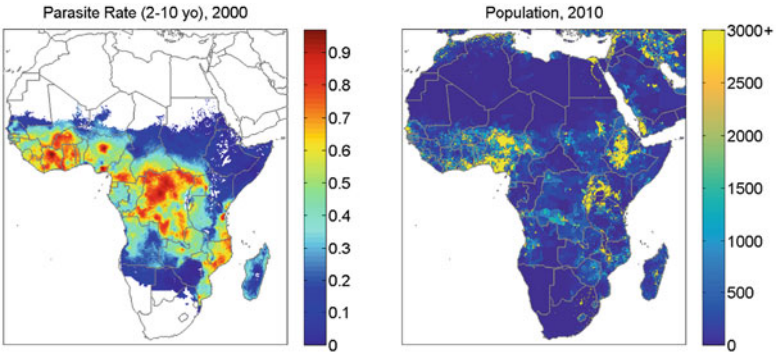


Fig. 4.2 (Left) *P. falciparum* parasite rate (fraction of 2–10 year olds) across continental Africa in 2000 [53], which is fairly similar to the global maximum. (Right) African population distribution in 2010 [25]. In combination, these maps show the populations (as opposed to the areas) most at risk

these protozoans were spread victim to victim by the female *Anopheles* mosquito, which requires standing water and heat to breed [29]. Ross elucidated the malaria transmission cycle first in birds (1897) and later in humans, in Freetown, Sierra Leone (1899).

Economic and agricultural modernization, urbanization, and large-scale malaria control programs in the twentieth century led to a dramatic retreat of malaria across the globe [83]. The exception was Africa, where disease burden stayed near its peak until quite recently [22]. Figure 4.2 shows the *P. falciparum* parasite rate (fraction of 2–10 year olds with detectable blood-stage parasites) across continental Africa in 2000. (Data from the Malaria Atlas Project [53].) Figure 4.2 also shows the modern population across Africa. (Data from the Gridded Population of the World database, version 4 [25].) Population is mainly concentrated in warm coastal West Africa, where malaria is highly endemic (especially in heavily populated Nigeria), and in the cooler eastern highland areas surrounding Lake Victoria and in Ethiopia, where the malaria burden is appreciably lower. Since 2000, parasite rates as well as mortality rates in Africa have fallen dramatically [13].

4.2.2 Basic Biology

Plasmodium parasites, the causative agents of malaria, are eukaryotic protozoans belonging to the large order haemosporidia—a diverse assemblage of parasites that infect and transition between an array of vertebrate hosts and blood-sucking dipteran insect vectors (flies and mosquitoes), and that likely have existed almost as long as the dipterans themselves, at least 150 million years [22]. The consensus is that the haemosporidia first evolved as free-living, sexually reproducing parasites, which colonized the midguts of aquatic insects via a form of extracellular sexual

reproduction known as *sporogony*. Subsequently, they evolved an additional form of intracellular asexual reproduction, known as *schizogony*, which occurs in the vertebrate host and dramatically increases the proliferation potential of the parasite [22, 88].

Plasmodia are known to infect mammals, lizards, birds, and, in the highly unusual case of the *Mesnilium* genus, amphibious fish via an unknown vector [88]. We therefore start the discussion of the malaria life cycle with the injection of motile *sporozoites* from the salivary glands of an infectious mosquito into the skin of its human victim. Within minutes, the sporozoites travel to the liver, where they infect hepatocytes, expand asexually via an initial round of "pre-erythrocyte" schizogony, and ultimately produce 30–40,000 *merozoites* per infected hepatocyte. The hepatocyte then ruptures, spilling the merozoites into the bloodstream, where they initiate the erythrocyte cycle of schizogony, repeatedly infecting and rupturing erythrocytes every 48 or 72 h, and thus yielding the classical tertian (in the cases of *P. vivax* and *P. falciparum*) and quartan (for *P. malariae*) malarial fevers [4].

Merozoites are not able to infect mosquitoes, so in a subset of infected erythrocytes, the invading merozoites ultimately terminally differentiate into male and female *gametocytes*, which are incapable of further schizogony but may initiate sporogony in the mosquito when they are ingested. Upon ingestion, the male and female gametocytes recombine extracellularly in the mosquito midgut, thus initiating the sexual sporogonic cycle, and undergo a series of transformations and invade into the mosquito body cavity, ultimately yielding an *oocyte* which, once mature, ruptures and releases many thousands of sporozoites that make their way to the unfortunate mosquito's salivary glands, for the cycle to continue [4, 31]. This basic process and some of the key differences between the sporogonic and schizogonic cycles are summarized graphically in Fig. 4.3.

The *Anopheles* mosquito also has a relatively complex life cycle, divided broadly into free-flying adult (imago) and aquatic juvenile stages. The adult female mosquito life cycle is dominated by the gonotrophic cycle, which entails the initial taking of a blood meal to fuel egg development, temperature-dependent blood digestion, and egg maturation, and is terminated with oviposition of eggs in an aquatic habitat, only to begin again. In water, the eggs hatch to become actively feeding, motile larvae divided into four instar stages, which eventually become nonfeeding pupae that yield adult mosquitoes [35].

4.2.3 Immunology and Epidemiology

Epidemiologically, malaria transmission was classified by Macdonald [66] in extremes as either *stable* or *unstable*. In the stable case, malaria is *endemic* (Greek for "in the people"). The population is very frequently exposed to infectious bites, inducing a basic level of immunity throughout the population (except in the youngest children), and thus, malaria incidence fluctuates but little, except for normal seasonal changes related to rainfall, temperature, etc. *Epidemics* (a marked

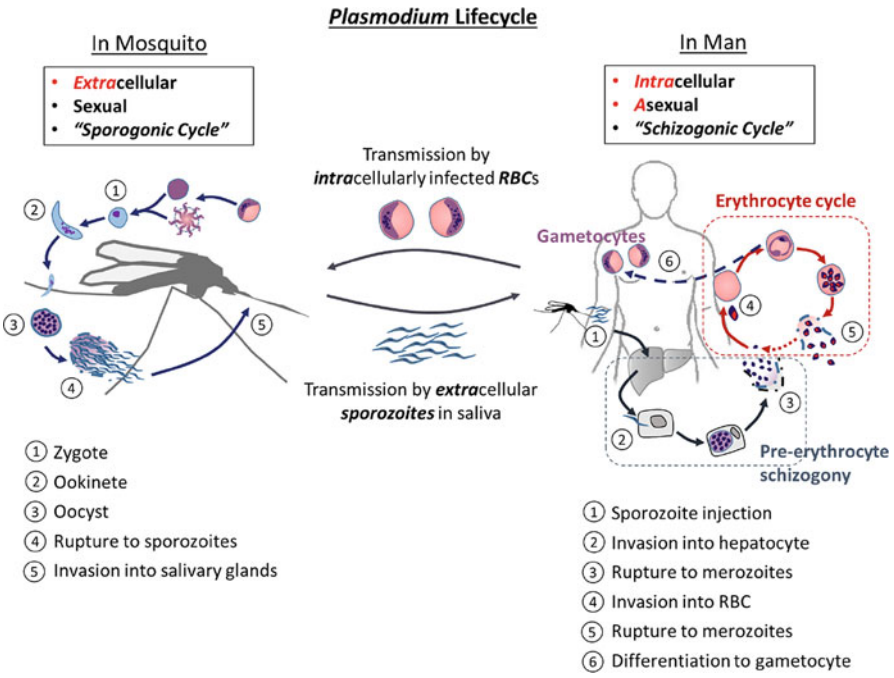


Fig. 4.3 The basic *plasmodium* life cycle in mosquito and man. Sporogony, which entails the sexual reproductive process beginning with ingested gametocytes and ending in salivary sporozoites, is depicted on the left, while asexual schizogony in man is shown on the right

increase in disease from the normal baseline) are unlikely, and malaria is also more difficult to control or, especially, eliminate. On the other hand, in the unstable case, malaria is characterized by low exposure to infectious bites, a varying and low level of population-level immunity, and hence vulnerability to dangerous and sudden epidemics. However, under such conditions, elimination is far more feasible.

Within endemic areas (and typically stable malaria), the unique immunology of malaria is of profound importance in shaping the burden of disease. When transmission is intense, clinical malaria is extremely common, severe, and frequently life-threatening during the first few years of life. It manifests itself as an uncomplicated febrile disease in adolescence and becomes quite rare by adulthood. Indeed, while historically on the order of 50% of all West African children succumbed to malaria before age five [22], Europeans long thought that adult Africans were incapable of acquiring or transmitting the disease [32]. However, adults very frequently have detectable parasites in their blood, yet no clinical symptoms. Thus, there exists a profound disparity between clinical immunity—protection against the clinical manifestations of disease such as fever, malaise, and, in more severe cases, profound anemia, multi-organ dysfunction, or cerebral malaria—and anti-parasite immunity—protection against infection by parasites per se [31]. Clinical immunity is slowly gained over the course of years as a consequence of numerous infectious

bites, while true anti-parasite immunity is rarely, if ever, attained. However, it should be noted that, while blood-borne parasites are detected at nearly the same rates (in a binary sense) in old and young in holoendemic areas, the *density* of parasites in the old is vastly lower [101]. Furthermore, clinical immunity is relatively short-lived and requires frequent re-exposure for maintenance. Adults who move away from endemic areas become vulnerable to severe disease within just a few years [42], although protection against the most severe and life-threatening forms of disease may be longer-lasting [51, 52]. As a consequence, malaria eradication efforts are complicated by the waning of immunity induced by increased control. Thus, initially beneficial control measures have the potential to increase disease later in time [47]. In the worst case, when control measures lapse, transmission can shift to an unstable scenario with the potential for devastating epidemics. The latter scenario is not merely hypothetical and has occurred on multiple occasions from the 1960s onward [28, 103].

4.2.4 Weather and Climate

Vector and parasite life histories depend in highly nonlinear ways on temperature. Adult and immature aquatic mosquito survival is maximized at temperatures in the mid-20s (°C), with survival tailing off rather symmetrically at higher and lower temperatures. The developmental rates of *plasmodium* parasites, immature anophelines, and mosquito eggs, however, all generally increase with temperature up to at least 30 °C. Temperature variability is also likely important in determining survival and development [12, 82], and temperatures may also vary appreciably across the micro-environments to which anophelines are regularly exposed [14, 77, 95].

Like temperature, precipitation and hydrology are fundamental environmental determinants of malarial disease. Malaria transmission often follows a highly seasonal pattern, where the most intense transmission occurs during the rainier season. For example, in the relatively arid African Sahel region, inter-annual variations in disease and *Anopheles* abundance are strongly linked to variations in rainfall [15]; in the Sahel and much of coastal West Africa, highly seasonal rainfall correlates with a highly seasonal pattern of malaria transmission [20]. Moreover, in the Sahel, clinical malaria incidence tends to track rainfall, but with a delay [16]. Monthly rainfall is also positively associated with malaria incidence in both the highlands [26] and coast of Kenya [59]. *An. gambiae*, which tends to breed in small, temporary pools associated with human activity, seems particularly sensitive to short-term rainfall patterns; Koenraadt et al. [60] observed a significant correlation between rainfall lagged by 1 week and adult *An. gambiae* numbers in a Kenyan village.

The relationship between anopheline abundance and rainfall is complex and varies across space and time. For example, the rainiest regions in Sri Lanka actually have the lowest malaria incidence, as strong rainfall results in constantly moving waters that make poor anopheline habitat, while drought may lead to standing waters that breed malaria mosquitoes [18]. However, the malaria incidence *pattern*

still follows that of rainfall, as seasonal malaria cases peak a few months after the seasonal rainfall peak [18]. Thomson and colleagues [98, 99] observed a nonlinear, quadratic relationship between seasonal rainfall and the logarithm of malaria incidence in Botswana: while rain is necessary for habitat, excessive precipitation could wash out anopheline breeding grounds. An experiment by Paaijmans et al. [76] reported a similar phenomenon at the microscale. These authors monitored larvae attrition in artificial habitats over the course of the rainy season in western Kenya and observed that larval death rates were much greater on rainy nights. Bomblies et al. [16] also found the temporal pattern of rainy days during the rainy season itself to be important in explaining inter-annual variations in *Anopheles* abundance.

Malaria, whose distribution depends so profoundly upon temperature and environment, now threatens to change with anthropogenic global warming—warming that is driven principally by the continuously accelerating combustion of fossil fuels, and secondarily by global land-use changes, including deforestation, and large-scale agriculture [97]. Increasing global temperatures are expected to directly affect the capacity of *Anopheles* mosquitoes to transmit malaria.

4.3 Mathematical Modeling

Almost since the first elucidation of its life cycle, malaria has been a subject of mathematical modeling and analysis. In the early 1900s, Ross proposed some simple mechanistic models for malaria transmission [96]. Subsequently, George Macdonald developed a simple model for transmission that included the delay from mosquito infection to infectivity [66]. Macdonald also introduced the *basic reproduction number* $\mathscr{R}_0$—the average number of secondary cases a single initial case generates in a completely susceptible (uninfected and non-immune) population—as an indicator of malaria potential. Macdonald showed that this number is most sensitive to changes in the adult mosquito daily survival probability [66]. His work provided a theoretical justification for the Global Malaria Eradication Programme (GMEP, 1955–1969) of the World Health Organization (WHO). GMEP relied mainly upon indoor residual insecticide spraying for adult vector control, along with mass drug administration [71]. Since the pioneering contributions of Ross and Macdonald, the interest in malaria modeling has expanded considerably, and while many complex models exist, the majority are based upon the Ross–Macdonald framework [89].

Temperature may be incorporated into Ross–Macdonald-style models by making key parameters, such as adult mosquito survival, functions of ambient temperature. Such thermal-response functions are typically determined from experimental data, and allow us to predict malaria potential as a function of current and projected temperature patterns. The inclusion of rainfall in climate-focused mathematical

models has varied. Some models ignore rainfall entirely, while others (generally when the focus is upon mapping malaria) apply some kind of mask or weighting according to either total or seasonal rainfall or an index of wetness such as NVDI [92] or soil moisture [100]. Several dynamical models have used relatively simple relations between rainfall and either oviposition or the larvae carrying capacity of the habitat [57, 104], while yet others—for example, [5, 6, 16]—have employed more physically realistic hydrodynamic modeling to drive the accumulation and loss of water within topographic depressions.

Even if we are interested in the role of temperature (the chief parameter altered by global warming) in determining malaria potential, it is unlikely that we can ignore rainfall and hydrodynamics completely, as the water temperature in aquatic microhabitats is an essential determinant of larval development time and survival. This temperature is, in general, not equal to the ambient temperature, especially in equatorial areas [77, 78]. In fact, the difference can vary with habitat size, time of year, and latitude.

In the last two decades, a large number of mathematical models, both statistical and mechanistic, have been developed to assess the possible impact of climate change upon malaria disease potential. A *partial* list of references includes [1, 3, 11, 12, 15, 16, 27, 30, 33, 39, 40, 57, 62–64, 67, 68, 70, 72, 74, 75, 84, 85, 92, 100, 104, 106], and one may also see Eikenberry and Gumel [37] for a recent review. These models have led to varying conclusions. Some of the earlier models predicted a large *expansion* in the global land area vulnerable to malaria [21, 67, 68, 84], while others predicted only smaller *shifts* in malaria range [45, 54, 90], as some areas where malaria is highly endemic become too hot to support the *Anopheles* vectors, while other cooler areas may become capable of more intense transmission. While the debate on expansion versus shift is still open, several recent process-based modeling efforts support the notion that, in western coastal Africa, the malaria burden may be minimally affected or decrease with global warming [92, 106], while central and eastern highland Africa may see greater disease potential [92]. The impact of global warming on the highlands of western Kenya (in eastern Africa) has been of particular interest in recent years, given the large population increase and concurrent large-scale deforestation [86, 87].

Our objective in this chapter is to introduce the basic mathematical tools necessary to address the question of how climate change may affect malaria potential. We focus mainly on models derived within the Ross–Macdonald framework.

4.3.1 Ross–Macdonald Framework

Sir Ronald Ross (1857–1932), the discoverer of the malaria life cycle, was a polymath who proposed several simple mathematical models for the transmission of malaria among humans and mosquitoes. The 1911 version is a system of two ordinary differential equations [96],

$$\begin{aligned} \frac{dX}{dt} &= abm\, z\,(H - X) - r\,X, \\ \frac{dZ}{dt} &= ac\, x\,(M - Z) - g\,Z, \end{aligned} \tag{4.3.1}$$

where H and X are, respectively, the total and infected human population, and similarly, M and Z the total and infected mosquito population. The parameters are a, the mosquito biting rate (bites/mosquito/day); b, the probability of human infection after an infectious bite ($b = 1$ in Ross's original formulation); c, the probability of a human infecting a mosquito upon biting; $m = M/H$, the number of mosquitoes per human; $z = Z/M$, the fraction of infectious mosquitoes; $x = X/H$, the proportion of infected humans or *parasite rate*; r, the human recovery rate (day^{-1}); and g, the mosquito death rate (day^{-1}). The last parameter is related to the daily survival probability p,

$$g = -\ln p. \tag{4.3.2}$$

The Ross Institute and Hospital for Tropical Diseases was established shortly before Ross's death in 1932. The British malariologist George Macdonald (1903–1967), who became its director in 1947, went on to develop a greatly influential model based upon the ideas of his predecessor but with two major modifications [66]. First, and most significantly, the delay from initial infection to infectivity in mosquito was included, with n denoting the time of the sporogonic cycle, also known as the *extrinsic incubation period* (EIP). Second, since humans may be infected by multiple *Plasmodium* strains, which are all cleared independently, Ross's recovery parameter r was replaced by $\rho(r, h)$, the rate at which new human infections occur. Here, r is the *strain-specific* rate of recovery and h the *inoculation rate* (to be defined shortly).

To obtain Macdonald's model, we recast the system (4.3.1) as a set of *delay differential equations*,

$$\begin{aligned} \frac{dx}{dt} &= abm\, z(t)(1 - x(t)) - \rho(h, r)x(t), \\ \frac{dz}{dt} &= ac\, x(t - n)(1 - z(t - n))e^{-gn} - g\, z(t), \end{aligned} \tag{4.3.3}$$

where $x(t)$ is the parasite rate and $z(t)$ now represents the fraction of mosquitoes that have completed the EIP to become infectious to humans. Thus, $z(t)$ is equivalent to the "sporozoite rate," the fraction of mosquitoes with sporozoites in saliva, denoted s by Macdonald. Note that the growth term in the z equation is now a function of x and z at time $t - n$, and also includes the factor e^{-gn} to account for those mosquito deaths that occur before sporogony is complete. For example, if $n = 10$ days, and the daily survival probability is $p = 0.9$ (hence, $g = 0.1054$), then 65% of initially infected mosquitoes survive to become infectious; if p falls to 0.7, then fewer than 3% of infected mosquitoes survive to infectiousness.

Macdonald formulated his model solely in terms of $x(t)$ by introducing the *inoculation rate*, $h \equiv h(t) = abm\, z(t)$,

$$\frac{dx}{dt} = h(1 - x(t)) - \rho(r, h)x(t), \tag{4.3.4}$$

While Macdonald originally gave an erroneous form for the overall recovery rate $\rho(r, h)$, the correct form is [36, 96]

$$\rho(r, h) = \frac{h}{e^{h/r} - 1}. \tag{4.3.5}$$

At equilibrium, the sporozoite rate is constant, $z(t) = s$, and

$$h = abms = \frac{a^2 bcmp^n x}{ax - \ln p} = \frac{a^2 bcmx}{ax + g} e^{-gn}. \tag{4.3.6}$$

This model yields the following expression for Macdonald's basic reproduction number:

$$\mathscr{R}_0 = \frac{a^2 bcmp^n}{-r \ln p} = \frac{a^2 bcme^{-gn}}{rg}. \tag{4.3.7}$$

The key conclusion is that $\mathscr{R}_0$ is most sensitive to p, the adult daily survival probability. Thus, the model provides a powerful theoretical justification for malaria eradication efforts based upon insecticidal measures, which underpinned the WHO's Global Malaria Eradication Programme (GMEP). Somewhat later, the closely related *vectorial capacity* (VC) metric was defined [43] as the number of new malarial cases (or infectious bites) that could result from a single case in a single day. In terms of the Ross–Macdonald parameters [44], this metric is given by

$$\text{VC} = \frac{a^2 cmp^n}{-\ln p} = \frac{b}{r}\mathscr{R}_0. \tag{4.3.8}$$

Thus, VC is the component of $\mathscr{R}_0$ that depends only on vectorial parameters and is independent of human parameters.

4.3.2 Thermal-Response Functions for Ross–Macdonald Parameters

Nearly all the Ross–Macdonald parameters depend in some way upon climate and weather. The most studied and basic relationships are the temperature dependence of survival (p or g), sporogonic duration and EIP (n), and mosquito biting rate

(a), often taken as the inverse of the duration of the gonotrophic cycle. Less appreciated is the fact that both b and c depend on temperature as well, although this dependence is not typically included in models. This leaves only m lacking explicit temperature dependence. While taken as an imposed parameter in the Ross–Macdonald framework, m is clearly a function of the temperature-dependent parameters a and p, as well as the weather-dependent immature *Anopheles* life cycle, which is completely neglected by Ross–Macdonald models.

Adult Mosquito Survival (p or g) Generally speaking, *Anopheles survival* peaks at temperatures in the mid- to low-20s celsius, and falls fairly symmetrically about this peak. However, rates of the *developmental* processes of gonotrophy (mosquito egg development) and sporogony (parasite development) hasten with increasing temperatures up to at least the low-30s, and may be modeled either by a hyperbolic, monotonically increasing function, or via an asymmetric, unimodal function that drops off rapidly at high temperatures. We briefly review some of the data and functional forms used to model these processes.

While several earlier models, most notably the epidemic potential models of Martens and colleagues, fitted a quadratic curve to three data points from 1949, more recent work has relied on a series of experiments by Bayoh [8], where adult *An. gambiae* was exposed to constant temperatures between 5 and 45 °C at several different relative humidities (RHs). The data are equally well described by a quadratic curve; other authors (e.g., [11]) have used Gaussian distributions. The results are summarized in Fig. 4.4.

If, as is typically (at least implicitly) assumed in most models, the probability of death does not vary with time—that is, p is independent of age—then the death rate g is the inverse of survival time, $p = e^{-g} = e^{-1/S}$, and survival time is *exponentially* distributed. However, this assumption is actually false, as the probability of death increases with age in both laboratory and wild mosquito populations [91]. Such a phenomenon may be described using a variety of probability distributions; the most important ones are the Gompertz distribution, which is widely used and the best fit for many datasets, and the Gamma distribution, which may be straightforwardly incorporated into ODE models, as done, e.g., by Christiansen-Jucht et al. [27]. A discussion of this complication falls outside the scope of the chapter.

Gonotrophic Cycle and Mosquito Biting Rate (a) The gonotrophic cycle is divided into three stages: (I) the search for a host and attack, (II) temperature-dependent blood meal digestion and egg maturation, and (III) oviposition in a suitable body of water [35]. Since only a single blood meal is usually necessary to nourish their eggs, and host attack is a risky energy-intensive process, female anophelines generally take only a single blood meal per cycle (stage I). The overall cycle length, which we denote G_C, can therefore be taken as the inverse of the biting rate, a (supposing, as with the adult survival time and death rate, that G_C is exponentially distributed and age-independent).

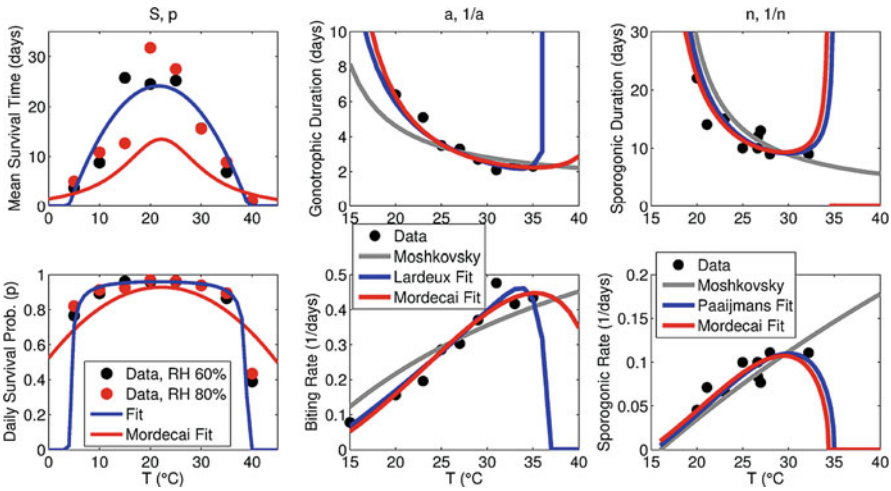


Fig. 4.4 Temperature-dependent Ross–Macdonald parameters. (Left) Mean daily survival S (top) and daily survival probability p (bottom), derived under the (false but expedient) assumption that survival is exponentially distributed. Data points show laboratory survival for *An. gambiae* at either 60 or 80% relative humidity (RH), obtained from [9], together with a quadratic fit to S (blue curve) and Mordecai et al.'s fit to p (red curve). (Middle) Biting rate a and its inverse, gonotrophic duration $(1/a)$. Data points from Lardeux et al. [61], and fits due to Moshkovsky's formula (plus 24 h for stages I and III) [35], Lardeux et al. [61], and Mordecai et al. [70]. (Right) Sporogonic duration n and rate $1/n$. Data points compiled from a variety of sources as reported in [79], and fits using either Moshkovsky's formula [35], or Briere functions due to either Paaijmans et al. [79] or Mordecai et al. [70]

Stage II of the gonotrophic cycle is dominant in terms of time and relates to ambient temperature hyperbolically (at least up to fairly high temperatures). The classical formula of Moshkovsky is based upon the "sum of heat" hypothesis—that is, a certain amount of heat, integrated over time, is necessary to complete development—has been widely used to model this phenomenon. Moshkovsky's expression for the duration of stage II, G_{II}, is

$$G_{II} = \frac{D}{T - T_{min}}, \tag{4.3.9}$$

where $T_{\min}$ is the minimum temperature for development, $T > T_{\min}$ is the mean ambient temperature (°C), and D is an empirical constant measured in degree-days. Detinova [35] gave $D = 37.1$ and $T_{\min} = 9.9$ for the European vector *An. maculipennis* at relative humidity RH = 70–80%, based on 1938 experiments by Shlenova [94].

The expression (4.3.9) is purely *monotonic*, while basic physiology suggests that very high temperatures should, at some point, impede egg development [70].

Thus, we may also use some asymmetric unimodal function, such as the Briere function [17], which relates the rate r of gonotrophy to temperature T as $r(T) = cT(T - T_0)(T_m - T)^{1/2}$. This relation was used by Mordecai et al. [70], with parameters based on much more recent experimental work by Lardeux et al. [61], who examined oviposition in *An. pseudopunctipennis* under different temperatures. Lardeux et al. [61] themselves employed a qualitatively similar unimodal function given graphically in Fig. 4.4.

Stage II generally dominates the gonotrophic cycle. To obtain the complete cycle length G_C, we may add about 24 h to G_{II} for stages I and III [35], although limited habitat availability may prolong stage III and additional time may be needed for the search for suitable waters [50].

Several aspects of mosquito biology can complicate the modeling of the biting rate and temperature. First, multiple blood meals may be taken per gonotrophic cycle, although this is generally only observed among newly emerged anophelines that lack sufficient nutritional reserves to fuel egg development from a single blood meal [93]. The time to first blood meal also takes 1–3 days and is temperature dependent [81]. Finally, and possibly quite significantly [24], malarial infection itself may alter anopheline feeding patterns, and infectious mosquitoes (i.e., those with sporozoites in their salivary glands) have been observed taking more frequent, smaller blood meals, while mosquitoes carrying pre-infectious plasmodium stages (e.g., oocysts) may take fewer blood meals [23].

Lastly, we note that several mathematical models have also assumed a constant hazard of death with each blood meal attempt, such that roughly half of all attempts end in death. This introduces a further "hidden" temperature dependency on adult mosquito survival. But this dependence is often ignored when a and p are considered as independent, imposed parameters.

Sporogonic Cycle Duration (n) Similarly to gonotrophy, the duration of sporogony (or EIP), n, has been described classically using the formula of Moshkovsky (Eq. (4.3.9)), with $D = 111$ degree-days for *P. falciparum* and $D = 105$ degree-days for *P. vivax*, with $T_{\min} = 14.5\,^{\circ}\text{C}$ for the relatively cold-tolerant *P. vivax* and $T_{\min} = 16\,^{\circ}\text{C}$ for all other *plasmodia*.

Note that it has long been recognized that temperatures above 30–32 °C can impede the sporogonic cycle [35, 66], although the major effect may be that higher temperatures impede the early stages of sporogony that immediately follow ingestion of a blood meal, but may not block development beyond the oocyst stage [38]. Okech et al. [73] also found that wild strains of West African *An. gambiae* developed under fluctuating field temperatures up to 33 °C without apparent difficulty. In any case, such observations motivate, as for gonotrophy, the adoption of a unimodal Briere function as done by Mordecai et al. [70] and Paaijmans et al. [79], or a modification of Eq. (4.3.9) to block sporogony above, say, 32–34 °C.

4.3.3 Temperature and Malaria Potential

Using the above relations, we compute the temperature-dependence of $\mathscr{R}_0$. The result, obtained with a quadratic fit to Bayoh's adult mosquito survival data [8] and unimodal Briere functions for gonotrophy and sporogony, is shown in Fig. 4.5.

Given the above (or similar) relations relating key Ross–Macdonald parameters to temperature, one can construct a map representing malaria potential as a function of temperature, as multiple authors have done, often employing some index of transmission potential derived either from Macdonald's $\mathscr{R}_0$ or from a newer model [30, 46, 67, 92], While such a map is sometimes framed as displaying (relative) $\mathscr{R}_0$ across space, as only vector- and parasite-specific parameters vary with temperature, it is perhaps more appropriate to cast it in terms of vectorial capacity (VC) (equivalent to $\mathscr{R}_0$ sans the human-specific components).

Probably the simplest possible approach to malaria mapping would be to calculate VC or $\mathscr{R}_0$ as a function of mean annual temperature. But this may be misleading, as temperatures are often only seasonally suitable for malaria and the relationship between temperature and $\mathscr{R}_0$ is nonlinear. As daily average temperatures are typically only available for global climate datasets at a monthly scale, this is the temporal resolution typically employed [30, 92], although an innovative work by Gething et al. [46] reconstructs approximate daily temperature variations, and daily temperature variation may be crucial to malaria potential [11, 82].

Using the WorldClim 2.0 database [41], we computed monthly values of Macdonald's $\mathscr{R}_0$. Even at this time-scale and disregarding precipitation, malaria potential can vary quite appreciably over the year. The mean monthly $\mathscr{R}_0$, shown in Fig. 4.6, yields a crude measure of annual malaria potential as a function solely of temperature at the global scale. Note that we could simply use mean *yearly* temperatures instead, to yield a single $\mathscr{R}_0$ value at each location. However, this single yearly value varies somewhat from the mean of the $\mathscr{R}_0$ values calculated

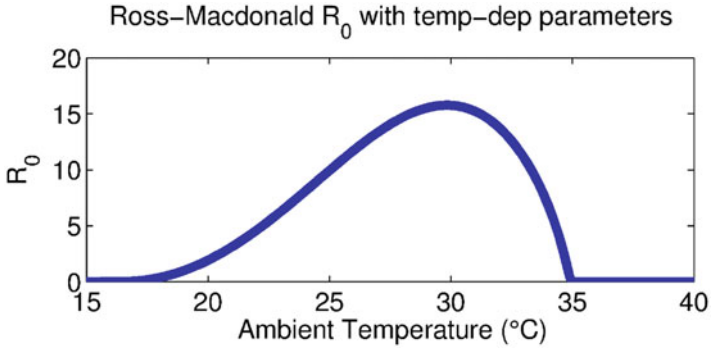


Fig. 4.5 Ross–Macdonald $\mathscr{R}_0$ as a function of temperature, based on the thermal-response functions detailed in the text for a, n, and p, while other parameters are representative of a malaria-endemic region

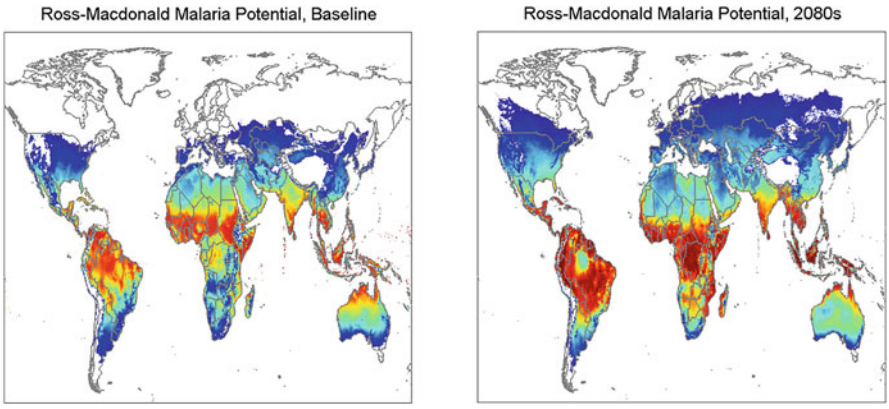


Fig. 4.6 (Left) Global malaria potential based on the Ross–Macdonald model, as a function of temperature alone, for 1970–2000 baseline conditions (per the WorldClim 2.0 database [41]). (Right) The same calculations using projected climate change under the HADCM3 model using the IPCC SRES A1B emissions scenario [102]

across each month individually, suggesting that weather variability is important to account for. In Fig. 4.6, we see surprisingly good visual concordance between the model and Lysenko's historical malaria map. The main points of disagreement are in the extremely arid Saharan region and in the far Eurasian north, where the relatively cold-tolerant *P. vivax* historically caused short-lived epidemic malaria [65]. Malaria potential is also overestimated in the Kalahari Desert of southern Africa.

A variety of other (generally more sophisticated) metrics and methods may be followed in relating temperature to malaria potential. The above efforts do not incorporate precipitation or land cover at all, and precipitation may be included as an independent index of malaria potential (as done in [30]), or we may simply apply a precipitation mask to our map, such that transmission is prohibited anywhere rainfall is sufficiently low. We could also scale $\mathscr{R}_0$ with precipitation in some way, presuming it to be a marker of immature anopheline habitat availability. Ryan et al. [92] applied a mask based on the NDVI, and precipitation and soil moisture may also be more directly incorporated into the underlying model [100]. While we have simply employed VC/$\mathscr{R}_0$, other indices may be used, such as epidemic potential (a metric derived from VC by Martens et al. [67]), or an explicitly time-varying index proportional to VC, as employed by Gething et al. [46].

How might the malaria potential be affected by climate change in the Ross–Macdonald framework? Using the HADCM3 model—a down-sampled GCM climate projection—and the IPCC SRES A1B emissions scenario [102], we have calculated the projected monthly mean $\mathscr{R}_0$ for the 2080s. The results, shown in the right panel of Fig. 4.6, indicate a global expansion of land area that is at some risk and a shift in potential within Africa, where the areas at greatest risk shift roughly south and eastwardly. A similar phenomenon is seen in Figs. 4.13 and 4.14

(Sect. 4.4.2), where we compare predictions under the baseline Ross–Macdonald framework, using an augmented model that considers immature mosquito dynamics.

4.4 Augmented Ross–Macdonald Framework

In our discussions above, we have already alluded to multiple complicating factors that make the Ross–Macdonald framework likely insufficient for capturing the full range of climate effects on malaria epidemiology. Potentially most important, in our view, is the neglect of the immature anopheline life cycle, which is affected by a broad array of environmental factors, including temperature, rainfall, and local hydrodynamics and land use. These factors have been modeled in different ways by multiple authors. Here, we augment the basic Ross–Macdonald framework to explicitly include immature mosquito dynamics as well as adult vectors.

4.4.1 Modeling Immature Anophelines

Immature mosquitoes develop from egg, through four actively feeding larval instar stages, and to a final pupal stage. In a differential equations setting, we may variously lump these developmental stages. Here, we consider a model framework with all four larval instar stages,

$$\begin{aligned}
\frac{dE}{dt} &= \Lambda - \sigma_E E - \mu_E E, \\
\frac{dL_1}{dt} &= \sigma_E E - \sigma_L L_1 - \mu_L L_1 - \Phi, \\
\frac{dL_i}{dt} &= \sigma_L L_{i-1} - \sigma_L L_i - \mu_L L_i - \Phi, \quad i = 2, 3, 4, \\
\frac{dP}{dt} &= \sigma_L L_4 - \sigma_P P - \mu_P P.
\end{aligned} \tag{4.4.1}$$

$E(t)$, $L_1(t), \ldots L_4(t)$, and $P(t)$ are the number of mosquitoes in the egg, first through fourth larval instar, and pupal stage, respectively, at time t; $\sigma_i \equiv \sigma_i(T_W)$ is the temperature-dependent development rate for stage i ($i = E, L, P$); T_W represents water temperature; $\mu_i \equiv \mu_i(T_W, R)$ is the death rate for stage i ($i = E, L, P$), which depends, in general, upon both the water temperature and rainfall history, denoted R. The rate at which eggs are oviposited by adult mosquitoes is given generically as Λ. The function $\Phi \equiv \Phi\left(\sum_i L_i, R\right)$ denotes density-dependent mortality among larvae, which is expected to depend upon rainfall, as this is the source of much anopheline habitat, and upon hydrodynamics more broadly, including topology, vegetation, soil type, etc.

This general framework for immature mosquito dynamics may be coupled to the Ross–Macdonald delay-differential description for adult mosquito dynamics by augmenting the system (4.3.3) by an equation for the total mosquito population, M,

$$\frac{dM}{dt} = \sigma_P P - gM. \tag{4.4.2}$$

We first examine the temperature dependence of development and mortality rates, which is the weather dependency that has been studied most extensively and incorporated into models. Then we examine how this dependence affects an "augmented" version of Macdonald's $\mathscr{R}_0$, which incorporates immature anopheline dynamics. We then turn to a discussion of rainfall and hydrodynamics, and examine the time-varying dynamics of anopheline populations and malarial infections when such behaviors are more fully accounted for. Furthermore, we use microscale hydrodynamic simulations to estimate the relation between air and water temperature as a function of time and latitude to refine the malaria potential map under the augmented Ross–Macdonald model.

Temperature-Dependent Parameters Similar to sporogony and gonotrophy, the rates of immature anopheline development, σ_i, generally increase hyperbolically with temperature, at least up to a point. A purely monotonic relation for larval development time based on work done by Jepson in 1947 [58] has been used in multiple papers, while Bayoh et al. [9] used a unimodal function on the basis of experimental data; the same data were used by Mordecai et al. [70] for a morphologically extremely similar Briere function. The resulting development rates are summarized graphically in Fig. 4.7. Note, however, that while the unimodal functions go to zero because, beyond about 34 °C, larvae fail to develop into adults, it is not clear whether this failure is actually due to increased attrition at high temperatures, rather than arrested development. Laboratory larval survival time as a function of temperature is also given in Fig. 4.7, based on [9, 10], with death rate and survival time described using a fourth-order polynomial fit.

Basic Reproduction Number, $\mathscr{R}_0$ Considering the model framework for immature Anopheles dynamics given above, but excluding any dependence upon rainfall, we extend Macdonald's $\mathscr{R}_0$ as follows. First, we simplify the general model by omitting density-dependent mortality in the larval compartment, and supposing that oviposition, Λ, is limited by a logistic term,

$$\Lambda = a\lambda M\left(1 - \frac{E}{K}\right). \tag{4.4.3}$$

Here, M is the total adult mosquito population, a the biting rate (necessarily equal to the oviposition rate), λ is eggs per oviposition, and K is a carrying capacity,

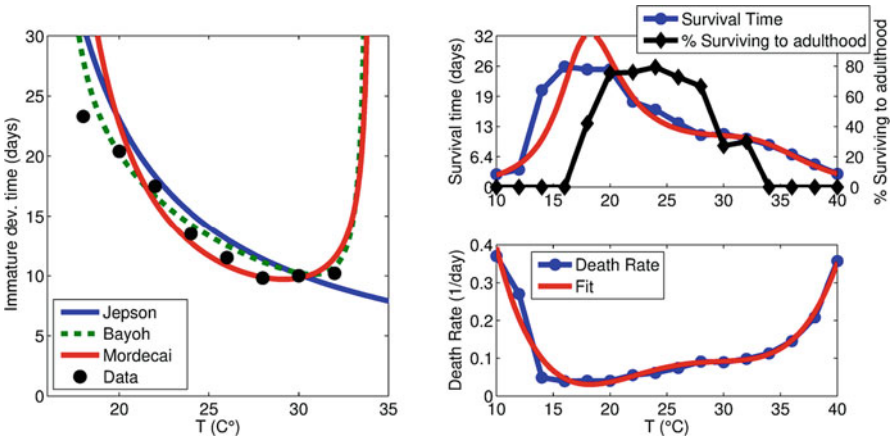


Fig. 4.7 (Left) Immature anopheline development rates, along with unimodal fits due to Bayoh et al. [9], Mordecai et al. [70], or the monotonic relation by Jepson [58]. (Right, top) *An. gambiae* survival times as well as the fraction surviving to adulthood [10]. (This curve is shifted to the right relative to crude survival, as development is faster at higher temperatures.) (Right bottom) Survival time transformed to death rate, assuming exponentially distributed survival, and a fourth-order polynomial fit to the data

which will generally be proportional to habitat availability as determined by land cover and precipitation patterns. We assume, furthermore, that all development (σ_i) and death (μ_i) parameters are constant, and that we have a single larval compartment. At steady-state, M is given by the expression

$$M = \left(\frac{\lambda a \sigma_E \sigma_L \sigma_P}{g(\mu_P + \sigma_P)(\sigma_L + \mu_L)} - \sigma_E - \mu_E \right) \frac{K}{\lambda a}, \tag{4.4.4}$$

and $m = M/H$. Using the thermal-response functions related above, and assuming for simplicity that the temperature-dependent death rates for all immature anophelines are equal, we get the curves for the normalized and absolute $\mathscr{R}_0$ as a function of the ambient temperature T_A given in Fig. 4.8.

When water and air temperatures are equal, immature dynamics may have only a small effect upon the optimum temperature for malaria transmission, as expressed by $\mathscr{R}_0$, but quite dramatically shift the $\mathscr{R}_0$ curve to the left when water is even a few degrees warmer than air. Interestingly, there is an asymmetry such that when water is colder than air, the temperature for peak $\mathscr{R}_0$ is affected minimally, but the absolute magnitude of $\mathscr{R}_0$ decreases across the entire temperature range, and $\mathscr{R}_0$ falls to zero at the lower temperature range. This pattern is also observed in Fig. 4.8.

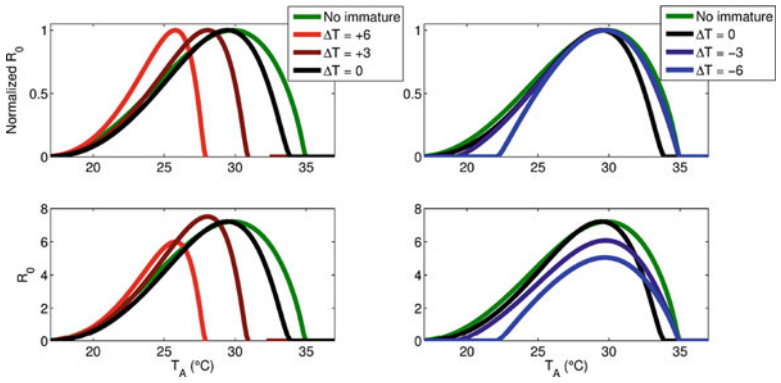


Fig. 4.8 Normalized (top panels) and absolute values (lower panels) of $\mathscr{R}_0$ as a function of ambient temperature, T_A, under the augmented Ross–Macdonald model, including immature anopheline dynamics, for different constant values of $\Delta T = T_W - T_A$. The standard Ross–Macdonald $\mathscr{R}_0$ curve is also given for comparison. As seen, higher water temperatures shift the curve markedly to the left, while colder water temperatures mainly reduce $\mathscr{R}_0$ at small T_A and decrease the magnitude of $\mathscr{R}_0$ across the entire T_A domain

4.4.2 Rainfall and Habitat Dependence

Rainfall Many important anophelines rely upon ephemeral bodies of water often associated with human activity (ruts, hoofprints, etc.), the availability of which is strongly linked to recent rainfall patterns. In our generic model framework, the rainfall time-series, R, must be translated into some measure of habitat availability. Most commonly, this is manifested either at the level of oviposition or via density-dependent larval mortality; see, for example, [27, 57, 85, 104]. Seasonal anopheline abundance and rainfall patterns often track closely, as is strikingly illustrated in an example dataset drawn from the Garki Project [69] in Fig. 4.9. Thus, we may conclude that the Ross–Macdonald parameter m (mosquitoes per human) is fundamentally related to rainfall, as are $\mathscr{R}_0$ and vectorial capacity. Furthermore, the parameter m varies with temperature; this relationship is generally not independent of rainfall, as water temperature is partially determined by rainfall and habitat size, as can be demonstrated both experimentally (see, for example, [77, 78]) and from detailed mathematical modeling.

Several works have employed complex, realistic physical models of water accumulation to form Anopheline habitat, but simpler options exist [27, 57, 104]. White et al. [104] took larval carrying capacity (loosely speaking) to be a convolution of recent rainfall and some weighting function, with the best (of those considered) determined to be an exponential weighting of past rainfall. Earlier work by Hoshen and Morse [57] took the rate of oviposition to be linearly proportional to the sum of rainfall over the last 10 days. A more data-driven approach can also be taken. For example, Lunde et al. [64] calculated immature anopheline carrying

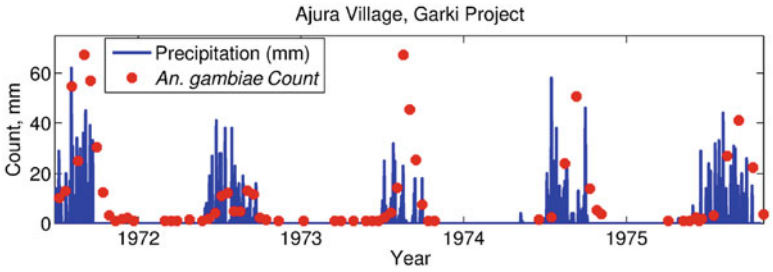


Fig. 4.9 Precipitation and *An. gambiae* counts (from pyrethrum spray collections) in the Ajura Village included in the Garki Project [69]. Seasonal peaks in mosquitoes clearly track the seasonal rainfall pattern, although within each season, total precipitation is a rather poor predictor of total mosquitoes collected

capacity for different spatial grid-cells as a composite function of soil moisture and *potential* river length. The latter quantity was derived from the HydroSHEDS database, which provides data on the potential for water accumulation, given the topology of the Earth's surface. Bomblies et al. [16] developed a more comprehensive model, explicitly including runoff, flow, and water accumulation within depressions at village scale topography, and coupled this to an agent-based model for malaria mosquitoes. This work also formed the basis for several more recent studies [15, 106].

Habitat Regardless of how the accumulation and loss of habitat volume and/or surface area is determined, this metric must be translated into some kind of carrying capacity or density-dependent death term, etc. Several authors have assumed a biomass carrying capacity for anopheline ponds of about 300 mg m^{-2}, with stage-four instars weighing 0.45 mg [16, 33, 100]. Under the augmented Ross–Macdonald model, we may also simply limit oviposition via a logistic term, with the egg carrying capacity proportional to water surface area.

At the microscale, we can apply first principles from physics to describe water in a suitable depression as habitat for immature mosquitoes. Such a model can yield a time-varying immature carrying capacity, help elucidate the relationship between water and air temperature, give insight into the relationship between habitat parameters such as depth or shading and anopheline numbers, as well as justify or motivate simpler phenomenological relationships between anopheline habitat and rainfall.

Our first task is to define some habitat geometry that relates depth (d), surface area (A), and volume (V). Options include, for example, a cylindrical geometry, right-angle cone, or a somewhat more general geometry proposed by Hayashi and colleagues [19, 56], which describes a depression by maximum surface area, maximum depth, and a dimensionless shape parameter, p.

The various mechanisms involved in microscale habitat hydrodynamics are schematically summarized in Fig. 4.10. Water volume is gained at a rate proportional

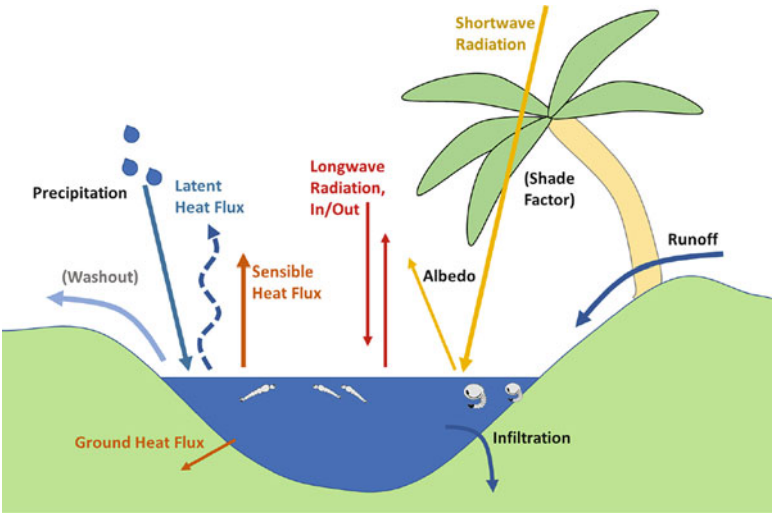


Fig. 4.10 Schematic for heat and volume balance in an anopheline microhabitat. Heat is gained and lost via both short- and long-wave radiation, precipitation, infiltration, runoff, and washout, while the same mechanisms lead directly or indirectly (i.e., via latent heat lost in the form of water vapor) to volume changes

to precipitation (both directly and via runoff over some catchment area), while it is lost through evaporation and infiltration into the soil. Volume balance is described by the ordinary differential equation

$$\frac{dV}{dt} = P(A + R_{\text{frac}}(A_{\text{catch}} - A)) - A(E + I), \tag{4.4.5}$$

where $P \equiv P(t)$ is the precipitation time-series (m); A_{catch} is the catchment area for precipitation runoff, with R_{frac} the fraction running off; E and I are volume losses due to evaporation (latent heat flux) and infiltration, respectively, with infiltration dominant in the Sahel [16, 34]. The infiltration rate varies nonlinearly; a simple expression is given in [5]. Washout, which happens when influx exceeds the maximum volume of the habitat, can serve as a source of larval mortality, as demonstrated experimentally in artificial habitats by Paaijmans et al. [76]. It has been incorporated into models, for example, by having larval mortality increase with precipitation [100] or via a quadratic relationship between egg survival and rainfall [85].

As evaporation represents the loss of both water and heat, the heat balance of a habitat is directly coupled to its volume balance. This suggests a modeling complication of potentially fundamental importance: water and ambient temperatures are not necessarily equal, nor do they necessarily differ by a constant offset. The heat balance in a habitat can be described by an ordinary differential equation for the total heat, Q (joules),

$$\frac{dQ}{dt} = A(R_n - \lambda E - H - G) + P_Q - I_Q. \tag{4.4.6}$$

Here, A is habitat surface area, R_n is net radiation per unit area, λE is latent heat flux (i.e., the heat contained in evaporating water), H is sensible heat flux, and G is heat flux through the surrounding soil [2, 6]. The units are m^2 for A and Wm^{-2} for the quantities inside the parentheses. Heat is gained via precipitation and runoff, P_Q, and lost via infiltration, I_Q.

Further mathematical details and model formulations can be found in [5–7, 16, 77, 78, 85, 100].

Figure 4.11 demonstrates how rainfall in the Ajura village of the Garki Project might translate into habitat volume and adult mosquito populations. The results were obtained with the model represented in Fig. 4.10, assuming a relatively large habitat (maximum surface area 100 m^2 and maximum depth of 1.5 m). The graphs show that the "impulse response function"—that is, the habitat volume time-series response following a single pulse of rainfall—is essentially an exponential, concordant with the conclusions of White et al. [104].

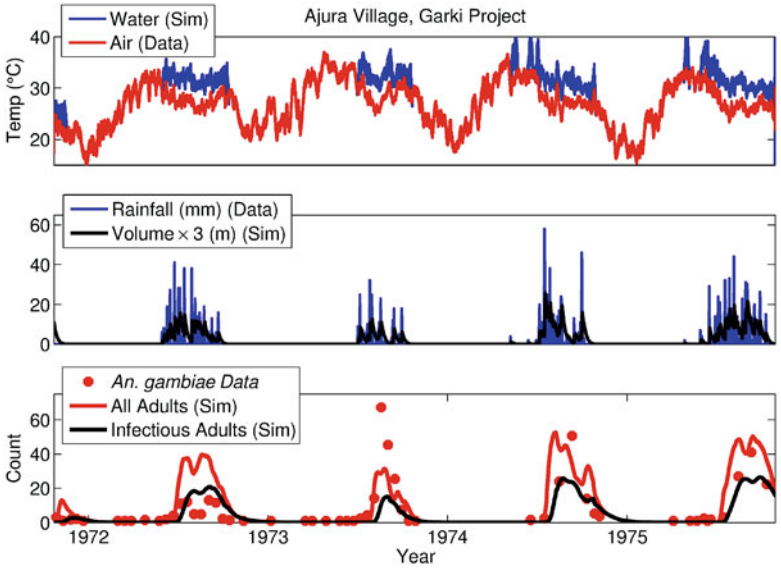


Fig. 4.11 Simulated water temperatures, habitat volume, and adult mosquito populations (both total and infectious), based on ambient air temperature and precipitation data for the Garki region [69]. Maximum and minimum air temperatures were used to develop sinusoidally varying daily temperature profiles, and daily solar insolation was calculated from time of year and latitude (12.4°N). (Top) Smoothed time-series of ambient and (simulated) water temperatures. (Center) Precipitation and simulated habitat volume. (Bottom) Modeled adult *Anopheles* populations and corresponding data-points from the Ajura village

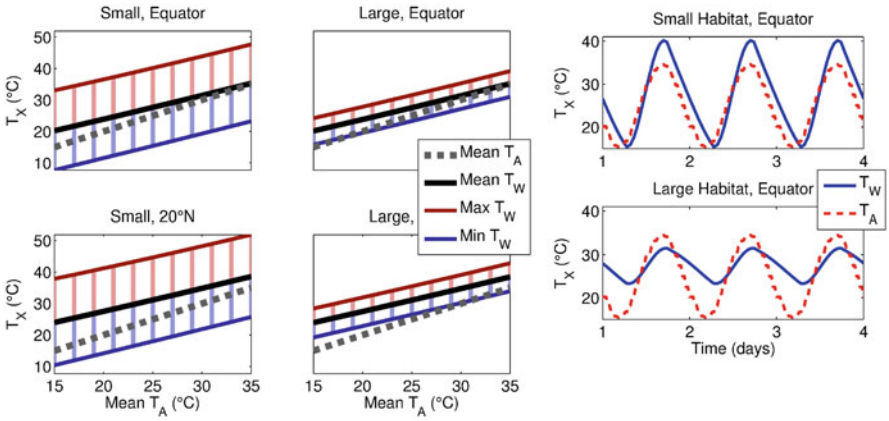


Fig. 4.12 Simulated water temperatures under sunny, low-wind conditions, for small and large habitats, at the equator and at 20°N, at the height of summer. The four panels on the left show how the minimum, mean, and maximum values of the water temperature, T_W, vary with mean air temperature, T_A. The two panels on the right show time-series of diurnally varying air and water temperatures in a small (top) and large (bottom) habitat. The average ΔT is similar across habitat sizes, but actual T_W is much more variable in the smaller habitat

The top panel in Fig. 4.11 indicates that ambient and water temperatures can differ significantly. Water temperatures are typically around 2–6 °C greater than air temperatures, and about 4 °C greater on average. This is consistent with experiments by Paaijmans et al. [77, 78], who recorded diurnally varying ambient and water temperatures in a nearly equatorial area of Kenya in artificial anopheline habitats and found water to be several degrees warmer, especially at the height of the day.

By modeling the heat balance of microhabitats, we can get a better idea of the likely difference $\Delta T = T_W - T_A$ across time and space and use this information to motivate an improved set of temperature-dependent malaria potential maps. Figure 4.12 shows how, under simulated diurnal temperature and solar radiation variation, water and ambient temperature vary over the course of a day. The variability in water temperature is greater for smaller habitats, although ΔT is fairly insensitive to habitat size. We also see from this figure that the average of ΔT is likely not constant but varies with T_A, such that ΔT is greater at lower ambient temperatures.

We emphasize that, while insolation at the equator is almost invariant throughout the year, there is nontrivial seasonal variability even at ±20° latitude. Simulations suggest that, at 20°N under low-wind and sunny conditions, ΔT may range from about −4 to +1 °C in winter, but around +4 to +8 °C during summer. We have generated a set of ΔT data points as a function of month (using the first Julian day of the month), latitude, and average T_A, which yields a multiple linear regression for ΔT at any time and spatial point.

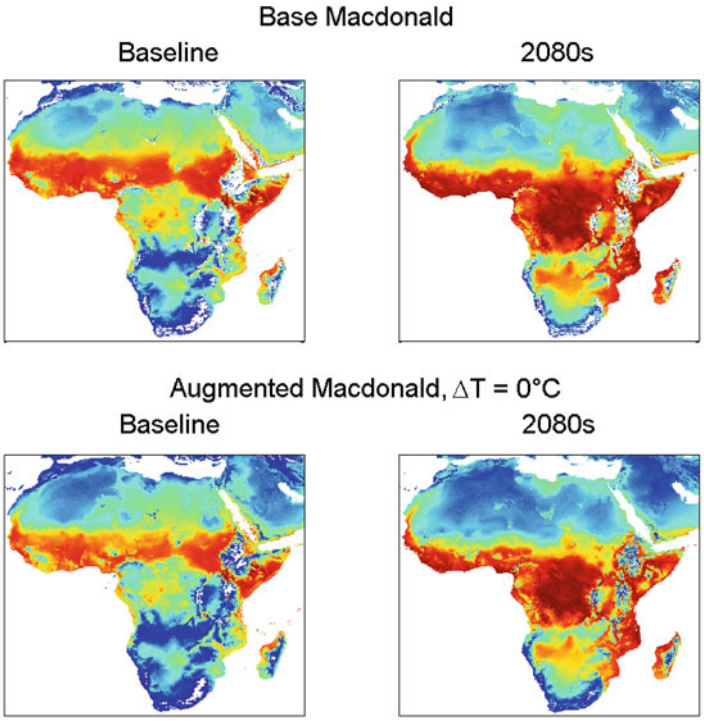


Fig. 4.13 Temperature-dependent malaria potential (as measured by $\mathscr{R}_0$) across continental Africa. (Left) Baseline conditions (based on WorldClim 2.0 [41]). (Right) 2080s global warming conditions (based on the HADCM3 model and SRES A1B scenario [102]). (Top) Basic Ross–Macdonald framework. (Bottom) Augmented Ross–Macdonald model with $\Delta T = 0\,^{\circ}\mathrm{C}$

4.4.3 Malaria Potential Across Africa

Figures 4.13 and 4.14 show how temperature-dependent malaria potential varies across continental Africa under three scenarios, two where ΔT is fixed ($\Delta T = 0\,^{\circ}\mathrm{C}$ and $\Delta T = 3\,^{\circ}\mathrm{C}$) and one where ΔT varies with date and latitude, under baseline conditions and a possible global warming scenario. All models predict a contraction in malaria potential under global warming in west coastal Africa. When ΔT is variable, the models predict appreciably more malaria potential in heavily populated eastern highland Africa compared to a fixed ΔT of $3\,^{\circ}\mathrm{C}$.

4.5 Summary and Conclusions

Malaria epidemiology is fundamentally linked to weather and climate, and it remains to be seen how anthropogenic global warming will ultimately influence this disease. Multiple mathematical models have addressed this question, with somewhat

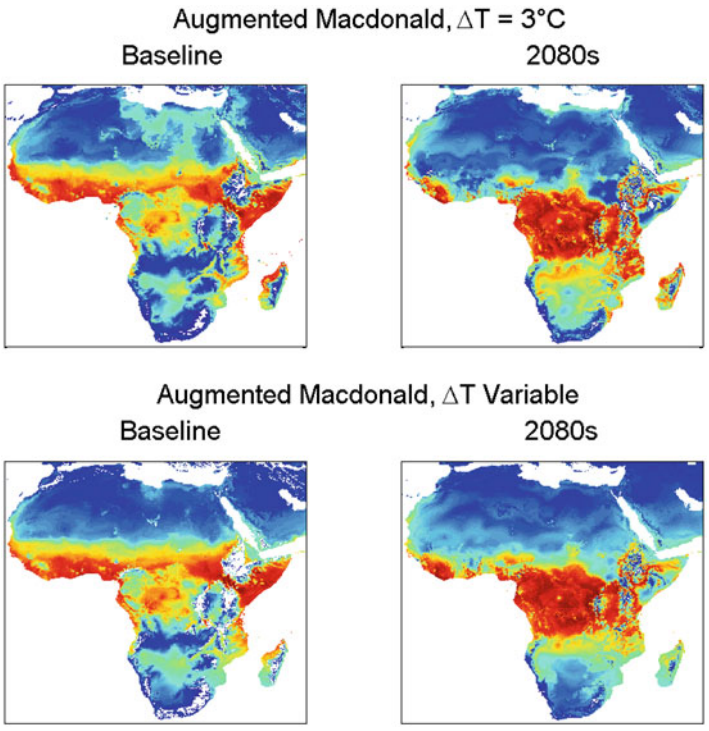


Fig. 4.14 Similar to Fig. 4.13. Temperature-dependent malaria potential under baseline conditions (left) and 2080s global warming conditions (right). (Top) Augmented Ross–Macdonald model with $\Delta T = 3\,^{\circ}\mathrm{C}$. (Bottom) ΔT varying with month and latitude

varying conclusions, although the most likely outcome is a modest expansion of the global geographic areas potentially at risk. Within Africa, where almost the entire malaria burden is currently felt, there may be a shift in the areas and populations most at risk, from western to central and eastern Africa, particularly in some populous highland areas in western Kenya, Uganda, and Ethiopia. Furthermore, there are likely to be seasonal shifts in disease transmission [92], and precipitation, land use, and hydrology are all likely to be important environmental factors at local and regional scales [16, 86]. The goal of this chapter has been to establish the basic biology of malarial disease, and present in some detail how this is translated into the Ross–Macdonald framework (and extensions thereof), which forms the basis for hundreds of mathematical models and yields a widely used expression for the basic reproduction number, $\mathscr{R}_0$.

The Ross–Macdonald number $\mathscr{R}_0$ relates malaria potential to several key quantitative parameters. These parameters all depend on temperature, and at least one, namely the mosquito-to-human ratio (m), depends additionally on rainfall

and land use. Because the Ross–Macdonald number forms the basis of many climate-focused mathematical malaria models, we have examined the thermal-response functions and data sources in detail for some of the key parameters, namely the duration of the sporogonic cycle (n), the mosquito biting rate (a), and the daily survival probability (p). Given these thermal-response functions, we can compute $\mathscr{R}_0$ as a function of temperature, and thus produce global and continental-scale maps of malaria potential, both under current and projected climatic conditions.

We extended the Ross–Macdonald framework to include a basic model for immature *Anopheles* dynamics. Using this *augmented Macdonald framework*, we show that the temperature-dependence of $\mathscr{R}_0$ may vary appreciably, depending upon how air and water temperature relate. This can significantly affect predicted malaria potential.

Precipitation and hydrodynamics are also fundamentally important to vectorial capacity. We examined how they may be reasonably modeled at the small scale to predict *Anopheles* abundance. We incorporated the effect of different air and water temperatures into the augmented Ross–Macdonald framework to generate malaria potential maps under various climate change scenarios. The results show that the populations of western Africa may be less susceptible to malaria under climate change, but those in the east may be more vulnerable. This finding is consistent with several more sophisticated modeling studies [92, 106]. Incorporating precipitation more directly into large-scale malaria potential maps is likely essential to reach a full understanding of the effect of global warming on malaria potential, but we defer that task to the future.

Finally, a variety of other phenomena affect malaria, and it is essential to at least mention some of these factors. From our discussion on modeling anopheline habitat, rainfall, and water heat- and volume-balance, we have already seen that water temperature may differ from ambient, and this difference can affect the optimum temperature for malaria transmission. In our discussion of mapping malaria potential under the Ross–Macdonald framework, we also demonstrated the importance of monthly temperature variations. It comes as no surprise then that daily variations in both ambient and water temperature also appreciably affect malaria potential, as has been demonstrated in several recent experimental and theoretical works [12, 14, 79, 80, 82]. Overall, temperature variability seems likely to asymmetrically affect malaria potential such that transmission is reduced at higher temperatures, while it may have a smaller effect at colder temperatures [12].

As elaborated in Sect. 4.2.3, the unique immunology of malaria is central to its epidemiology. But this fact has generally been ignored in mathematical malaria models that focus on the potential impact of climate change (but see, e.g., [106] for an exception). However, there exists a substantial mathematical literature focused on this aspect of the disease, dating at least to the influential Garki Model developed by Dietz et al. [36]. More recent works are [42, 47–49, 51, 52]. A full accounting for the interaction between climate and immunity is likely to be a fundamental challenge for the future.

References

1. Agusto, F., Gumel, A., Parham, P.: Qualitative assessment of the role of temperature variations on malaria transmission dynamics. J. Biol. Syst. **23**(4), 597–630 (2015)
2. Allen, R.G., Pereira, L.S., Raes, D., et al.: FAO Crop evapotranspiration (guidelines for computing crop water requirements irrigation and drainage paper 56. Technical Report, Food and Agriculture Organization of the United Nations (FAO), Rome, (1998)
3. Alonso, D., Bouma, M.J., Pascual, M.: Epidemic malaria and warmer temperatures in recent decades in an East African highland. Proc. R. Soc. Lond. B Biol. Sci. **278**(1712), 1661–1669 (2010)
4. Antinori, S., Galimberti, L., Milazzo, L., et al.: Biology of human malaria plasmodia including Plasmodium knowlesi. Mediter. J. Hematol. Infect. Dis. **4**(1) (2012)
5. Asare, E.O., Tompkins, A.M., Amekudzi, L.K., et al.: A breeding site model for regional, dynamical malaria simulations evaluated using in situ temporary ponds observations. Geospat. Health **11**(1s), 391 (2016)
6. Asare, E.O., Tompkins, A.M., Amekudzi, L.K., et al.: Mosquito breeding site water temperature observations and simulations towards improved vector-borne disease models for Africa. Geospat. Health **11**(1s), 67–77 (2016)
7. Asare, E.O., Tompkins, A.M., Bomblies, A.: A regional model for malaria vector developmental habitats evaluated using explicit, pond-resolving surface hydrology simulations. PLoS One **11**(3), e0150626 (2016)
8. Bayoh, M.N.: Studies on the Development and Survival of *Anopheles Gambiae* Sensu Stricto at Various Temperatures and Relative Humidities. Ph.D. thesis, Durham University, Durham (2001). http://etheses.dur.ac.uk/4952/
9. Bayoh, M., Lindsay, S.: Effect of temperature on the development of the aquatic stages of *Anopheles gambiae* sensu stricto (diptera: Culicidae). Bull. Entomol. Res. **93**(5), 375–381 (2003)
10. Bayoh, M.N., Lindsay, S.W.: Temperature-related duration of aquatic stages of the Afrotropical malaria vector mosquito *Anopheles gambiae* in the laboratory. Med. Vet. Entomol. **18**(2), 174–179 (2004)
11. Beck-Johnson, L.M., Nelson, W.A., Paaijmans, K.P., et al.: The effect of temperature on anopheles mosquito population dynamics and the potential for malaria transmission. PLoS One **8**(11), e79276 (2013)
12. Beck-Johnson, L.M., Nelson, W.A., Paaijmans, K.P., et al.: The importance of temperature fluctuations in understanding mosquito population dynamics and malaria risk. R. Soc. Open Sci. **4**(3), 160969 (2017)
13. Bhatt, S., Weiss, D., Cameron, E., et al.: The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015. Nature **526**(7572), 207–211 (2015)
14. Blanford, J.I., Blanford, S., Crane, R.G., et al.: Implications of temperature variation for malaria parasite development across Africa. Sci. Rep. **3**, 1300 (2013)
15. Bomblies, A.: Modeling the role of rainfall patterns in seasonal malaria transmission. Clim. Chang. **112**(3–4), 673–685 (2012)
16. Bomblies, A., Duchemin, J.B., Eltahir, E.A.: Hydrology of malaria: model development and application to a Sahelian village. Water Resour. Res. **44**(12) (2008)
17. Briere, J.F., Pracros, P., Le Roux, A.Y., et al.: A novel rate model of temperature-dependent development for arthropods. Environ. Entomol. **28**(1), 22–29 (1999)
18. Briët, O.J., Vounatsou, P., Gunawardena, D.M., et al.: Temporal correlation between malaria and rainfall in Sri Lanka. Malar. J. **7**(1), 77 (2008)
19. Brooks, R.T., Hayashi, M.: Depth-area-volume and hydroperiod relationships of ephemeral (vernal) forest pools in southern New England. Wetlands **22**(2), 247–255 (2002)
20. Cairns, M., Roca-Feltrer, A., Garske, T., et al.: Estimating the potential public health impact of seasonal malaria chemoprevention in African children. Nat. Commun. **3**, 881 (2012)

21. Caminade, C., Kovats, S., Rocklov, J., et al.: Impact of climate change on global malaria distribution. Proc. Natl. Acad. Sci. **111**(9), 3286–3291 (2014)
22. Carter, R., Mendis, K.N.: Evolutionary and historical aspects of the burden of malaria. Clin. Microbiol. Rev. **15**(4), 564–594 (2002)
23. Cator, L.J., Lynch, P.A., Read, A.F., et al.: Do malaria parasites manipulate mosquitoes? Trends Parasitol. **28**(11), 466–470 (2012)
24. Cator, L.J., Lynch, P.A., Thomas, M.B., et al.: Alterations in mosquito behaviour by malaria parasites: potential impact on force of infection. Malar. J. **13**(1), 164 (2014)
25. Center for International Earth Science Information Network (CIESIN)–Columbia University: Gridded Population of the World, version 4 (GPWv4): Population Count, Revision 10. Technical report, NASA Socioeconomic Data and Applications Center (SEDAC), Palisades (2017). https://doi.org/10.7927/H4PG1PPM|. Accessed 1 February 2018
26. Chaves, L.F., Hashizume, M., Satake, A., et al.: Regime shifts and heterogeneous trends in malaria time series from Western Kenya Highlands. Parasitology **139**(1), 14–25 (2012)
27. Christiansen-Jucht, C., Erguler, K., Shek, C.Y., et al.: Modelling *Anopheles gambiae ss* population dynamics with temperature-and age-dependent survival. Int. J. Environ. Res. Public Health **12**(6), 5975–6005 (2015)
28. Cohen, J.M., Smith, D.L., Cotter, C., et al.: Malaria resurgence: a systematic review and assessment of its causes. Malar. J. **11**(1), 122 (2012)
29. Cox, F.E.: History of the discovery of the malaria parasites and their vectors. Parasit. Vectors **3**(1), 5 (2010)
30. Craig, M.H., Snow, R., le Sueur, D.: A climate-based distribution model of malaria transmission in sub-Saharan Africa. Parasitol. Today **15**(3), 105–111 (1999)
31. Crompton, P.D., Moebius, J., Portugal, S., et al.: Malaria immunity in man and mosquito: insights into unsolved mysteries of a deadly infectious disease. Annu. Rev. Immunol. **32**, 157–187 (2014)
32. Curtin, P.D.: Medical knowledge and urban planning in tropical Africa. Am. Hist. Rev. **90**(3), 594–613 (1985)
33. Depinay, J.M.O., Mbogo, C.M., Killeen, G., et al.: A simulation model of African Anopheles ecology and population dynamics for the analysis of malaria transmission. Malar. J. **3**(1), 29 (2004)
34. Desconnets, J.C., Taupin, J.D., Lebel, T., et al.: Hydrology of the HAPEX-Sahel central supersite: surface water drainage and aquifer recharge through the pool systems. J. Hydrol. **188**, 155–178 (1997)
35. Detinova, T.S., Bertram, D., et al.: Age-Grouping Methods in Diptera of Medical Importance: With Special Reference to Some Vectors of Malaria. World Health Organization, Geneva (1962)
36. Dietz, K., Molineaux, L., Thomas, A.: A malaria model tested in the African savannah. Bull. World Health Organ. **50**(3–4), 347 (1974)
37. Eikenberry, S.E., Gumel, A.B.: Mathematical modeling of climate change and malaria transmission dynamics: a historical review. J. Math. Biol. **77**, 857–933 (2018)
38. Eling, W., Hooghof, J., van de Vegte-Bolmer, M., et al.: Tropical temperatures can inhibit development of the human malaria parasite Plasmodium falciparum in the mosquito. In: Proceedings of the Section Experimental and Applied Entomology–Netherlands Entomological Society, vol. 12, pp. 151–156 (2001)
39. Ermert, V., Fink, A.H., Jones, A.E., et al.: Development of a new version of the Liverpool Malaria Model. I. Refining the parameter settings and mathematical formulation of basic processes based on a literature review. Malar. J. **10**(1), 35 (2011)
40. Ermert, V., Fink, A.H., Jones, A.E., et al.: Development of a new version of the Liverpool Malaria Model. II. Calibration and validation for West Africa. Malar. J. **10**(1), 62 (2011)
41. Fick, S.E., Hijmans, R.J.: WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int. J. Climatol. **37**(12), 4302–4315 (2017)

42. Filipe, J.A., Riley, E.M., Drakeley, C.J., et al.: Determination of the processes driving the acquisition of immunity to malaria using a mathematical transmission model. PLoS Comput. Biol. **3**(12), e255 (2007)
43. Garrett-Jones, C.: Prognosis for interruption of malaria transmission through assessment of the mosquito's vectorial capacity. Nature **204**(4964), 1173–1175 (1964)
44. Garrett-Jones, C., Shidrawi, G.: Malaria vectorial capacity of a population of anopheles gambiae: an exercise in epidemiological entomology. Bull. World Health Organ. **40**(4), 531–545 (1969)
45. Gething, P.W., Smith, D.L., Patil, A.P., et al.: Climate change and the global malaria recession. Nature **465**(7296), 342–345 (2010)
46. Gething, P.W., Van Boeckel, T.P., Smith, D.L., et al.: Modelling the global constraints of temperature on transmission of Plasmodium falciparum and P. vivax. Parasit. Vectors **4**(1), 92 (2011)
47. Ghani, A.C., Sutherland, C.J., Riley, E.M., et al.: Loss of population levels of immunity to malaria as a result of exposure-reducing interventions: consequences for interpretation of disease trends. PLoS One **4**(2), e4383 (2009)
48. Griffin, J.T., Hollingsworth, T.D., Okell, L.C., et al.: Reducing Plasmodium falciparum malaria transmission in Africa: a model-based evaluation of intervention strategies. PLoS Med. **7**(8), e1000324 (2010)
49. Griffin, J.T., Hollingsworth, T.D., Reyburn, H., et al.: Gradual acquisition of immunity to severe malaria with increasing exposure. Proc. R. Soc. Lond. B Biol. Sci. **282**(1801), 20142657 (2015)
50. Gu, W., Regens, J.L., Beier, J.C., et al.: Source reduction of mosquito larval habitats has unexpected consequences on malaria transmission. Proc. Natl. Acad. Sci. **103**(46), 17560–17563 (2006)
51. Gupta, S., Snow, R.W., Donnelly, C., et al.: Acquired immunity and postnatal clinical protection in childhood cerebral malaria. Proc. R. Soc. Lond. B Biol. Sci. **266**(1414), 33–38 (1999)
52. Gupta, S., Snow, R.W., Donnelly, C.A., et al.: Immunity to non-cerebral severe malaria is acquired after one or two infections. Nat. Med. **5**(3), 340–343 (1999)
53. Hay, S.I., Snow, R.W.: The malaria atlas project: developing global maps of malaria risk. PLoS Med. **3**(12), e473 (2006). Malaria Atlas Project (MAP). https://map.ox.ac.uk/, Accessed 18 August 2018
54. Hay, S.I., Cox, J., Rogers, D.J., et al.: Climate change and the resurgence of malaria in the East African highlands. Nature **415**(6874), 905–909 (2002)
55. Hay, S.I., Guerra, C.A., Tatem, A.J., et al.: The global distribution and population at risk of malaria: past, present, and future. Lancet Infect. Dis. **4**(6), 327–336 (2004)
56. Hayashi, M., Van der Kamp, G.: Simple equations to represent the volume–area–depth relations of shallow wetlands in small topographic depressions. J. Hydrol. **237**(1-2), 74–85 (2000)
57. Hoshen, M.B., Morse, A.P.: A weather-driven model of malaria transmission. Malar. J. **3**(1), 32 (2004)
58. Jepson, W., Moutia, A., Courtois, C.: The malaria problem in Mauritius: the bionomics of Mauritian anophelines. Bull. Entomol. Res. **38**(1), 177–208 (1947)
59. Karuri, S.W., Snow, R.W.: Forecasting paediatric malaria admissions on the Kenya Coast using rainfall. Glob. Health Action **9**(1), 29876 (2016)
60. Koenraadt, C., Githeko, A., Takken, W.: The effects of rainfall and evapotranspiration on the temporal dynamics of Anopheles gambiae s.s. and Anopheles arabiensis in a Kenyan village. Acta Trop. **90**(2), 141–153 (2004)
61. Lardeux, F.J., Tejerina, R.H., Quispe, V., et al.: A physiological time analysis of the duration of the gonotrophic cycle of Anopheles pseudopunctipennis and its implications for malaria transmission in Bolivia. Malar. J. **7**(1), 141 (2008)
62. Lindsay, S., Birley, M.: Climate change and malaria transmission. Ann. Trop. Med. Parasitol. **90**(5), 573–588 (1996)

63. Lunde, T.M., Bayoh, M.N., Lindtjørn, B.: How malaria models relate temperature to malaria transmission. Parasit. Vectors **6**(1), 20 (2013)
64. Lunde, T.M., Korecha, D., Loha, E., et al.: A dynamic model of some malaria-transmitting anopheline mosquitoes of the Afrotropical region. I. Model description and sensitivity analysis. Malar. J. **12**(1), 28 (2013)
65. Lysenko, A., Semashko, I.: Geography of malaria. a medico-geographic profile of an ancient disease. Itogi Nauk. Med. Geogr., 25–146 (1968)
66. Macdonald, G.: The Epidemiology and Control of Malaria. Oxford University Press, London (1957)
67. Martens, W., Niessen, L.W., Rotmans, J., et al.: Potential impact of global climate change on malaria risk. Environ. Health Perspect. **103**(5), 458–464 (1995)
68. Martens, P., Kovats, R., Nijhof, S., et al.: Climate change and future populations at risk of malaria. Glob. Environ. Chang. **9**, S89–S107 (1999)
69. Molineaux, L., Gramiccia, G., et al.: The Garki project: research on the epidemiology and control of malaria in the Sudan savanna of West Africa. Technical report, World Health Organization (WHO), Geneva (1980). http://garkiproject.nd.edu/access-garki-data.html, Accessed 25 January 2018
70. Mordecai, E.A., Paaijmans, K.P., Johnson, L.R., et al.: Optimal temperature for malaria transmission is dramatically lower than previously predicted. Ecol. Lett. **16**(1), 22–30 (2013)
71. Nájera, J.A., González-Silva, M., Alonso, P.L.: Some lessons for the future from the global malaria eradication programme (1955–1969). PLoS Med. **8**(1), e1000,412 (2011)
72. Nikolov, M., Bever, C.A., Upfill-Brown, A., et al.: Malaria elimination campaigns in the Lake Kariba region of Zambia: a spatial dynamical model. PLoS Comput. Biol. **12**(11), e1005192 (2016)
73. Okech, B.A., Gouagna, L.C., Walczak, E., et al.: The development of Plasmodium falciparum in experimentally infected Anopheles gambiae (Diptera: Culicidae) under ambient microhabitat temperature in western Kenya. Acta Trop. **92**(2), 99–108 (2004)
74. Okuneye, K., Gumel, A.B.: Analysis of a temperature-and rainfall-dependent model for malaria transmission dynamics. Math. Biosci. **287**, 72–92 (2017)
75. Okuneye, K., Eikenberry, S.E., Gumel, A.B.: Weather-driven malaria transmission model with gonotrophic and sporogonic cycles. J. Biol. Dynm. **13**(S1), 288–324 (2019).
76. Paaijmans, K.P., Wandago, M.O., Githeko, A.K., et al.: Unexpected high losses of Anopheles gambiae larvae due to rainfall. PLoS One **2**(11), e1146 (2007)
77. Paaijmans, K.P., Heusinkveld, B.G., Jacobs, A.F.: A simplified model to predict diurnal water temperature dynamics in a shallow tropical water pool. Int. J. Biometeorol. **52**(8), 797–803 (2008)
78. Paaijmans, K., Jacobs, A., Takken, W., et al.: Observations and model estimates of diurnal water temperature dynamics in mosquito breeding sites in western Kenya. Hydrol. Proced. Int. J. **22**(24), 4789–4801 (2008)
79. Paaijmans, K.P., Read, A.F., Thomas, M.B.: Understanding the link between malaria risk and climate. Proc. Natl. Acad. Sci. **106**(33), 13844–13849 (2009)
80. Paaijmans, K.P., Blanford, S., Bell, A.S., et al.: Influence of climate on malaria transmission depends on daily temperature variation. Proc. Natl. Acad. Sci. **107**(34), 15135–15139 (2010)
81. Paaijmans, K.P., Cator, L.J., Thomas, M.B.: Temperature-dependent pre-bloodmeal period and temperature-driven asynchrony between parasite development and mosquito biting rate reduce malaria transmission intensity. PLoS One **8**(1), e55777 (2013)
82. Paaijmans, K.P., Heinig, R.L., Seliga, R.A., et al.: Temperature variation makes ectotherms more sensitive to climate change. Glob. Chang. Biol. **19**(8), 2373–2380 (2013)
83. Packard, R.M.: The Making of a Tropical Disease: A Short History of Malaria. JHU Press, Baltimore (2007)
84. Parham, P.E., Michael, E.: Modeling the effects of weather and climate change on malaria transmission. Environ. Health Perspect. **118**(5), 620–626 (2010)
85. Parham, P.E., Pople, D., Christiansen-Jucht, C., et al.: Modeling the role of environmental variables on the population dynamics of the malaria vector Anopheles gambiae sensu stricto. Malar. J. **11**(1), 271 (2012)

86. Pascual, M., Bouma, M.J.: Do rising temperatures matter? Ecology **90**(4), 906–912 (2009)
87. Pascual, M., Ahumada, J.A., Chaves, L.F., et al.: Malaria resurgence in the East African highlands: temperature trends revisited. Proc. Natl. Acad. Sci. **103**(15), 5829–5834 (2006)
88. Perkins, S.L.: Malaria's many mates: past, present, and future of the systematics of the order haemosporida. J. Parasitol. **100**(1), 11–25 (2014)
89. Reiner, R.C., Perkins, T.A., Barker, C.M., et al.: A systematic review of mathematical models of mosquito-borne pathogen transmission: 1970–2010. J. R. Soc. Interface **10**(81), 20120921 (2013)
90. Rogers, D.J., Randolph, S.E.: The global spread of malaria in a future, warmer world. Science **289**(5485), 1763–1766 (2000)
91. Ryan, S.J., Ben-Horin, T., Johnson, L.R.: Malaria control and senescence: the importance of accounting for the pace and shape of aging in wild mosquitoes. Ecosphere **6**(9), 1–13 (2015)
92. Ryan, S.J., McNally, A., Johnson, L.R., et al.: Mapping physiological suitability limits for malaria in Africa under climate change. Vector Borne Zoonotic Dis. **15**(12), 718–725 (2015)
93. Scott, T.W., Takken, W.: Feeding strategies of anthropophilic mosquitoes result in increased risk of pathogen transmission. Trends Parasitol. **28**(3), 114–121 (2012)
94. Shlenova, M.: The speed of blood digestion in female A. maculipennis messae at stable effective temperature. Med. Parazit. Mosk. **7**, 716–735 (1938)
95. Singh, P., Yadav, Y., Saraswat, S., et al.: Intricacies of using temperature of different niches for assessing impact on malaria transmission. Indian J. Med. Res. **144**(1), 67–75 (2016)
96. Smith, D.L., Battle, K.E., Hay, S.I., et al.: Ross, Macdonald, and a theory for the dynamics and control of mosquito-transmitted pathogens. PLoS Pathog. **8**(4), e1002588 (2012)
97. Stocker, T.F., Qin, D., Plattner, G., et al.: Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC)–Summary for Policymakers. Climate Change 2013: The Physical Science Basis (2013)
98. Thomson, M.C., Mason, S.J., Phindela, T., et al.: Use of rainfall and sea surface temperature monitoring for malaria early warning in Botswana. Am. J. Trop. Med. Hyg. **73**(1), 214–221 (2005)
99. Thomson, M., Doblas-Reyes, F., Mason, S., et al.: Malaria early warnings based on seasonal climate forecasts from multi-model ensembles. Nature **439**(7076), 576–579 (2006)
100. Tompkins, A.M., Ermert, V.: A regional-scale, high resolution dynamical malaria model that accounts for population density, climate and surface hydrology. Malar. J. **12**(1), 65 (2013)
101. Trape, J.F., Rogier, C., Konate, L., et al.: The Dielmo project: a longitudinal study of natural malaria infection and the mechanisms of protective immunity in a community living in a holoendemic area of Senegal. Am. J. Trop. Med. Hyg. **51**(2), 123–137 (1994)
102. Vermeulen, S., Zougmore, R., Wollenberg, E., et al.: Climate Change, Agriculture and Food Security (CCAFS): A global partnership to link research and action for low-income agricultural producers and consumers.Curr. Opin. Environ. Sustain. **4**(1), 128–133 (2012). GCM Downscaled Data Portal. http://www.ccafs-climate.org/data_spatial_downscaling/
103. Webb Jr, James L.A.: The Long Struggle Against Malaria in Tropical Africa. Cambridge University Press, Cambridge (2014)
104. White, M.T., Griffin, J.T., Churcher, T.S., et al.: Modelling the impact of vector control interventions on Anopheles gambiae population dynamics. Parasit. Vectors **4**(1), 153 (2011)
105. World Health Organization (WHO): World Malaria Report 2015. Technical report, World Health Organization (WHO), Geneva (2015). http://www.who.int/malaria/publications/world-malaria-report-2015/report/en/
106. Yamana, T.K., Bomblies, A., Eltahir, E.A.: Climate change unlikely to increase malaria burden in West Africa. Nat. Clim. Chang. **6**(11), 1009 (2016)

Chapter 5
A Risk-Structured Mathematical Model of Buruli Ulcer Disease in Ghana

Christina Edholm, Benjamin Levy, Ash Abebe, Theresia Marijani, Scott Le Fevre, Suzanne Lenhart, Abdul-Aziz Yakubu, and Farai Nyabadza

Abstract This chapter discusses a mathematical model for the spread of an infectious disease with transmission through a pathogen in an environment, including the effects of human contact with the environment. The model assumes a structured susceptible population consisting of both "low-risk" and "high-risk" individuals. It also includes the effects of shedding the pathogen by the infected population into the environment. The model has a disease-free equilibrium state, and a linear stability analysis shows three possible transmission routes. The model is applied to Buruli ulcer disease, a debilitating disease induced by *Mycobacterium ulcerans*. There is some uncertainty about the exact transmission path, but the bacteria is known to live in natural water environments. The model parameters are estimated from data

C. Edholm · S. Lenhart (✉)
University of Tennessee, Knoxville, TN, USA

Department of Mathematics, Scripps College, Claremont, CA, USA
e-mail: cedholm@scrippscollege.edu; lenhart@math.utk.edu

B. Levy
Department of Mathematics, Fitchburg State University, Fitchburg, MA, USA
e-mail: blevy1@fitchburgstate.edu

A. Abebe
Department of Mathematics and Statistics, Auburn University, Auburn, AL, USA
e-mail: abebeas@auburn.edu

T. Marijani
Department of Mathematics, University of Dar es Salaam, Dar es Salaam, Tanzania

S. Le Fevre
Norwich University, Northfield, VT, USA
e-mail: slefevre@stu.norwich.edu

A.-A. Yakubu
Department of Mathematics, Howard University, Washington, DC, USA
e-mail: ayakubu@Howard.edu

F. Nyabadza
Department of Mathematics, University of Stellenbosch, Stellenbosch, South Africa
e-mail: nyabadzaf@sun.ac.za

H. G. Kaper, F. S. Roberts (eds.), *Mathematics of Planet Earth*, Mathematics of Planet Earth 5, https://doi.org/10.1007/978-3-030-22044-0_5

on Buruli ulcer disease in Ghana. This chapter includes a sensitivity analysis of the total number of infected individuals to the parameters in the model.

Keywords Infectious disease · Buruli ulcer disease · Pathogen in environment · Risk-structured model · Ghana

5.1 Introduction

Modeling infectious diseases with an indirect transmission route through pathogens in the environment is an area of significant current interest. The mechanisms and types of interaction terms to include vary depending on the disease. Some diseases have both direct transmission via infected individuals and indirect transmission via a contaminated environment. For example, cholera is frequently modeled with both direct and indirect transmission due to *cholera vibrios* in drinking or bathing water [20, 31, 43]. For Johne's disease in dairy cattle, the length of time that the pathogen (*Mycobacterium avium* subspecies *paratuberculosis*) can survive in a pasture depends on temperature. In fact, this disease has several transmission routes besides the environmental one, such as vertical transmission and contaminated milk and colostrum [11, 28]. As Breban showed, the length of time that a pathogen can survive in the environment gives an indication which transmission routes should be included in a model [8].

In this chapter, we turn our attention to diseases that have only one transmission route, which is indirect through pathogens in the environment. We illustrate this case with data and simulations for Buruli ulcer disease.

Outline of the Chapter Section 5.2 gives details of the Buruli ulcer (BU) disease and describes results of earlier research. Section 5.3 introduces a mathematical model that applies to BU and similar diseases. The model consists of five coupled ordinary differential equations, one for each of five state variables. The state variables are the number of susceptible individuals at "high" risk for coming into contact with the pathogen, the number of susceptible individuals at "low" risk for coming into contact with the pathogen, the number of infected individuals, the number of individuals undergoing treatment, and the concentration of pathogens in the environment. The model includes the effects of shedding by the infected population of the pathogen into the environment. Section 5.4 gives some basic properties of the model. The disease-free equilibrium (DFE) state is introduced in Sect. 5.5, and the basic reproductive number in Sect. 5.6. Section 5.7 is devoted to a linear stability analysis of the DFE state. The model involves a number of parameters, which are estimated in Sect. 5.8 from data for BU disease in Ghana. Section 5.9 contains numerical results based on these estimated parameter values. The endemic equilibrium state is discussed in Sect. 5.10. Section 5.11 contains a sensitivity analysis for the total number of infected individuals to variations in the parameters. This chapter concludes with a summary and outline of future work in Sect. 5.12.

5.2 Buruli Ulcer Disease

Buruli ulcer (BU) disease has been reported in over 30 countries on four continents. It is not limited to tropical environments, as can be seen in Fig. 5.1, but most cases occur in Africa, especially in the western and central regions [51]. Cases of BU disease are concentrated in rural areas near still-water sources [21, 37, 40]. After tuberculosis and leprosy, BU disease is the third most common human mycobacterial disease, and after tuberculosis the second most prevalent mycobacterial disease in Ghana [2].

Cases of BU disease were recorded as far back as 1897 [9], but more extensive research on this disease was not completed until the 1940s [19, 23, 37]. In the 1980s, BU disease gained attention because of an increase of cases, possibly linked to changes in the environment [30]. In 1998, the World Health Organization (WHO) created the Global Buruli Ulcer Initiative, which implemented control programs that contributed to the reduction of cases in more recent years [37, 47]. We also note that climate plays a role in BU disease due to the fact that the bacteria live in water environments, which are affected by climate change [32, 46].

Mycobacterium ulcerans (MU), which causes dermal BU disease, is becoming a debilitating affliction in many countries worldwide. MU causes an infection of the skin and soft tissue, which can lead to permanent disfigurement and disability [52]. The disease first presents itself as a small bump under the skin, varying in appearance based on the strain of MU, and then, if left untreated, BU disease can progress to a large lesion, possibly requiring amputation [14]. MU is slow-growing but does not grow in moving water [37]. MU can adapt to a dark environment; it has

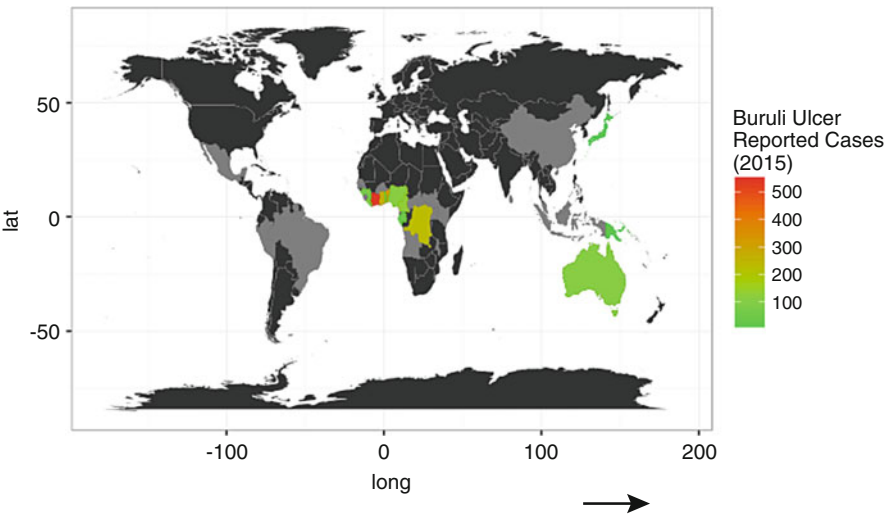


Fig. 5.1 Global Buruli ulcer disease case map. Data from the World Health Organization (WHO) (http://apps.who.int/gho/data/node.main.A1631)

been identified in various organisms around the water such as fish, plants, or insects. There are different strains, which may affect the likelihood of human infection; in general, low-risk strains are prevalent in America and Asia, while Africa and Australia have more high-risk infective strains [37].

Transmission of MU from the environment to humans remains a topic of uncertainty, despite much recent research. Portaels et al. [36] put forth the idea that water bugs from several plants in swamps in Benin and Ghana ingest water-filtering organisms with concentrations of MU. Such bugs would infect humans by biting, and the trauma associated with the bite seems to be necessary for infection. Subsequently, various scientific groups studied environments with hosts and water bugs to explore the possibility that water insect vectors are needed for infection. Marsollier et al. [27] conducted a study of water bugs biting the tails of mice that had been submerged in water and interpreted the development of BU disease in the mice as evidence of transmission by insect bite. De Silva et al. [10] reported that there is no human-to-human infection, and that water bugs previously associated with BU infection do not deliberately bite humans. Benbow et al. [3, 4] expressed caution due to the uncertainty about the transmission of MU.

Williamson et al. [48] conducted a study in Ghana focusing on both endemic and non-endemic villages. Oftentimes, villages in the study were relatively close, using the same watershed, but one village would have BU disease, while the other villages would not. They concluded that key elements of MU infection and transmission are human contact with water and focal demography. Marion et al. [26] investigated MU dynamics in Cameroon using mice to explore water bugs as vectors for MU. Williamson et al. [49] also showed that people do contribute to the spread and distribution of MU in the environment. In a follow-up study, Williamson et al. [50] investigated whether skin abrasions would lead to a BU infection in pigs and humans. They found that, to become infected, there needed to be an injection underneath the skin, which could result from specific puncture wounds or possibly an insect bite. Most recently, Morris et al. [33] studied the location and spread of MU and discovered that MU infects many organisms, none of which had been considered hosts, and these organisms are critical to the MU life cycle. They also found that, since MU does not reproduce rapidly, the bacterium had evolved to be able to move between various sources such as humans.

The idea of water insect vectors (water bugs) having a role in infection has been used in two epidemiological models with systems of ordinary differential equations [7, 34]. These models include transmission from bites of water bugs, the pathogen in the environment, and fish populations that prey on infected water bugs. Garchitorena et al. [17] considered different routes of transmission, using Cameroon data and spatial statistical models. Water bugs infecting humans were included as a mode of transmission along with environmental contamination. They concluded that environmental transmission was more important than transmission through water bugs. Lastly, we mention the application of an individual-based model by Garchitorena et al. [16], which focuses on the economic impacts of the disease on population-level inequalities, without the inclusion of water bugs.

5.3 Model Formulation

In this section, we formulate a model to study how increased risk of a human subpopulation contact with an environment containing a pathogen affects the spread of a disease, with only indirect transmission.

The state of the system is described at any time by five variables. Two state variables are S_L, the number of susceptible individuals (susceptibles) at "low" risk of infection from contact with the pathogen in the environment, and S_H, the number of susceptibles at "high" risk of infection from contact with the pathogen in the environment. The three additional state variables are I, the number of infected individuals; T, the number of infected individuals undergoing treatment; and B, the pathogen density in the environment; see Table 5.1.

The state variables change over time due to various interactions among the classes of individuals and the environment, as indicated in the diagram in Fig. 5.2. Low-risk susceptibles can transfer into the class of high-risk susceptibles and vice versa. Recruitment into the class of low-risk susceptibles takes place through

Table 5.1 The model variables and their description

Variable	Description
S_L	Number of susceptible individuals at low risk
S_H	Number of susceptible individuals at high risk
I	Number of infected individuals
T	Number of infected individuals undergoing treatment
B	Concentration of pathogen in the environment

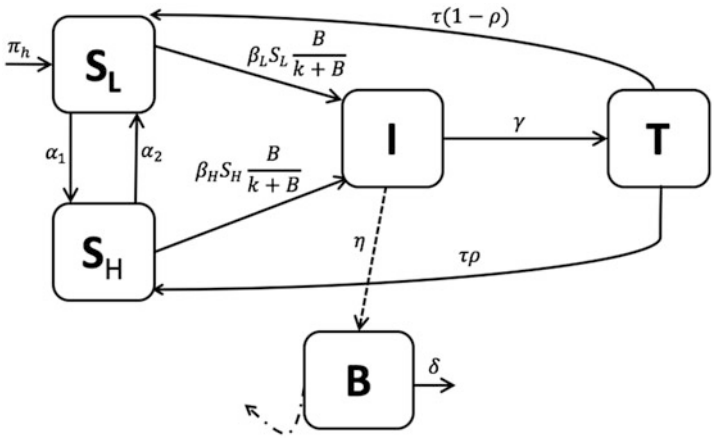


Fig. 5.2 Flow diagram for the BU disease model. Solid lines represent flows between classes, the straight dashed line represents the infected class shedding pathogen into the environment, and the solid curved lines represent the source of new infections resulting from susceptibles interacting with the pathogen in the environment

external sources like birth and immigration. There is indirect transmission into the infected class when individuals from either class of susceptibles make contact with the environment. Since BU disease is treatable, infected individuals transition into the class of individuals being treated, from which they transition back into the either of the two susceptible classes. Lastly, individuals may die a natural death at every stage.

We assume that the class of pathogens in the environment has logistic growth. It is possible that the carrying capacity changes over time due to variations in rainfall, temperature, etc. The pathogens in the environment spread the disease to susceptibles when the latter come into contact with the environment. Importantly, the infected class can also shed the pathogen into the environment. The pathogen in the environment decays at its natural rate.

These interactions are captured in the following system of ordinary differential equations (ODEs, the prime indicates differentiation with respect to time):

$$\begin{aligned}
S_L' &= -\left(\alpha_1 + \beta_L \frac{B}{k+B} + \mu\right) S_L + \alpha_2 S_H + (1-\rho)\tau T + \pi_h, \\
S_H' &= \alpha_1 S_L - \left(\alpha_2 + \beta_H \frac{B}{k+B} + \mu\right) S_H + \rho\tau T, \\
I' &= \beta_L \frac{B}{k+B} S_L + \beta_H \frac{B}{k+B} S_H - (\mu+\gamma)\, I, \\
T' &= \gamma\, I - (\mu+\tau)\, T, \\
B' &= r\left(1 - \frac{B}{K_1(t)}\right) S_H + \eta\, I - \delta\, B.
\end{aligned} \tag{5.3.1}$$

The parameters in the model are described in Table 5.2.

Table 5.2 The model parameters and their description

Parameter	Description
α_1	Transition rate from S_L to S_H
α_2	Transition rate from S_H to S_L
β_L	Transmission rate from S_L to I
β_H	Transmission rate from S_H to I
τ	Recovery rate with treatment
ρ	Proportion of those transitioning out of T into S_H
γ	Treated rate for infected individuals
μ	Natural death rate
π_h	Recruitment of susceptibles
K_1	Carrying capacity of B in the environment
r	Intrinsic growth rate of B
η	Shedding rate of infected individuals
δ	Rate of decay of B
k	Half-saturation constant

5.4 Basic Properties

Here, we examine some basic properties of the system (5.3.1). The total human population is made up of susceptibles, infected individuals, and infected individuals undergoing treatment, so if $N(t)$ is the total number of individuals, then

$$N(t) = S_L(t) + S_H(t) + I(t) + T(t)$$

at any time t. Adding the ODEs for S_L, S_H, I, and T, we obtain an ODE for N,

$$N' = \pi_h - \mu N.$$

Since π_h is constant in time, the equation is readily integrated,

$$N(t) = \frac{\pi_h}{\mu} + \left(N(0) - \frac{\pi_h}{\mu}\right) e^{-\mu t}.$$

In particular,

$$N_\infty = \lim_{t\to\infty} N(t) = \pi_h/\mu,$$

so $\pi_h = \mu N_\infty$. Furthermore, if $0 < N(0) \leq N_\infty$, then $0 < N(t) \leq N_\infty$ for all $t > 0$.

If there exist positive constants K_2 and K_3 such that $K_2 \leq K_1(t) \leq K_3$ for all $t > 0$, then B satisfies the differential inequality

$$B' \leq r(1 - B/K_3)B - \delta B + \eta N_\infty.$$

Let $B_\infty = (K_3/2r)\left(r - \delta + \left((r-\delta)^2 + 4\eta r N_\infty/K_3\right)^{1/2}\right)$. Note that B_∞ is the stable equilibrium solution of the differential inequality for B with strict equality (i.e., $\leq$ replaced by $=$). If $0 < B\,(0) \leq B_\infty$, then $0 < B(t) \leq B_\infty$ for all $t > 0$.

Hence, any solution of the system (5.3.1) that starts inside the domain Ω,

$$\Omega = \left\{(S_L, S_H, I, T, B) \in \mathbb{R}^5_+ \,|\, 0 < S_L + S_H + I + T \leq N_\infty,\ 0 < B \leq B_\infty\right\}$$

remains inside Ω as time progresses [39].

5.5 Disease-Free Equilibrium

We assume for simplicity that the carrying capacity K_1 of the pathogen in the environment is constant.

The system (5.3.1) has a disease-free equilibrium (DFE) solution, where every human is susceptible and there are no *Mycobacterium ulcerans* in the environment. This solution corresponds to the point P_0 in state space,

$$P_0 = \left(S_L^*, S_H^*, 0, 0, 0\right), \tag{5.5.1}$$

where

$$S_L^* = \frac{\alpha_2 + \mu}{\alpha_1 + \alpha_2 + \mu} N_\infty, \quad S_H^* = \frac{\alpha_1}{\alpha_1 + \alpha_2 + \mu} N_\infty.$$

Here we have used the relation $\pi_h = \mu N_\infty$. In Sect. 5.8, we will use estimated values for the parameters for BU disease in Ghana; in that case, $\alpha_2 + \mu > \alpha_1$, so $S_L^* > S_H^*$.

5.6 Basic Reproduction Number

To find the basic reproduction number, $\mathscr{R}_0$, we use the next-generation matrix approach [12, 13, 44, 45]. We rewrite the system (5.3.1) as a system of equations for the vectors x and y,

$$x = \begin{pmatrix} I \\ B \end{pmatrix}, \quad y = \begin{pmatrix} S_L \\ S_H \\ T \end{pmatrix}.$$

Assuming that shedding is not a new infection, we are dealing with a system of the form

$$\begin{aligned} x' &= \mathscr{F}(x, y) - \mathscr{V}(x, y), \\ y' &= g(x, y), \end{aligned}$$

where $\mathscr{F}$ incorporates new infections and $\mathscr{V}$ transition terms,

$$\mathscr{F} = \begin{pmatrix} \beta_L \frac{B}{k+B} S_L + \beta_H \frac{B}{k+B} S_H \\ r\left(1 - \frac{B}{K_1}\right) B \end{pmatrix}, \quad \mathscr{V} = \begin{pmatrix} (\gamma + \mu) I \\ -\eta I + \delta B \end{pmatrix}.$$

(Later, we will compare with the case where shedding is considered a new infection.)

The Jacobian matrices F and V associated with $\mathscr{F}$ and $\mathscr{V}$, respectively, at P_0 are

$$F = \begin{pmatrix} 0 & (\beta_L S_L^* + \beta_H S_H^*)/k \\ 0 & r \end{pmatrix}, \quad V = \begin{pmatrix} \gamma + \mu & 0 \\ -\eta & \delta \end{pmatrix}.$$

Since F is nonnegative and V is a nonsingular M-matrix, the basic reproduction number $\mathscr{R}_0$ is the spectral radius of FV^{-1}. Thus,

$$\mathscr{R}_0 = \mathscr{R}_{0,B} + \mathscr{R}_{0,L} + \mathscr{R}_{0,H},$$

with

$$\mathscr{R}_{0,B} = \frac{r}{\delta}, \quad \mathscr{R}_{0,L} = \frac{\beta_L \eta}{(\gamma+\mu)\delta k} S_L^*, \quad \mathscr{R}_{0,H} = \frac{\beta_H \eta}{(\gamma+\mu)\delta k} S_H^*.$$

That is, $\mathscr{R}_0$ is determined by the pathogen growth rate in the environment and the transmission rates from the pathogen to the susceptible classes S_L and S_H. If there is no shedding by the infected individuals into the environment ($\eta = 0$), then $\mathscr{R}_0 = \mathscr{R}_{0,B}$. In this case, if $r > \delta$, the disease will persist as the pathogen grows in the environment. If there is shedding from infected individuals into the environment ($\eta > 0$), the two additional terms characterize how the susceptible classes become infected. Specifically, the term $\mathscr{R}_{0,L}$ contains the parameters that account for the transition from low-risk susceptible to infected individuals; the sum $\gamma + \mu$ is the rate at which the infected individuals leave the infected class. Meanwhile, β_L/k relates to the transition of susceptibles into the infected class as a result of an interaction with the pathogen in the environment. Lastly, η/δ is the fraction of the rate of infected shedding MU over the pathogen decay rate. The interpretation of the term $\mathscr{R}_{0,H}$ is similar.

5.7 Stability Analysis

The following local stability result follows from [44, Theorem 2].

Theorem 1 *If $\mathscr{R}_0 < 1$, the disease-free equilibrium state P_0 is locally asymptotically stable. If $\mathscr{R}_0 > 1$, P_0 is unstable.*

In other words, if $\mathscr{R}_0 < 1$, a small outbreak of the BU disease will be eradicated in the course of time, but if $\mathscr{R}_0 > 1$, the BU disease will persist.

Clearly, $\mathscr{R}_0 > 1$ if any one of $\mathscr{R}_{0,B}$, $\mathscr{R}_{0,L}$, or $\mathscr{R}_{0H}$ is greater than 1. Hence, in addition to the pathogen concentration in the environment, both high- and low-risk transmission paths must be controlled in order to eliminate the BU disease.

The next theorem addresses the global stability of P_0.

Theorem 2 *If $\mathscr{R}_0 < 1$, the disease-free equilibrium state P_0 is globally asymptotically stable in Ω.*

In other words, if $\mathscr{R}_0 < 1$, the BU disease will be eradicated.

Proof Again using the x and y notation, we let $Q = \omega^T V^{-1} x$, where ω^T is the left Perron eigenvector of $V^{-1}F$ corresponding to the eigenvalue $\mathscr{R}_0$. (Recall that $V^{-1}F$

is irreducible.) Proceeding as in Ref. [38], we write the x differential equations in the form

$$\frac{dx}{dt} = \mathscr{F}(x, y) - \mathscr{V}(x, y) = (F - V)\, x - f(x, y),$$

where F and V are the Jacobian matrices associated with $\mathscr{F}$ and $\mathscr{V}$, respectively, at P_0, and

$$f(x, y) = \begin{pmatrix} \left(\beta_L \left(S_L^*/k - S_L/(k+B)\right) + \beta_H \left(S_H^*/k - S_H/(k+B)\right)\right) B \\ r B^2/K_1 \end{pmatrix}.$$

Clearly, $r B^2/K_1 \geq 0$. Let $A = \min\{\beta_L, \beta_H\}$. Since $S^* + S_H^* = N_\infty$ and $S_L + S_H \leq N_\infty$,

$$\left(\beta_L \left(\frac{S_L^*}{k} - \frac{S_L}{k+B}\right) + \beta_H \left(\frac{S_H^*}{k} - \frac{S_H}{k+B}\right)\right) B \geq \frac{AB}{k+B} \left(N_\infty - (S_L + S_H)\right) \geq 0.$$

Hence, $f(x, y) \geq 0$. It follows that the vector Q satisfies the differential inequality

$$\frac{dQ}{dt} \leq (\mathscr{R}_0 - 1)\, \omega^T x - \omega^T V^{-1} f(x, y) \leq (\mathscr{R}_0 - 1)\, \omega^T x \leq 0 \text{ in } \Omega.$$

Hence, Q is a Lyapunov function. The statement of the theorem follows from LaSalle's invariance principle [22].

Recall that, in the definition of $\mathscr{F}$ and $\mathscr{V}$, we assumed that shedding was not a new infection. If shedding is regarded as a new infection, then

$$F = \begin{pmatrix} 0 & \left(\beta_L S_L^* + \beta_H S_H^*\right)/k \\ \eta & r \end{pmatrix}, \quad V = \begin{pmatrix} \gamma + \mu & 0 \\ 0 & \delta \end{pmatrix},$$

and $\hat{\mathscr{R}}_0$ is the positive root of the quadratic equation

$$f(\lambda) = \lambda^2 - \mathscr{R}_{0,B}\lambda - (\mathscr{R}_{0,L} + \mathscr{R}_{0,H}) = 0.$$

From the signs of the coefficients, this equation has a unique positive root, $\hat{\mathscr{R}}_0$. If $\hat{\mathscr{R}}_0 < 1$, then $f(1) > 0$ and $\mathscr{R}_0 = \mathscr{R}_{0,B} + \mathscr{R}_{0,L} + \mathscr{R}_{0,H} < 1$. Similarly, if $\hat{\mathscr{R}}_0 > 1$, then $f(1) < 0$ and $\mathscr{R}_0 = \mathscr{R}_{0,B} + \mathscr{R}_{0,L} + \mathscr{R}_{0,H} > 1$.

Note that $f(1) = 1 - (\mathscr{R}_{0,B} + \mathscr{R}_{0,L} + \mathscr{R}_{0,H}) = 1 - \mathscr{R}_0$, giving the same threshold as derived for $\mathscr{R}_0$. Thus, using $\hat{\mathscr{R}}_0$ instead of $\mathscr{R}_0$, Theorems 1 and 2 remain true when shedding is regarded as new infection.

In the next section, we calculate the basic reproduction numbers $\mathscr{R}_0$ and $\hat{\mathscr{R}}_0$ for BU disease in Ghana, using estimated parameter values obtained from actual data.

5.8 Parameter Estimation

Table 5.3 shows the monthly number of BU disease cases in Ghana for the period 2008–2015. The last column lists the total population of Ghana during the same period.

Figure 5.3 shows the total population, together with a best linear fit.

We use the slope of the linear function to represent the growth of susceptibles, π_h. For the natural death rate, we use the life expectancy in Ghana, which is approximately 61 years [41]; hence, $\mu = 1/(61 \times 365)$, using days as the underlying time unit. The remaining parameters in the model are estimated by minimizing the L^2 norm of the difference between the simulation results and the data at the corresponding time points,

$$J(\theta) = \|MI(\theta) - MI^*\|_2.$$

Here, θ is the vector of (unknown) parameter values, $MI(\theta)$ is the vector with the number of monthly infections of BU disease in the numerical simulation, and MI^* is the vector with the corresponding data. The minimization is done with the function `fmincon` in the MATLAB optimization toolbox. Since `fmincon` is a local solver, the multistart algorithm allows for a full exploration of the parameter space, using a large number of starting points to find the global minimum. The function `fmincon` also allows for linear inequality constraints on some of the parameters, which we use to mimic ecological constraints. For instance, with respect to the transitions between S_L and S_H, we assume that there are far fewer people transitioning to an area, job, or lifestyle where they will experience a drastic increase in exposure to the water environment. We model this constraint by imposing the constraint $\alpha_1 < 0.2\alpha_2$ as part of the parameter-fitting process. Additionally, the S_H class contains susceptibles with increased water contact. Assuming that their transmission rate to infected is much greater than that of susceptibles with normal water contact, S_L, we impose the constraint $\beta_L < 0.2\beta_H$ on the transmission rates.

Using a total population figure of 23 million individuals, we start the minimization procedure by assigning 2% of the individuals to the class of high-risk susceptibles S_H and set the initial values $B(0) = 1000$, $I(0) = 0$, and $T(0) = 0$. The resulting estimates of the parameter values are listed in the third column of Table 5.4.

5.9 Numerical Results

Figure 5.4 shows the number of infected individuals computed with the parameter values given in Table 5.4 and the monthly data of infected individuals given in Table 5.3. The computational results match the data quite well; therefore, it is reasonable to adopt the numerical values of the parameters given in Table 5.4.

Table 5.3 Number of BU disease cases by month, together with the annual number, in Ghana for the period 2008–2015

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Annual	Population
2015	19	18	13	17	20	17	16	26	22	12	19	13	212	27.4
2014	8	26	20	13	10	18	33	16	14	11	12	30	211	26.8
2013	43	57	58	38	41	27	36	25	53	27	28	35	468	26.2
2012	84	69	72	45	57	31	28	37	30	51	38	25	567	25.5
2011	77	74	76	47	48	77	65	130	59	98	95	75	921	24.9
2010	74	66	84	64	62	70	46	60	88	76	82	59	831	24.3
2009	81	49	40	48	77	98	64	93	60	95	63	56	824	23.7
2008	110	41	29	30	57	57	113	68	46	57	53	42	703	23.1

Data from the Ministry of Health, Ghana Health Service [18]. The last column lists the total population (in millions) of Ghana during the same period. Data from Worldometers [53]

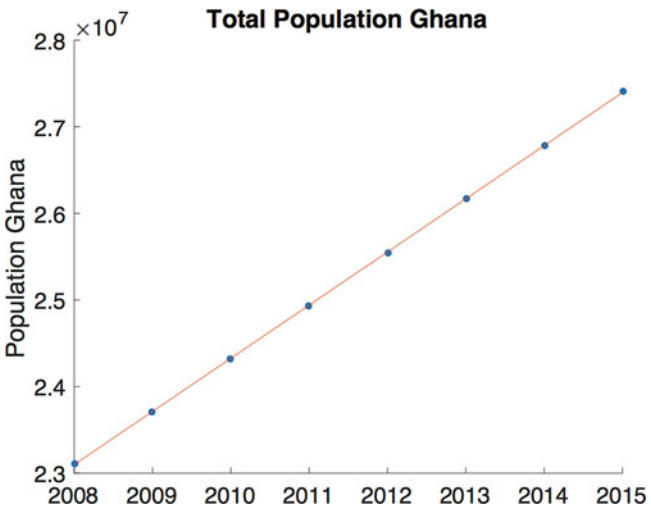


Fig. 5.3 Scatterplot of the total population (in millions) of Ghana for the period 2008–2015, together with a best linear fit

Table 5.4 Model parameters and their estimated values obtained with the `fmincon` function in the MATLAB optimization toolbox

Parameter	Description	Value
α_1	Transition rate from S_L to S_H	0.00065
α_2	Transition rate from S_H to S_L	0.1
β_L	Transmission rate from S_L to I	1.00×10^{-10}
β_H	Transmission rate from S_H to I	1.47×10^{-5}
τ	Recovery rate with treatment	0.17097
ρ	Proportion of those transitioning out of T into S_H	0.89916
γ	Treated rate for infected individuals	0.01000008
μ	Natural death rate	$1/(61 \times 365)$
π_h	Recruitment of susceptibles	$(1/365) \times 10^6$
K_1	Carrying capacity of B in the environment	9671.63899
r	Intrinsic growth rate of B	0.001000073
η	Shedding rate of infected	3.37×10^{-6}
δ	Rate of decay of B	0.00563
k	Half-saturation constant	0.2500004

We remark that the same procedure applied to a model with a single susceptible population did not fit the data as well as the model with two susceptible populations.

Connecting back to Sect. 5.5, we calculate the value of the basic reproductive numbers with these parameter values. If shedding is considered a new infection, then $\hat{\mathscr{R}}_0 = 1.2644$; otherwise, $\mathscr{R}_{0,B} = 0.1778$, $\mathscr{R}_{0,L} = 0.0014$, and $\mathscr{R}_{0,H} = 1.3725$, giving $\mathscr{R}_0 = 1.5518$. Therefore, the disease-free equilibrium is unstable in both cases.

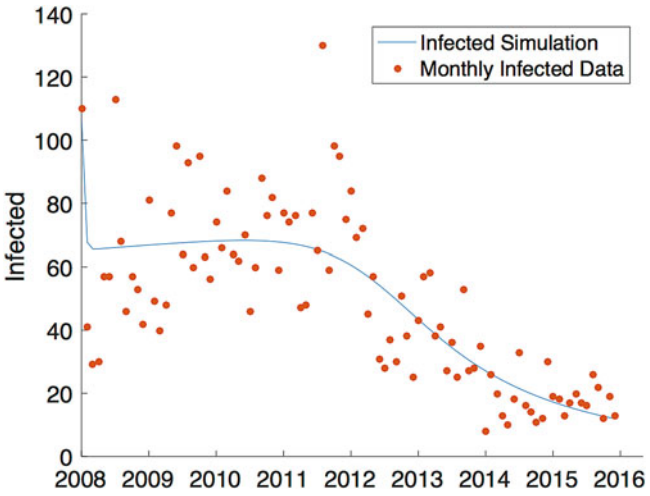


Fig. 5.4 BU disease cases in Ghana for the period 2008–2015. Blue curve: computed from Eqs. (5.3.1); red dots: actual monthly data from Table 5.3

5.10 Endemic Equilibrium

As before, we assume for simplicity that the carrying capacity K_1 of the pathogen in the environment is constant.

The model (5.3.1) admits an endemic equilibrium state, in addition to the disease-free equilibrium state discussed in Sect. 5.5. It satisfies the system (5.3.1) where the left-hand sides are all zero and is represented by the point P_1 in state space,

$$P_1 = (S_L^{**}, S_H^{**}, I^{**}, T^{**}, B^{**}). \tag{5.10.1}$$

For BU disease in Ghana, using the parameter values given in Table 5.4, we find

$$P_1 = (27,000,000,\ 3344,\ 32,118,527,\ 1,878,129,\ 169,669).$$

The Jacobian of the system (5.3.1) at P_1 (omitting double asterisks) is

$$\begin{pmatrix} -(\alpha_1 + \beta_L \frac{B}{k+B} + \mu) & \alpha_2 & 0 & (1-\rho)\tau & -\beta_L \frac{k}{(k+B)^2} S_L \\ \alpha_1 & -(\beta_H \frac{B}{k+B} + \alpha_2 + \mu) & 0 & \rho\tau & -\beta_H \frac{k}{(k+B)^2} S_H \\ \beta_L \frac{B}{k+B} & \beta_H \frac{B}{k+B} & -(\gamma+\mu) & 0 & 0 \\ 0 & 0 & \gamma & -(\tau+\mu) & 0 \\ 0 & 0 & \eta & 0 & (1-2B/K_1)r - \delta \end{pmatrix}.$$

With the values found above, the eigenvalues of the Jacobian are all negative, which indicates that the endemic equilibrium is stable. Note, however, that the equilibrium values are far from the current BU situation in Ghana.

5.11 Sensitivity Analysis

We performed a global sensitivity analysis, using the total number of infected individuals I as the outcome variable. We applied Latin Hypercube Sampling (LHS) to sample the parameter space and Partial Rank Correlation Coefficients (PRCC) to evaluate the sensitivity of the outcome variable to variations in the input variables [6, 25]. We included all the parameters from Table 5.2, except μ and π_h which are known demographic parameters for Ghana. Each parameter was varied over an interval from 50% below to 50% above the value given in Table 5.4 with a uniform probability distribution over the interval. Following the recommendation in [29], we took $N > 4K/3$ draws for the LHS sampling scheme, where K is the number of input parameters; in our case, $K = 12$ and $N = 50$.

PRCC is a powerful technique for assessing the monotonicity of the relationship between the total number of infected individuals and a particular parameter while holding the remaining parameters constant. Unlike the Pearson product–moment correlation, PRCC provides meaningful results even when the relationship is not linear. We applied the Fisher transformation to the PRCC as in [15, 24]. We calculated the PRCC between the output (total number of infected individuals) and a parameter $\hat{\rho}_i$, $i = 1, \ldots, K$, using the partial regression approach described in [25]. The Fisher transformation is

$$F_i = \frac{1}{2} \log \left(\frac{1 + \hat{\rho}_i}{1 - \hat{\rho}_i} \right),$$

and F_i follows an approximate Gaussian distribution with zero mean and variance $(N - K - 3)^{-1}$ if the true mean of $\hat{\rho}_i$ is 0 [24]. To test the significance of the ith PRCC, we calculated the p-values corresponding to $\hat{\rho}_i$ by calculating the probability that the absolute value of a randomly sampled value from the standard Gaussian distribution exceeds $|F_i|\sqrt{N - K - 3}$. Since we performed multiple ($K = 12$) hypothesis tests, we corrected the resulting p-values using the false discovery rate (FDR) method of [5]. PRCC values with an adjusted p-value less than 0.01 are considered significantly different from 0. The results of the sensitivity analysis are displayed in Table 5.5; statistically significant PRCC values are indicated in red.

Note that the PRCC values for α_1, α_2, β_H, γ, δ, and k are statistically significant based on the adjusted p-values. The parameters α_1 and β_H show significant positive correlation with the total number of infected individuals, while α_2, γ, δ, and k show significant negative correlation with the number of infected individuals.

Next, we ranked the six parameters identified as significant in Table 5.5 with respect to their correlation with the total number of infected individuals, using

Table 5.5 Model parameters, corresponding PRCC and FDR adjusted p-values resulting from the sensitivity analysis

Variable	PRCC	p-value
α_1	0.9299	<0.0001
α_2	−0.9426	<0.0001
β_L	−0.2334	0.2634
β_H	0.9243	<0.0001
τ	−0.2214	0.2650
ρ	−0.0584	0.7699
γ	−0.4902	0.0027
η	−0.0487	0.7699
δ	−0.9292	<0.0001
K_1	−0.1791	0.3697
r	0.1327	0.5080
k	−0.4882	0.0027

Significant values ($p < 0.01$) are highlighted in red

Table 5.6 FDR adjusted p-values using comparisons of Fisher transforms of the PRCC values

	α_2	β_H	γ	δ	k
α_1	<0.0001	0.9279	<0.0001	<0.0001	<0.0001
α_2		<0.0001	<0.0001	0.7435	<0.0001
β_H			<0.0001	<0.0001	<0.0001
γ				<0.0001	0.9910
δ					<0.0001

pairwise testing of all the $\binom{6}{2}$ PRCC values. To evaluate the difference $\hat{\rho}_i - \hat{\rho}_j$, we use $F_i - F_j$, which follows an approximate Gaussian distribution $N(0, \sigma^2)$ with $\sigma^2 = 2(N - K - 3)$ if the true PRCC values are equal. Thus, pairwise difference p-values are calculated and FDR corrections are applied accordingly as described above for PRCC significance testing. The results are given in Table 5.6; PRCC values that differ significantly are indicated in red.

We were unable to differentiate the PRCC values with the total number of infected individuals of the parameter pairs $\{\alpha_1, \beta_H\}$, $\{\alpha_2, \delta\}$, and $\{\gamma, k\}$; however, all other pairs are identified as significantly different. The PRCC values given in Table 5.5 yield the following ranking of the parameters with respect to their correlation with the total number of infected individuals: (1) α_1 and β_H have a strong positive influence on the output, (2) α_2 and δ have a strong negative influence on the output, (3) γ and k have a moderate negative influence on the output.

The strong influence of α_1 and β_H is to be expected; α_1 moves individuals into S_H, and β_H is a transmission rate from bacteria into the environment. On the other hand, the moderate influence of α_2 and δ can be explained by the fact that α_2 moves individuals out of S_H into S_L and δ is a decay rate for the bacteria in the environment which is responsible for transmission.

5.12 Conclusions

In this chapter, we studied a Buruli ulcer (BU) disease model where susceptibles are separated into two classes based on their risk for exposure to a pathogen in the environment. The model admits three transmission paths, and the resulting expression for the basic reproduction number $\mathscr{R}_0$ consists accordingly of three parts. The model admits a disease-free equilibrium (DFE) state, which is unstable whether or not the shedding of the pathogen into the environment is considered a new infection in the Next Generation Matrix Method.

We applied the model to the situation in Ghana, using monthly data for the BU disease for the period 2008–2015. The data were supplied by the Ghana Health Service. Ghana is a BU endemic region, albeit the number of new BU infections appears to be decreasing. From the data we estimated values of all the model parameters. The results of numerical computations showed a reasonable fit with the data. However, it should be kept in mind that the model may not accurately represent the disease transmission mechanisms and the environmental conditions.

When shedding is regarded as a new infection, we obtain the value $\hat{\mathscr{R}}_0 = 1.2644$ for the basic reproduction number. However, when shedding is not regarded as a new infection, we obtain the value $\mathscr{R}_0 = 1.5518$, and each of the three constituent parts of $\mathscr{R}_0$ is less than one. Therefore, targeting high-risk groups such as people in rural areas with no access to a continuous supply of clean water should be considered in formulating public health policies to reduce the impact of BU disease.

A global sensitivity analysis of the impact of parameters on the total number of infected individuals leads to the following conclusions about correlations. The transmission rate of new infections from both susceptible classes has a high positive impact on the output. As expected, reduction of human contact with MU-contaminated water leads to fewer infections. Additionally, we found a strong negative influence associated with the decay rate of MU in the environment, implying that a reduction of the amount of MU in the environment reduces the number of infections. The transition rates between the two susceptible classes also strongly influence the output, since fewer susceptibles in the higher water-contact class lead to reduced contact with the pathogen.

Future work will include a more detailed study of BU disease dynamics. Specifically, data at a smaller spatial scale, like villages or small towns, will enable a better representation of the interaction of humans and their water environment.

The model framework presented in this chapter, with two classes of susceptibles based on the risk of contact with a pathogen in the environment, can be applied to a variety of diseases. This type of indirect transmission may be strongly affected by changes in temperature, precipitation, and other environmental features. Much of the research connecting climate change and infectious diseases has focused on the role of temperature, more than other environmental features [1]. But as some recent work has demonstrated, there are many links between climate change and the dynamics of infectious diseases requiring attention. For example, warmer temperatures and heavy rainfall are linked to diarrheal diseases through contaminated water and

undercooked seafood [35]. Similarly, a loss of biodiversity affects the spread of infectious diseases among wildlife [42].

Environmental health and the health of wildlife, domestic animals, and humans are all part of *One Health*—the concept that the health (and ultimate survival) of any one group is intimately related to the health of the others. Viewing the effects of climate change and the dynamics of infectious diseases through the lens of One Health is one of the urgent challenges facing the scientific community.

Acknowledgements The authors acknowledge partial support from the National Science Foundation (NSF) under Grant No. 1343651 through the Southern Africa Mathematical Sciences Association (SAMSA) Masamu Program—a program that aims to enhance research in the mathematical sciences by serving as a platform for US–Africa research collaborations. This work was also partially supported by the National Institute for Mathematical and Biological Synthesis (NIMBioS)—one of the several Mathematical Sciences Institutes sponsored by the NSF Division of Mathematical Sciences—through NSF Award DBI-1300426, with additional support from The University of Tennessee, Knoxville. The authors appreciate the support for travel expenses from the Society of Mathematical Biology. They acknowledge an invaluable conversation with Pam Small and Heather Williamson about transmission mechanisms.

References

1. Altizer, S., Ostfeld, R.S., Johnson, P.T., et al.: Climate change and infectious diseases: from evidence to a predictive framework. Science **341**(6145), 514–519 (2013)
2. Amofah, G., Bonsu, F., Tetteh, C., et al.: Buruli Ulcer in Ghana: results of a national case search. Emerg. Infect. Dis. **2**, 167–170 (2002)
3. Benbow, M.E., Williamson, H., Kimbirauskas, R., et al.: Aquatic invertebrates as unlikely vectors of Buruli ulcer disease. Emerg. Infect. Dis. **14**(8), 1247 (2008)
4. Benbow, M.E., Kimbirauskas, R., McIntosh, M.D., et al.: Aquatic macroinvertebrate assemblages of Ghana, West Africa: understanding the ecology of a neglected tropical disease. Ecohealth **11**(2), 168–183 (2014)
5. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B **57**(1), 289–300 (1995)
6. Blower, S.M., Dowlatabadi, H.: Sensitivity and uncertainty analysis of complex models of disease transmission: an HIV model, as an example. Int. Stat. Rev. **62**(2), 229–243 (1994)
7. Bonyah, E., Dontwi, I., Nyabadza, F.: A theoretical model for the transmission dynamics of the Buruli ulcer with saturated treatment. Comput. Math. Methods Med. **2014**, 576039 (2014)
8. Breban, R.: Role of environmental persistence in pathogen transmission: a mathematical modeling approach. J. Math. Biol. **66**, 535–546 (2013)
9. Cook, A.: The Mengo Hospital Notes. Makerere College Medical School Library, Kampala (1897)
10. De Silva, M.T., Portaels, F., Pedrosa, J.: Aquatic insects and mycobacterium ulcerans: an association relevant to Buruli ulcer control. PLoS Med. **4**, e63 (2007). https://doi.org/10.1371/journal.pmed.0040063
11. De Silva, K.R., Eda, S., Lenhart, S.: Modeling environmental transmission of MAP infection in dairy cows. Math. Biosci. Eng. **4**, 1001–1017 (2017)
12. Diekmann, O., Heesterbeek, J.P.: Mathematical Epidemiology of Infectious Diseases. Wiley, Chichester (2000)
13. Diekmann, O., Heesterbeek, H., Britton, T.: Mathematical Tools for Understanding Infectious Disease Dynamics. Princeton University Press, Princeton (2012)

14. Duker, A.A., Portaels, F., Hale, M.: Pathways of Mycobacterium ulcerans infection: a review. Environ. Int. **32**, 567–573 (2006)
15. Feller, E.C., Pearson, E.S.: Tests for rank correlation coefficients: II. Biometrika **48**, 29–40 (1961)
16. Garchitorena, A., Ngonghala, C.N., Guegan, J.F., et al.: Economic inequality caused by feedbacks between poverty and the dynamics of a rare tropical disease: the case of Buruli ulcer in sub-Saharan Africa. Proc. R. Soc. B **282**, 20151426 (2015)
17. Garchitorena, A., Ngonghala, C.N., Texier, G., et al.: Environmental transmission of Mycobacterium ulcerans drives dynamics of Buruli ulcer in endemic regions of Cameroon. Sci. Rep. **5**, 18055 (2015)
18. Ghana Health Service: https://www.ghanahealthservice.org/
19. Janssens, P.G., Quertinmont, M.J., Sieniawski, J., et al.: Necrotic tropical ulcers and Mycobacterial causative agents. Trop. Geogr. Med. **11**, 293–312 (1959)
20. Kelly Jr, M.R., Tien, J.H., Eisenberg, M.C., et al.: The impact of spatial arrangements on epidemic disease dynamics and intervention strategies. J. Biol. Dyn. **10**, 222–249 (2016)
21. Kenu, E., Nyarko, K.M., Seefeld, L., et al.: Risk factors for Buruli ulcer in Ghana–a case control study in the Suhum-Kraboa-Coaltar and Akuapem South Districts of the eastern region. PLoS Negl. Trop. Dis. **8**(11), e3279 (2014)
22. LaSalle, J.P.: The Stability of Dynamical Systems, vol. 25. SIAM, Philadelphia (1976)
23. MacCallum, P., Tolhurst, J.C.: A new Mycobacterial infection in man. J. Pathol. Bacteriol. **60**, 93–122 (1948)
24. Macklin, J.T.: An investigation of the properties of double radio sources using the Spearman partial rank correlation coefficient. Mon. Not. R. Astron. Soc. **199**, 1119–1136 (1982)
25. Marino, S., Hogue, I.B., Ray, C.J., et al.: A methodology for performing global uncertainty and sensitivity analysis in systems biology. J. Theor. Biol. **254**, 178–196 (2008)
26. Marion, E., Eyangoh, S., Yeramian, E., et al.: Seasonal and regional dynamics of M. ulcerans transmission in environmental context: deciphering the role of water bugs as hosts and vectors. PLoS Negl. Trop. Dis. **4**, e731 (2010)
27. Marsollier, L., Robert, R., Aubry, J., et al.: Aquatic insects as a vector for *Mycobacterium ulcerans*. Appl. Environ. Microbiol. **68**(9), 4623–4628 (2002)
28. Martcheva, M., Lenhart, S., Eda, S., et al.: An immuno-epidemiological model for Johne's disease in cattle. Vet. Res. **46**, 69 (2015)
29. McKay, M.D., Beckman, R.J., Conover, W.J.: A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics **21**, 239–245 (1979)
30. Merritt, R.W., Walker, E.D., Small, P.L., et al.: Ecology and transmission of Buruli ulcer disease: a systematic review. PLoS Negl. Trop. Dis. **4**, e911 (2016)
31. Miller Neilan, R.L., Schaefer, E., Gaff, H., et al.: Modeling optimal intervention strategies for cholera. Bull. Math. Biol. **72**, 2004–2018 (2010)
32. Morris, A., Gozlan, R.E., Hassani, H., et al.: Complex temporal climate signals drive the emergence of human water-borne disease. Emerg. Microbes Infect. **3**, e56 (2014)
33. Morris, A., Guégan, J.F., Benbow, M.E., et al.: Functional diversity as a new framework for understanding the ecology of an emerging generalist pathogen. EcoHealth **13**, 570–581 (2016)
34. Nyabadza, F., Bonyah, E.: On the transmission dynamics of Buruli ulcer in Ghana: insights through a mathematical model. BMC Res. Notes **8**, 656 (2015)
35. Pascual, M., Bouma, M.J., Dobson, A.P.: Cholera and climate: revisiting the quantitative evidence. Microbes Infect. **4**, 237–245 (2002)
36. Portaels, F., Elsen, P., Guimaraes-Peres, A., et al.: Insects in the transmission of Mycobacterium ulcerans infection. Lancet **353**, 986 (1999)
37. Röltgen, K., Pluschke, G.: Epidemiology and disease burden of Buruli ulcer: a review. Res. Rep. Trop. Med. **6**, 59–73 (2016)
38. Shuai, Z., van den Driessche, P.: Global stability of infectious disease models using Lyapunov functions. SIAM J. Appl. Math. **73**, 1513–1532 (2013)

39. Siewe, N., Yakubu, A.A., Satoskar, A.R., et al.: Immune response to infection by Leishmania: a mathematical model. Math. Biosci. **276**, 28–43 (2016)
40. Sopoh, G.E., Johnson, R.C., Chauty, A., et al.: Buruli ulcer surveillance, Benin, 2003–2005. Emerg. Infect. Dis. **9**, 1374–1376 (2007)
41. The World Bank: Ghana data. Technical report, The World Bank (2016). Retrieved from http://data.worldbank.org/country/ghana
42. Thomas, C.D., Cameron, A., Green, R.E., et al.: Extinction risk from climate change. Nature **427**, 145–148 (2004)
43. Tien, J.H., Earn, D.J.: Multiple transmission pathways and disease dynamics in a waterborne pathogen model. Bull. Math. Biol. **72**, 1506–1533 (2010)
44. van den Driessche, P., Watmough, J.: Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. Math. Biosci. **180**, 29–48 (2002)
45. van den Driessche, P., Watmough, J.: Mathematical Epidemiology: Further Notes on the Basic Reproduction Number. Springer, Berlin (2008)
46. van Ravensway, J., Benbow, M.E., Tsonis, A.A., et al.: Climate and landscape factors associated with Buruli ulcer incidence in Victoria, Australia. PLoS One **7**, e51074 (2012)
47. Wansbrough-Jones, M., Phillips, R.: Buruli ulcer: emerging from obscurity. Lancet **367**(9525), 1849–1858 (2006)
48. Williamson, H.R., Benbow, M.E., Nguyen, K.D., et al.: Distribution of *Mycobacterium ulcerans* in Buruli ulcer endemic and non-endemic aquatic sites in Ghana. PLoS Negl. Trop. Dis. **2** (2008)
49. Williamson, H.R., Benbow, M.E., Campbell, L.P., et al.: Detection of *Mycobacterium ulcerans* in the environment predicts prevalence of Buruli ulcer in Benin. PLoS Negl. Trop. Dis. **6**, e1506 (2012)
50. Williamson, H., Mosi, L., Donnell, R., et al.: Mycobacterium ulcerans fails to infect through skin abrasions in a guinea pig infection model: implications for transmission. PLoS Negl. Trop. Dis. **8**, e2770 (2014)
51. World Health Organization (WHO): Weekly epidemiological record. Technical report, World Health Organization (2002). http://www.who.int/wer/2002/en/wer7732.pdf
52. World Health Organization (WHO): Buruli ulcer (Mycobacterium ulcerans infection), Fact sheet. Technical report, World Health Organization (2016). http://www.who.int/mediacentre/factsheets/fs199/en/
53. Worldometers: Population (2016). Retrieved from http://www.worldometers.info/world-population/ghana-population/

Chapter 6
Data-Informed Modeling in the Health Sciences

Antonios Zagaris

Abstract The adoption of automation and technology by health professionals is triggering an explosion of databases and data streams in that sector. The emergence of this data torrent creates the pressing need to mine it for value, which in turn requires investment for the development of modeling and analysis tools. In view of this, dynamicists are presented with the terrific opportunity to enrich their discipline by supplying it with new tools, expanding its scope, and elevating its social impact. This chapter is written in that spirit, examining three concrete case studies encountered *in the field*: quantifying the salmonellosis risk posed by distinct food sources, assimilating genetic data into a dynamical model for avian influenza transmission, and statistically decontaminating gas chromatography/mass spectroscopy time series. We review available prototypical models and build on them guided by data and mathematical abstraction, demonstrating in the process how to root a model into data. This takes us quite naturally into the realm of probabilistic and statistical modeling and reopens a decades-old discussion on the role of discrete models in applied mathematics. We also touch briefly on the timely subject of mathematicians being employed as such outside math departments and attempt a short outlook on their prospects and opportunities.

Keywords Probabilistic and data-driven modeling · Parameter inference · Extramural mathematics · Infection source attribution · Mathematical epidemiology · Data decontamination · Bayesian hierarchical models

6.1 Introduction

May you live in interesting times. There is, perhaps, no better description of this age in mathematics than this apocryphal mixed blessing. Broadly seen, our discipline

A. Zagaris (✉)
Department of Bacteriology and Epidemiology, Wageningen Bioveterinary Research, Wageningen University and Research, Lelystad, The Netherlands
e-mail: antonios.zagaris@asml.com

H. G. Kaper, F. S. Roberts (eds.), *Mathematics of Planet Earth*, Mathematics of Planet Earth 5, https://doi.org/10.1007/978-3-030-22044-0_6

remains a cornerstone of civilization, its place in the pantheon of human intellect as secure as ever. Mathematics continues to push boundaries, inspire, and bewilder, and it will do so for as long as it can attract young talent. Applied mathematicians, however, are necessarily caught up in the frenzy of our times: *data*. Although the scientific enterprise—let alone daily life—revolves around data collection and analysis, mathematics does not: at best, it handles data and then again not very often. Should we strive to incorporate data into everyday mathematical practice? And what would that entail?

The situation may seem unprecedented, but a parallel can be drawn to the arrival of cheap computational power. On the one hand, the study of many analytical tools of the past was so tied to computational limitations that the removal of the latter caused many tools to fall by the wayside. On the other hand, an entire cocoon of analysis grew around computers to form the uniquely powerful tool known as scientific computing. (And if truth be told, old tools still nod from the wayside every so often, with a curious tenacity.) These developments reestablished the social relevance of mathematics and created vast opportunities for mathematical practitioners, attracting talent, rejuvenating academic programs, and enriching scholarship. Numerical mathematics today is such an integral part of mathematical reality that it is hard to imagine a time when it was not.

Importantly, the seminal character of that advent lies *not* in that it sped up calculation but that it *enabled* it, precipitating a steep increase in the complexity of problems amenable to modeling and analysis and a commensurate expansion into other fields. Technology adoption depends on the end user, however, and mathematicians embraced computers to transgress insurmountable analytic difficulties and not as a stratagem for expanding mathematical reign. This is a salient difference with the present, as data is *not* a mathematical tool. And yet, it unlocks the same doors: relevance, expansion, improved interdisciplinarity, and the promise of new mathematics, science, and technology. The sooner this truth is embraced, the more effectively that promise can be fulfilled.

This chapter aims to present real-world examples of modeling driven by specific data types, and it primarily targets junior dynamicists and mathematical analysts. The problems we treat here were encountered *in the field* by the scientists acknowledged in this chapter, and modeling was done in full collaborative mode and outside the confines of a mathematical faculty. The material is far more about answering questions *with* mathematics than *in* mathematics, reflecting the author's belief that modeling starts with a question and not with an expertise. This attitude is evident in many researchers who use mathematics as a probe, prioritize problem over tool, and appropriate theory on demand. The most extraordinary among them are active at both ends, resembling primary producers in the mathematical ecosystem, who keep generating mathematics by doing science. This chapter makes no such claim; instead, it highlights a particular mode of thinking and the substantial challenges a data-minded modeler is likely to encounter. By bringing such considerations to the fore, we hope to make *mathematics of planet earth* viable in the here and now, so that we can dream of mathematics on other planets in the near future.

Outline of the Chapter We start in Sect. 6.2 with a historical dataset of salmonellosis (a disease caused by ingested *Salmonella* bacteria), which lists human cases and data on food-related Salmonella sources. We develop a stochastic model to formulate and calibrate a source attribution problem: Which food source(s) contribute the most infections? In Sect. 6.3, we explore the problem of resolving avian influenza transmission and incorporating genetic evidence into the solution. Using a Bayesian scaffold, we differentiate individual infectors and incorporate genetic data in a natural manner. In Sect. 6.4, we focus on a large dataset of molecular count time series and ask the question: Why do standard models fail to describe these data? Using signal processing and statistical tools, we identify the culprit in the form of a molecular contaminant and decontaminate the samples algorithmically.

6.2 Source Attribution

Our first example comes from mathematical epidemiology, a classical field tackling questions on the spread of infectious diseases. Most dynamicists are familiar with SIR models [20], which describe the (spatio)temporal evolution of epidemics. Our task in this section is somewhat different, in that we are *not* asked to predict how a disease spreads but to assess the relative importance of distinct transmission pathways. We start with a historical dataset of salmonellosis, which lists human cases and data on food-related *Salmonella* sources. To assess the infectious potential of those pathways, we develop below a toy model to attribute *cases* to *sources*; the more cases a source is projected to cause, the more important the corresponding pathway. Problem definition is a sizable part of the problem and the main motivation for this section, which is meant as a springboard for exploring the chapter theme.

Source attribution has a celebrated history, dating back to John Snow [26], who famously traced a cholera outbreak to a public water pump; see [30] for a critical exposition. Since that prototypical scientific whodunit, attribution has relied both on lab and field work and on modeling—the former produces data, which the latter transforms to quantifiable, actionable insights. Source attribution is today also part of the major epidemiological theme of food safety, which informs regulatory frameworks, trade, and policy. For an extreme example of how that theme impacts society, we refer the interested reader to the recent outbreak of E. coli O104:H4 and its effect on trade and international relations [19]. Even in the absence of catastrophic events, attribution serves to identify major drivers of pathogen transmission in logistics chains and to improve public health by shaping meaningful intervention strategies. Within that field, modeling is perceived as an *exploratory analytical tool* that provides causative clues, identifies data gaps and intervention points and, importantly, *predicts* the efficacy of envisioned intervention.

6.2.1 Data and Modeling

We start with a few thoughts on the problem of assigning infection events to specific infectors. Infection plainly presupposes transmission, which is the main theme of the aforementioned SIR framework. In models of that variety, the modeler partitions populations meaningfully, for example, into susceptible and infected subpopulations, and models the dynamic interaction between them. These interactions include an element of chance (*stochasticity*), which however often averages out over large populations. Once *parameterized*, SIR models are used to make inferences on disease progression. Data play an important role in that enterprise, because epidemiological model parameters admit no universal values; instead, they depend strongly on the situation at hand and are *estimated* by data fitting. Because of this, *the nature of a model is largely dictated by the available data*. We will take heed of this during model formulation below.

Here, we specifically consider *historical* data from a setting commonly encountered in practice, in which one monitors human *Salmonella* infections caused by distinct food sources. Transmission occurs by ingestion, but further details—including the precise source of each infection—are not part of the dataset. Although dynamics are present in the data (each case occurs at a specific time and place), resolving it is not part of our task. Hence, the SIR framework is not directly applicable and we will start nearly from scratch.

Before we proceed, it is important to outline the role of modeling in such problems; for a more extensive, characteristically lucid view on the matter, we refer the reader to [6]. First, attribution is a *causal* problem at heart, yet statistical and mathematical considerations alone hardly allow causal inferences; see [23] for a detailed discussion. Field epidemiological methods, such as contact tracing, are much better suited to answer causal questions but may be hampered by their reliance on infrastructures and resources that are not universally present. Such methods are of no use here specifically, because such information is absent from our dataset—a common situation in historical data collections. In that sense, modeling work in the health sciences often acts as *surrogate* for more direct but unavailable assessment methods. When that is the case, it is exceedingly important that modelers steer their work to answer the overarching question and understand the limitations of modeling. In that spirit, the reader is invited to pinpoint substantial weaknesses in our work below.

6.2.1.1 Data

For this expository presentation, we consider the *Salmonella* dataset in Table 6.1, copied verbatim from [16]. Indeed, our discussion here owes much to [16, 17], and we direct the interested reader there for a starting point into this domain. The dataset identifies four distinct infection sources—pork, beef, and two kinds of poultry—and treats humans as a uniform population. A *case* is a confirmed, human *Salmonella* infection, with the total coming to 3880 for the duration of the study. Our task

Table 6.1 Data [16] used to formulate and calibrate the source attribution model

Types	Percentages (%)				
Serotype	Humans	Pork	Beef	Broiler flocks	Layer flocks
S. Enteritidis	67.2	0.0	0.0	24.5	85.4
S. Typhimurium	17.5	54.0	10.5	18.2	6.8
S. Hadar	1.6	0.0	0.0	3.7	0.0
S. Manhattan	1.1	0.0	0.0	3.0	0.0
S. Infantis	0.9	17.2	0.0	17.5	4.9
S. Virchow	0.8	0.2	0.0	0.4	0.0
S. Agona	0.8	0.0	0.0	1.1	0.0
S. Derby	0.6	6.3	10.5	1.1	0.0
S. Newport	0.6	0.0	0.0	0.0	0.0
S. Java	0.5	0.0	0.0	0.0	0.0
S. Stanley	0.5	0.0	0.0	0.0	0.0
S. Braenderup	0.4	0.0	0.0	0.0	0.0
S. Bovismorbificans	0.4	0.0	0.0	0.0	0.0
S. Glostrup	0.3	0.0	0.0	0.0	0.0
S. Heidelberg	0.3	0.0	0.0	0.0	0.0
S. Saintpaul	0.3	0.0	0.0	0.0	0.0
S. Dublin	0.3	0.2	68.5	0.0	0.0
Others incl. nontypable	6.1	21.9	10.5	30.5	2.9
	Number typed				
	3880	448	19	269	103

is to assess how sources contribute to those cases—that is, to estimate the *total* number of cases caused by each source and *not* to assign individual cases to sources. Since all sources harbor the same pathogen, no sensible model can do so without additional factors *differentiating* sources. Our saving grace here is that pathogens assume multiple forms (*types*) differing in specifics, for example, in surface antigens (*serotypes*). These types can be differentiated in the lab, and nearly all cases in the table are of a specific type. Types cannot identify a source directly, as they are found in multiple sources (cf. *S. Typhimurium* found in all sources). However, distinct sources harbor types in different relative abundances; below, we develop a simple attribution scheme that uses these *type distributions* as source signatures.

To introduce notation, we consider an arbitrary number M of infection sources harboring N pathogen types and causing L infections in a single human compartment over a certain time period T. In the context of Table 6.1, $L = 3880$, $M = 4$, and $N = 17$. We write $\mathscr{C} = \{c_1, \ldots, c_L\}$ for the cases, $\mathscr{S} = \{s_1, \ldots, s_M\}$ for the pathogen sources, and $\mathscr{T} = \{t_1, \ldots, t_N\}$ for the pathogen types. We denote arbitrary cases, sources, and types by c_ℓ, s_m, and t_n or, where no confusion can arise, simply by ℓ, m, and n. We assume that cases can be *typed* unambiguously through some typing function mapping cases to pathogen types. In the context of Table 6.1, this means that we discard untyped cases (but see Sect. 6.2.3). Finally, we assume that the typing distribution of each source m is known, through lab

analysis of *infected* food samples, and represented by $\gamma_{1|m}, \ldots, \gamma_{N|m}$; in the context of Table 6.1, these numbers comprise the column of the mth source. These values represent relative type abundances in source m, are nonnegative, and sum to one; we write this compactly as $\mathbb{1}^{\mathrm{T}}\Gamma = \mathbb{1}^{\mathrm{T}}$, with $\mathbb{1}^{\mathrm{T}} = (1, \ldots, 1)$ and Γ the $N \times M$ matrix having entries $\gamma_{n|m}$. Note that Γ carries *no* information on pathogen *prevalence*, as it only quantifies *positive* (i.e., infected) food samples; this is consistent with the given data.

The approach we take below is to set up an infection model, similar to SIR infection modules; parameterize it, using the data in Table 6.1; and finally infer the infectious potential of the sources from the parameterized model.

6.2.1.2 Infection Model

The essential functionality of an infection module is to summarize the transmission mechanism of the pathogen. To illustrate the complexity of this task for our problem, we note that the path mediating food consumption includes factors that can hardly be modeled—let alone parameterized—with any degree of certainty; for example, farm, slaughter house, and selling point conditions; transport; cooking practices; and many more. The situation is further exacerbated by unknown transmission rates, the impossibility of performing controlled experiments, and the fact that consumers are exposed differentially to food sources. Given the scarcity of detail in Table 6.1, resolving transmission pathways at that level is out of the question; we resort, instead, to a highly abstracted model.

To formulate a minimal model of pathogen transmission that operates at the level of our data, we treat humans and sources as homogeneous compartments lacking detail. Specifically, we assume each source m to cause an infection of type n in the human compartment at some constant but *unspecified* rate $\lambda_{n|m}$. In other words, we model infection as a standard memoryless stochastic process, see also Fig. 6.1. Stochasticity is non-essential but mathematically palatable, as we can easily show that a number of processes derived from these simple building blocks are also memoryless. In particular, m causes an infection (of any type) at rate $\mu_m = \sum_{n=1}^{N} \lambda_{n|m}$, whereas infections of type n (caused by any source) occur at rate $\nu_n = \sum_{m=1}^{M} \lambda_{n|m}$; infections in general occur at rate $\sum_{m=1}^{M} \sum_{n=1}^{N} \lambda_{n|m}$. Defining the $N \times M$ matrix Λ and column vectors μ, ν in the obvious way, we have $\nu = \Lambda\mathbb{1}$ and $\mu = \Lambda^{\mathrm{T}}\mathbb{1}$—that is, ν and μ are the row and column sums of Λ. This simple model forms the basis of our attribution scheme.

6.2.1.3 Attribution Scheme

The memoryless model above describes the infection process but does not address our original problem. To attribute infections, we fix an arbitrary time τ and define I, a Boolean *random variable* (r.v.) that determines whether an infection occurred or not in that time ($I = 1$ or 0). If it did, then the probability that it was caused

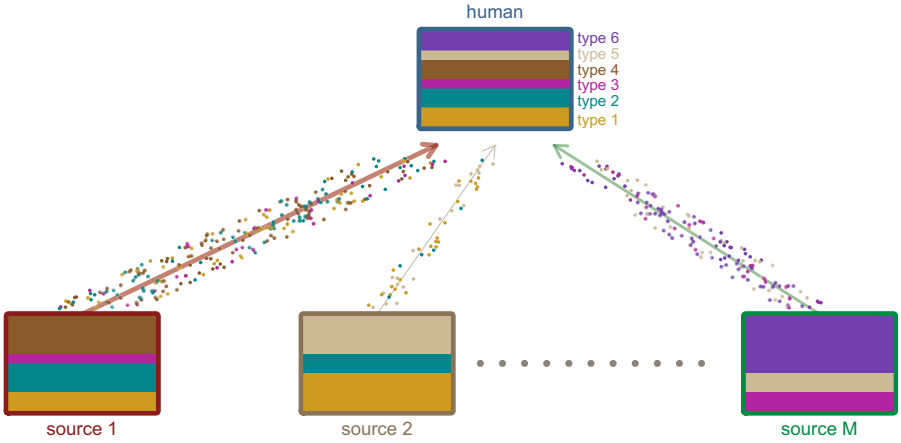


Fig. 6.1 Schematic representation of the model discussed in Sect. 6.2.1. Each source (bottom) harbors pathogen types in different proportions and causes human infections at different rates. The contribution of sources to the human compartment (top) yields the stable distribution of types in it that is witnessed in the data

by source m and was of type n is given by the *conditional probability mass function* (pmf)

$$f_{S,T|I}(m,n|1) = \lim_{\tau\downarrow 0} \frac{f_{S,T,I}(m,n,1)}{f_I(1)}.$$

Here, the limit $\tau \downarrow 0$ eliminates the probability of multiple infections. In this equation, $f_{S,T,I} : \mathscr{S} \times \mathscr{T} \times \{0,1\} \to \mathbb{R}_+$ is a pmf determining the probability $f_{S,T,I}(m,n,1)$ that source m caused a type-n infection in the fixed time τ. The *marginal* f_I specifies the probability that an infection occurred at all (also in time τ) and the overall equation defines the conditional. Since event times for memoryless processes are exponentially distributed, we have

$$f_{S,T|I}(m,n|1) = \lim_{\tau\downarrow 0} \frac{1-\mathrm{e}^{-\tau\lambda_{n|m}}}{1-\mathrm{e}^{-\tau\sum_{m,n}\lambda_{n|m}}} = \frac{\lambda_{n|m}}{\sum_{m',n'}\lambda_{n'|m'}}. \tag{6.2.1}$$

This equation represents the probability that an infection has a certain type and is caused by a certain source; the denominator is a normalization constant. Note again how the stochastic framework led directly to this first result.

For our attribution scheme, we are interested in quantifying the probability that an infection is caused by a specific source *no matter what its type*. We can compute that probability by marginalizing Eq. (6.2.1),

$$f_{S|I}(m|1) = \sum_n f_{S,T|I}(m,n|1) = \frac{\sum_n \lambda_{n|m}}{\sum_{m',n'}\lambda_{n'|m'}} = \frac{\mu_m}{\sum_{m'}\mu_{m'}}. \tag{6.2.2}$$

Thus, given L cases, our scheme assigns on average L_m infections to source m, where

$$L_m = f_{S|1}(m|1)L = \frac{\mu_m}{\sum_{m'} \mu_{m'}} L. \tag{6.2.3}$$

Note that we assign infection totals and not individual cases; this is consistent with the lack of individuality in the dataset informing our model.

To employ scheme (6.2.3), we must estimate μ-values by fitting them to the available data. Parameter estimation is a field of its own, and we shall not delve deep into it here; a conceptual primer is included in the Appendix at the end of this chapter. For use below, we note here various conditionals and marginals of Eq. (6.2.1). Writing all pmfs as column vectors and omitting notation signifying that all probabilities are conditioned on an infection occurring, we have the compact forms

$$f_T = \frac{\nu}{\mathbb{1}^{\mathrm{T}}\nu}, \quad f_S = \frac{\mu}{\mathbb{1}^{\mathrm{T}}\mu}, \quad f_{T|S} = \Lambda\,\mathrm{diag}(\mu)^{-1}, \quad f_{S|T} = \mathrm{diag}(\nu)^{-1}\Lambda. \tag{6.2.4}$$

These are, respectively, the probabilities that an infection was of a specific type, caused by a specific source, of a specific type given the source, and caused by a specific source given the type. The simplicity of these relations can be traced back to that of *Bayes' law*—another advantage of using a probabilistic approach.

6.2.1.4 Parameter Inference

Parameterizing scheme (6.2.3) means estimating the column sums μ of Λ, which in turn entails some sort of *optimization*. For example, *maximum likelihood estimation* (MLE; see Appendix) returns parameter values that maximize the probability of observing the given data (*likelihood*). Optimization in high-dimensional spaces is problem-prone because of local minima, wide confidence intervals (*sloppiness* [15]), and more. Here, we only work in $M = 4$ dimensions, yet estimation shall turn out to be entirely *impossible* in that the model parameters are *non-identifiable* [24].

To see this play out in practice, we build and maximize the data likelihood for our model. For memoryless processes, the number of events occurring in a fixed time T is Poisson-distributed. Assuming, also, that the infection processes of distinct types are *independent* and writing L_n for the number of type-n cases in the data, we find the likelihood

$$\mathscr{L}(\Lambda|L_1,\ldots,L_N) = \mathrm{prob}(L_1,\ldots,L_N|\Lambda) = \prod_{n=1}^{N} \frac{(T\nu_n)^{L_n}\mathrm{e}^{-T\nu_n}}{L_n!}. \tag{6.2.5}$$

The Poisson components in Eq. (6.2.5) are evaluated at parameter values Λ, making the likelihood $\mathscr{L}$ a function of them. According to MLE, maximizing $\mathscr{L}$ yields

the *optimal* parameter values in view of the existing data; for other approaches to parameter inference, see, for example, [4]. However, Eq. (6.2.5) involves type-specific infection rates ν (row sums) and *not* source-specific rates μ (column sums); this could have been anticipated, as *our data differentiates between types but aggregates sources*. As a consequence, parameter estimates can only be obtained at that aggregation level. Indeed, maximizing the log-likelihood,

$$\ln \mathscr{L} = \sum_{n=1}^{N} L_n \ln \nu_n - T \sum_n \nu_n - \sum_{n=1}^{N} L_n! + L \ln T, \tag{6.2.6}$$

yields $\hat{\nu}_n = L_n/T$. Unfortunately, these empirical type rate estimates yield little information on μ-entries. To proceed further, we need to add detail to our model.

6.2.2 Additional Modeling

To estimate the parameters in our attribution scheme, we need to express our likelihood in terms of μ. Since ν and μ are row and column sums of the same matrix, there is no one-to-one relation between them generally speaking. Instead, we need additional modeling assumptions that go beyond memorylessness and well into specifics. For our exposition here, we describe a simple approach which, realistically, should be replaced by input from *domain experts* involved in the study. The detail we add to the model is that sources do not "emit" pathogen types at whichever rates but, instead, at rates proportional to their prevalence in those sources. Since the type proportions are $\gamma_{1|m}, \ldots, \gamma_{N|m}$, our assumption is expressed mathematically as $\lambda_{n|m} = \gamma_{n|m}\, \mu_m$. In words, the potential of a source m to cause nth type infections is proportional to the relative abundance of that type in the source. In matrix format,

$$\Lambda = \Gamma \operatorname{diag}(\mu), \quad \nu = \Gamma \mu. \tag{6.2.7}$$

This leaves the attribution scheme (6.2.3) unaltered; its sole purpose is to enable estimation of μ by tying it to that of ν, with the latter having been carried out in Eq. (6.2.6). The matrix Γ relating the two can be read off the data, cf. Sect. 6.2.1.

6.2.2.1 Inferring the New Parameters

Returning to the parameter inference problem, we use Eq. (6.2.7) and the chain rule $\nabla_\mu \mathscr{L} = \nabla_\nu \mathscr{L}\, \nabla_\mu \nu$ to obtain the MLE condition

$$0 = \nabla_\mu \ln L = \left(\frac{L_1}{\nu_1} - T, \ldots, \frac{L_N}{\nu_N} - T \right) \Gamma.$$

Using the identity $\mathbb{1}^{\mathrm{T}}\Gamma = \mathbb{1}^{\mathrm{T}}$ and transposing, we arrive at

$$\Gamma^{\mathrm{T}} \begin{pmatrix} L_1/\nu_1 \\ \vdots \\ L_N/\nu_N \end{pmatrix} = T\,\mathbb{1} \quad \text{subject to } \nu = \Gamma\mu. \tag{6.2.8}$$

This system is comprised of M equations, with T acting as a scaling parameter that is immaterial for attribution. If $M > N$ (more sources than types), then solution (6.2.6) applies generically, but the system has no unique solution in terms of μ. In the more interesting case $M < N$, which also corresponds to Table 6.1, solution (6.2.6) does not apply in general. In that case, the system is polynomial in μ and can be solved numerically; see [14] for a demonstration and below for the question of existence and uniqueness of solutions.

6.2.2.2 A Reinterpretation of the Model

The stochastic model set up above assumed a specific infection *mechanism* that relied on memoryless processes. Another way to look at it is through the lens of a probabilistic identity which it obeys by virtue of Eq. (6.2.4),

$$f_T(n) = \sum_{m=1}^{M} f_S(m) f_{T|S}(n|m). \tag{6.2.9}$$

Here again, we omitted notation pertaining to an infection occurring. The left member of this equation (f_T) represents the typing pmf in the human compartment—that is, the first column of Table 6.1; the conditional pmfs in the right member ($f_{T|S}$) represent the typing distributions of the different sources—that is, the remaining four table columns; and the entries of f_S are model parameters subject to inference, as they are proportional to μ by Eq. (6.2.4). In other words, data fitting boils down to approximating a given pmf by a *linear combination* of $M = 4$ given pmfs (the four rows of Γ). Such models are known in the statistical literature as *mixture models*; see [3] and the references given there for an overview.

The interpretation of (6.2.9) as a mixture model offers an abstract view into the associated existence and uniqueness problem. Specifically, any pmf lies in the unit $(N-1)$-simplex $\mathbb{D} = \{f \mid \mathbb{1}^{\mathrm{T}} f = 1\} \cap \overline{\mathbb{R}^N_+}$; hence, linear combinations of $f_{T|S}(\cdot|1), \ldots, f_{T|S}(\cdot|M)$ define an $(M-1)$-simplex $\mathbb{D}' \subset \mathbb{D}$. If $M < N$, then $\mathbb{D}'$ is of lower dimension than $\mathbb{D}$; generically, then, $f_T \in \mathbb{D}\backslash\mathbb{D}'$ and thus the vector f_T cannot be expressed as a linear combination of the vectors $f_{T|S}(\cdot|1), \ldots, f_{T|S}(\cdot|M)$. Therefore, data fitting becomes the *approximation* problem of finding the point(s) in $\mathbb{D}'$ that are closest to f_T. Since $\mathbb{D}'$ is a compact and convex set, any strictly convex notion of closeness leads to a unique minimizer in $\mathbb{D}'$—that is, to a unique optimal solution of Eq. (6.2.9).

It is instructive to note here that the added value of this short section is its first paragraph, as it translates our original (mechanistic) model in statistical terms which are more accessible to domain experts. The paragraph above, on the other hand, may be mathematically stimulating but unlikely to come up in an applied discussion; in fact, the author would argue it is an overkill for a tentative toy model, as a numerical investigation would suffice. And yet, it does have value *for the modeler* for two reasons: first, it provides a *mental* model in terms of known quantities (vectors, subspaces, distances); and second, it introduces one to the important problem of quantifying *closeness* between probability distributions. Indeed, it is not self-evident how to define a meaningful distance between pmfs or even whether a *distance*, with its significant overhead, is the right notion. We worked above with likelihoods and shall revisit the problem in Sect. 6.4.4, but there is nothing exhaustive about our treatment. To the contrary, this question has generated significant literature that presents an excellent opportunity to expand one's mathematical knowledge; see, for example, the classic reference [22]. This lends strong advocacy to applied work, as noted much earlier: "*The researcher's purely mathematical ingenuity is likely to be exercised more, not less, by the fact of his dealing with genuine problems*" [6].

6.2.2.3 Self- and Multidirectional Attribution

Although this model is a fair starting point for pathogens such as *Salmonella*, there are situations where infection spreads also within the general population or from it to other compartments. A particularly relevant example is *antimicrobial resistance*, a property of bacteria to acquire and transfer genetic material encoding antibiotic-fighting mechanisms. These resistance genes can be highly mobile, both vertically (across bacterial generations) and horizontally between bacteria of the same or other species; indeed, monitoring programs have produced evidence of transmission between human and animal reservoirs [12].

High mobility suggests a *network* of interactions, in which resistance genes travel both within and across compartments. An additional layer of complexity is that these genes are found in multiple bacterial species, making aggregation possible at many levels. To apply directly the framework above, one can introduce an additional r.v. R for the *receiver* compartment of the infection, work with an $M \times M \times N$ array Λ holding infection rates $\lambda_{r,s,t}$ of type-t transmission from source s to receiver r, compute the pmf $f_{R,S,T}$, and set up the attribution scheme $f_{S|R}$. In doing that, the data type remains the same: resistance gene counts, possibly supplemented by bacterial types. The model is parameterized by the $M \times M$ matrix μ of transmission rates $f_{S|R}$ within and between compartments, which increases dramatically the number of model parameters (curse of dimensionality). However, the thorny issue here is less the sharp increase in the number of rates and more that attribution becomes *impossible*, as any number of infections within a compartment can be attributed intracompartmentally. This is easiest seen in the context of Eq. (6.2.9), which now becomes

$$f_{T|R}(n|m) = \sum_{m'=1}^{M} f_{S|R}(m'|m) f_{T|S,R}(n|m', m).$$

Making the normative assumption that each source emits types in identical proportions to all receivers, we can rewrite this equation as

$$f_{T|R}(n|m) = \sum_{m'=1}^{M} f_{S|R}(m'|m) f_{T|S}(n|m').$$

Assuming now more strongly that that proportion corresponds to the relative abundance of types in the source, as in Eq. (6.2.7), we can finally write

$$f_{T|R}(n|m) = \sum_{m'=1}^{M} f_{S|R}(m'|m) f_{T|R}(n|m').$$

Here, $f_{T|R}(\cdot|m)$ encodes type prevalences for an arbitrary compartment m and $f_{S|R}(m'|m)$ the strength of the connection from it to compartment m'. Interpreting all terms in this equation as matrices, we obtain the matrix equation $f_{T|R} = f_{T|R} f_{S|R}$, with $f_{S|R}$ unknown and $f_{T|R}$ encoding the data. This system has the solution $f_{S|R} = \mathrm{diag}(\mathbb{1})$, which is exact but unfortunately trivial: each compartment emits only to itself, thus generating its own distribution perfectly. We defer the problem of formulating an attribution scheme for network transmission that can resolve self-attribution to the interested reader, noting that it is of substantial interest in the applied community studying antimicrobial resistance.

6.2.3 Discussion

In this section, we developed an infection attribution scheme tailored to a specific data collection. Our main motivation was to demonstrate the *process* of building a data-informed model, which in effect consisted of mixing abstraction, pragmatism, and a pinch of domain expertise. The model itself was underpinned by a mechanistic, intuitive premise, but it also turned out to be interpretable in the language of choice in source attribution literature (statistical modeling). Such conceptual clarity is particularly compelling, as it enables one to assess model limitations and applicability; the link to statistics is also highly desirable, as it enables communication between different communities.

Of these three ingredients, abstraction is the one most accessible to mathematicians and pragmatism the easiest to acquire by practicing mathematics "in the field." The one trait modelers are not expected to emulate is domain knowledge: that presupposes sustained communication with domain experts, ideally in their natural habitat. Modelers undertaking that effort often encounter suspicion, and

complaints on the inapplicability (or sheer impenetrability) of applied mathematical work often ring true; there is no shortage of studies where unrealistic assumptions meet mathematical overkill to answer irrelevant questions. However, it is important to remember that this is a double-edged sword: domain experts will always lack mathematical refinement, yet they are often reluctant to relinquish creative control and likely to overestimate the power of modeling. Where mathematicians promise too much, because they fall in love with their models, practitioners expect too much because they do the same with their data. Famously, "*all models are wrong*" [6] but, also, all data is bad; and yet, the two must successfully fuse into science. The completion of this seemingly impossible task relies on involving modelers early on in study, preferably during study design already. Doing this enables them to share in the common goal, understand its complexity, and contribute to experimental design. At the same time, it helps experts understand what modeling can or cannot do and plan their study accordingly. To not do this is to run an oft-quoted danger: "*To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of*" [13].

Reasonable as the above may seem, it often fails to happen—all the more so because of data reuse. Our work above is an extreme example, since our model was built to accommodate immutable historical data. In such cases, and as trite as it may sound, modelers must remember to construct a model around what they have and not what they wish they did—if for no other reason, then merely to avoid "*the choice of a [...] model being determined by the researcher's background, the tradition prevalent within a discipline, or because the modeller is unaware of or unfamiliar with other modelling techniques*" [2]. Experts might be able to glean additional information channels, which again speaks for establishing communication early on, but data nature and quality are likely to pose strong methodological constraints anyhow. Such constraints may present an appreciable challenge to budding dynamicists trained primarily to use differential equations (DEs), but they also represent veritable opportunities to work with new concepts and expand mathematical knowledge.

Finally, it is important to emphasize here that our modeling work was not completed with (6.2.3) nor with (6.2.9). At the very least, *every* proposed model must be validated before being deployed and, ideally, supplemented with a sensitivity analysis for random/systematic model error and a strategy to account for missing data. Model validation is by far the most important of these yet conspicuously absent from much of mathematical literature and virtually all mathematics curricula; that is one important aspect in which applied mathematics needs to catch up with faster-moving disciplines, such as data science or machine learning. Sensitivity analysis quantifies the effect that uncertainty in model input (data) has on model output (here, μ) and thus also the reliability of our estimates. In our case, such uncertainty may arise from the moderate size of the data collection in Table 6.1 or from typing errors; see [14] for a demonstration of their impact. Systematic biases can be generally harder to identify, but in this section there is at least one obvious source of possible bias: the origin of *Salmonella* bacteria. Indeed, we took for granted the assertion that all cases are due to one of the four given sources, despite

lack of factual substantiation. (The existence of "other" types in the data may very well indicate undocumented infection sources.) This was our primary motivation in calling this a *toy model*, with the lack of a concrete strategy for dealing with missing data ("other" types, asymptomatic transmission, and more) a close second. These modeling aspects are out of scope here but of the utmost importance in real life.

6.3 Genetic Evidence and Epidemic Models

We saw above how data can be used to parameterize a mechanistic infection model and the source attribution scheme derived from it. The second example we treat also comes from the field of mathematical epidemiology but has a modern twist. We specifically consider highly pathogenic avian influenza (HPAI), a communicable veterinary disease that infects poultry routinely and disastrously. The 2003 HPAI epidemic in The Netherlands left in its wake approximately 30 million dead birds, 20,000 affected flocks, 89 infected people including one fatality, and economic losses in the billions of Euros [28]. This prompted the establishment of a preparedness network providing accurate diagnostics within hours of a warning signal and assessing the risk posed by disease reintroductions. The action for infected farms is largely prescribed [9] and entails transport restrictions, as well as depopulation of infected (and conditionally also of neighboring) flocks. Such control measures can be devastating for farmers and animals alike, so proper risk assessment is of the utmost importance. This is where *model-based decision support* comes into play.

In light of the above, a model could conceivably help quantify the risk that an HPAI reintroduction will *spread* under various control scenarios. That is indeed one of the tasks set for the epidemiological team at Wageningen Bioveterinary Research (WBVR) in The Netherlands. The data that can be accessed are both current and historic, with the latter pertaining to the 2003 epidemic and smaller outbreaks. These multimodal collections specifically encompass, on the individual farm level: locations and populations, dates of diagnosis and depopulation, genetic sequences of virus isolates, transport events between farms, and possibly other, more or less structured information such as weather data or field notes.

6.3.1 A Modeling Scaffold

Since spatiotemporal variability and disease transmission are both present, the problem can and has been interpreted as being of the SIR variety. A dynamicist attempting an off-the-shelf implementation of an SIR model, however, would encounter significant technical incongruences: time is not represented continuously but quantized; farms stay put instead of mixing homogeneously; farm densities are meaningless, since action is undertaken at the single-farm level; and so on. These considerations highlight the *inherent discreteness* of our setup and precipitate the

realization that—just as in the previous section—DEs are altogether inappropriate here and other tools are required. We must, however, acknowledge that the methods employed in these two sections *do* share something with DEs, namely the notion of *mechanism*. In the next section, we will foray into statistical modeling and that connection will be all but lost. The work in Sects. 6.3.2 and 6.3.3 is an abridged re-interpretation of [5], with a few modifications that simplify our presentation. The genetic component in Sect. 6.3.4 is new.

The *independent* variable in our setting is time t and takes values in the timeframe $\mathscr{T} = \{0, 1, \ldots, T\}$ (in days). This is the grid in which historic data are reported, with $t = 0$ marking the beginning of an outbreak and $t = T$ its end. Although spatial information exists and is significant, space is *not* considered an independent variable because farm locations are *fixed*. The *dependent* variable is the *system state*—that is, the totality of individual states of all farms in the ensemble $\{0, \ldots, K\}$. For our purposes here, we assume that the kth farm can be *unambiguously* declared to be in state $X_k^t \in \mathscr{X} = \{S, I, R\}$, at any time $t \in \mathscr{T}$. The vector $X^t = (X_0^t, \ldots, X_K^t)$ is the system state at that time and takes values in the *system state space* $\mathscr{X}^{K+1}$. This mimics classic SIR models but also abstracts the problem to its limits as, in reality, the farm–virus relation is much more nuanced. For example, is a farm exposed if virus is present in animals or merely in particulate matter? In the former case, how many exposed animals and which viral loads are required to call the farm exposed? What are the effects of a delayed diagnosis? Do farms enter the R-state upon culling or does infectiousness wane gradually? These and other questions are part and parcel of applied work, often debated upon at length in expert circles; the ability to navigate such ambiguity is a key success factor in applicable work.

The model we develop below is considerably more involved than that in the last section. To assist the presentation, we supplement it with a prototypical three-farm system ($K = 2$) wherever that helps illustrate model development.

6.3.2 State Evolution

To obtain a simple model, we must dress the scaffold above with *dynamics*—that is, prescribe laws for the state evolution function $X : \mathscr{T} \to \mathscr{X}^{K+1}$. In our discrete setting, this means specifying rules for state transitions $X^t \mapsto X^{t+1}$ and requires detailed knowledge of infection mechanisms, disease progression, and regulatory affairs. By the nature of the epidemic we model, each farm can only transition according to the linear chain $S \mapsto I \mapsto R$; this rules out the majority of the possible (system) state transitions. The rules for the permissible transitions are constructed from *single-farm* transition rules, by assuming that changes in daily farm states occur *independently*. In mathematical language,

$$f_{X^{t+1}|X^t}\left(x^{t+1}\middle|\, x^t\right) = \prod_{k=0}^{K} f_{X_k^{t+1}|X^t}\left(x_k^{t+1}\middle|\, x^t\right). \tag{6.3.1}$$

Here, the left member is the probability distribution for the state of the kth farm at time $t+1$, conditioned on the *system* state at time t; this reduces our problem to specifying distributions $f_{X_k^{t+1}|X^t}(x_k^{t+1}|x^t)$, for $x_k^{t+1} \in \mathscr{X}$ and $x^t \in \mathscr{X}^{K+1}$. In terms of our prototypical three-farm system, Eq. (6.3.1) becomes

$$
\begin{aligned}
&f_{X^{t+1}|X^t}\left(x_0^{t+1}, x_1^{t+1}, x_2^{t+1} \middle| x_0^t, x_1^t, x_2^t\right) \\
&= f_{X_0^{t+1}|X^t}\left(x_0^{t+1} \middle| x_0^t, x_1^t, x_2^t\right) f_{X_1^{t+1}|X^t}\left(x_1^{t+1} \middle| x_0^t, x_1^t, x_2^t\right) f_{X_2^{t+1}|X^t}\left(x_2^{t+1} \middle| x_0^t, x_1^t, x_2^t\right).
\end{aligned}
$$

Note that the probability of observing a farm state on day $t+1$ only depends on (the totality of) farm states the day before.

The introduction of probabilities here is reminiscent of our approach in Sect. 6.2.1 and warranted by our *de facto* inability to predict virus ($S \mapsto I$) transitions. Historical data additionally show considerable variability of susceptibility periods between similar farms, a fact that is evocative of "chance factors." By such factors we understand here transmission determinants that we do not explicitly model, such as farm hygiene, weather and environment, traffic, farm visits, and numerous others. The existence of such factors in the problem also points to probabilistic modeling.

6.3.2.1 Communal and Noncommunal State Transitions

To model the conditional distributions in Eq. (6.3.1), we further impose the rule

$$
f_{X_k^{t+1}|X^t}(x_k^{t+1}|x^t) = f_{X_k^{t+1}|X_k^t}(x_k^{t+1}|x_k^t), \quad x_k^t \neq S.
$$

In plain terms, we assume that the transitions of I- and R-farms depend *solely* on their own state. Although evidently false for susceptible farms (infection is a communal phenomenon), this is a viable and useful simplification for all other categories. For example, for farm $k=0$ in our three-farm system, this rule reads

$$
\begin{aligned}
f_{X_0^{t+1}|X^t}\left(x_0^{t+1} \middle| I, x_1^t, x_2^t\right) &= f_{X_0^{t+1}|X_0^t}\left(x_0^{t+1} \middle| I\right), \\
f_{X_0^{t+1}|X^t}\left(x_0^{t+1} \middle| R, x_1^t, x_2^t\right) &= f_{X_0^{t+1}|X_0^t}\left(x_0^{t+1} \middle| R\right).
\end{aligned}
$$

All distributions here are Bernoulli (i.e., binary) by virtue of chain linearity, as farms either retain their state or switch to the next one in the chain. For example, removed farms remain depopulated for the duration of the epidemic,

$$
f_{X_k^{t+1}|X_k^t}(S|R) = f_{X_k^{t+1}|X_k^t}(I|R) = 0, \quad f_{X_k^{t+1}|X_k^t}(R|R) = 1. \tag{6.3.2}
$$

Similarly, an infected farm can only remain infected or enter the R-state,

$$f_{X_k^{t+1}|X_k^t}(S|I) = 0, \quad f_{X_k^{t+1}|X_k^t}(I|I) = 1 - f_{X_k^{t+1}|X_k^t}(R|I). \tag{6.3.3}$$

The probability of the $I \mapsto R$ transition appearing here needs additional modeling, as it encodes disease progression and policy response. The same is true of $S \mapsto I$, which amounts to our transmission model. Before we delve into that, we note the probability of a specific transition in our three-farm illustration,

$$\begin{aligned} &f_{X^{t+1}|X^t}(R, I, I|I, I, S) \\ &\quad = f_{X_0^{t+1}|X^t}(R|I, I, S)\ f_{X_1^{t+1}|X^t}(I|I, I, S)\ f_{X_2^{t+1}|X^t}(I|I, I, S) \\ &\quad = f_{X_0^{t+1}|X_0^t}(R|I)\ f_{X_1^{t+1}|X_1^t}(I|I)\ f_{X_2^{t+1}|X^t}(I|I, I, S)\,. \end{aligned} \tag{6.3.4}$$

6.3.2.2 Transitions $S \mapsto I$

The step most relevant to risk assessment is infection spread, $S \mapsto I$. Above, we obliterated extraneous factors by assuming that the infection process is only informed by the overall system state,

$$f_{X_k^{t+1}|X^t}(I|x) = p_{kt}, \quad f_{X_k^{t+1}|X^t}(S|x) = 1 - p_{kt}, \quad x_k^t = S. \tag{6.3.5}$$

To model p_{kt}, we mechanistically envision susceptible farms being *independently challenged* by each farm in the *infectious* set $\mathscr{I}_t = \{k' \,|\, X_{k'}^t = \mathrm{I}\}$. Letting $p_{kk'}$ denote the probability that farm $k' \in \mathscr{I}_t$ infects farm k in the interval $(t-1, t)$, we can then write

$$1 - p_{kt} = \prod_{k' \in \mathscr{I}_t} (1 - p_{kk'}) = \mathrm{e}^{-\mathrm{FOI}_{\mathrm{kt}}}, \quad \mathrm{FOI}_{\mathrm{kt}} = -\sum_{k' \in \mathscr{I}_t} \ln(1 - p_{kk'}) > 0. \tag{6.3.6}$$

The quantity $\mathrm{FOI}_{\mathrm{kt}}$ is the (time-integrated/cumulative) *force of infection* exerted on the kth farm in the time interval $(t-1, t)$.

In the context of our three-farm model, and at the time t that we used in Eq. (6.3.4), we have a sole susceptible farm ($k = 2$) and two infectious ones ($k = 0, 1$). As a result, the infectious set is $\mathscr{I}_t = \{0, 1\}$ and the FOI on $k = 2$ becomes $\mathrm{FOI}_{2\mathrm{t}} = -\ln[(1 - p_{20})(1 - p_{21})]$. By Eqs. (6.3.5) and (6.3.6), then, the probability in Eq. (6.3.4) of that farm succumbing within the day is

$$f_{X_2^{t+1}|X^t}(I|I, I, S) = 1 - (1 - p_{20})(1 - p_{21}). \tag{6.3.7}$$

Infection spread is the only process in the model that couples farms; time only enters through the infectious set $\mathscr{I}_t$. For a discussion of our modeling assumptions, see Sect. 6.3.5. Note also that the infection module in [5] differs in specifics.

6.3.2.3 Transitions $I \mapsto R$

The switch $I \mapsto R$ is modeled via the transition time $T^{R|I}$ and usually has memory; this is evident in that $T^{R|I}$ follows a localized, instead of an exponential, distribution. To regain the computationally and analytically palatable memorylessness, one often represents the transition through a chain of memoryless "elementary" steps

$$I_1 \mapsto \ldots \mapsto I_M \mapsto R, \quad \text{with i.i.d. transition times } T_1, \ldots, T_M. \tag{6.3.8}$$

(As usual, i.i.d. stands for *independent and identically distributed* r.v.s.) If r denotes the daily transition probability for each step, then elementary transition times follow the geometric distribution:

$$T_m \sim f_{T_m}(t) = (1-r)^{t-1} r = r\, \mathrm{e}^{(t-1)\ln(1-r)}, \quad t \in \mathbb{N}. \tag{6.3.9}$$

As a result, the overall transition time $T^{R|I} = T_1 + \ldots + T_M$ follows a gamma distribution $f_{T^{R|I}}$, which is indeed localized.

This transition necessarily also covers regulatory aspects, namely policy response to detection of an infected flock. In our model, this response always ends with removal of the flock and stochasticity only affects the time intervening between detection and removal (a matter of a few days in The Netherlands). Note also that the simplistic model above parameterizes this transition simply in terms of r and (possibly) of M. Alternatively, one could set up a *within-farm* model describing the flock in detail up to detection (usually by observation of clinical signs). Although such a multiscale model may represent reality more accurately, the data must support it; either way, a model of that sort falls far outside the scope of this presentation.

6.3.3 Parameterization

Once supplemented with parameter values $p_{kk'}$, r, and M, the model above can be simulated to provide insight into infection spread. Here also, as in Sect. 6.2, epidemiological parameters must be estimated by data fitting, a process that effectively reverses simulation by inferring input (parameter values) from output (outbreak evolution). There, working the parameterization out in detail effectuated the realization that our data supported a more specific infection model than the original. The situation is similar here: through additional modeling, we will reduce drastically the number of elemental infection probabilities $p_{kk'}$.

6.3.3.1 Revisiting the Infection Module

Our FOI in Eq. (6.3.6) uses the laws of probability to aggregate individual influences $p_{kk'}$ but says nothing about what those influences might be. A physical infection

process should *constrain* those terms to a lower-dimensional set of *hyperparameters* θ and reduce accordingly the search directions in parameter space during optimization (*data fitting*). Since infection first spreads locally, a decisive factor is the between-farm distance $r_{kk'}$. To reflect that, we make the modeling choice to constrain infection probabilities through an algebraic relation $p_{kk'} = K(r_{kk'} \mid \theta_{I|S})$, for some positive decreasing function K controlled by hyperparameters $\theta_{I|S}$. Common sense dictates that these hyperparameters should include the amplitude of K and a length scale r_0 representing interaction range, but additional parameters tuning other features of K may be included as needed.

A significant feature of a kernel of this form is that it approximates transmission by an isotropic process, as it incorporates no other (e.g., angular) information on farm locations. Specifically, it disregards anisotropic and network-like components, such as wind- or transport-mediated transmission. Interestingly, isotropic kernels can describe both local (diffusion) and nonlocal infection spread (Lévy flights), so data fitting can potentially answer whether a given dataset *supports* nonlocal spread or not. Here also, such information can supplement contact tracing methods in understanding infection dynamics. A solid mathematical understanding *within this context* of the link between kernel-based approaches and PDEs would be a welcome addition to the domain literature.

6.3.3.2 Parameter Inference

Once the infection module has been reparameterized by $\theta_{I|S}$ and disease progression and culling by some separate set $\theta_{R|I}$, we must return to parameter inference. As we saw, MLE specifically yields parameter values maximizing the probability of observing the given data. In technical terms, the probability of observing the specific instantiation $(x^0, \ldots, x^T)$ of system states $(X^0, \ldots, X^T)$ reported in the data is

$$\begin{aligned} L\left(\theta \,\middle|\, x^0, \ldots, x^T\right) &= \prod_{t=0}^{T-1} f_{X^{t+1}|X^t,\Theta}\left(x^{t+1} \,\middle|\, x^t, \theta\right) \\ &= \prod_{t=0}^{T-1} \prod_{k=0}^{K} f_{X_k^{t+1}|X^t,\Theta}\left(x_k^{t+1} \,\middle|\, x^t, \theta\right). \end{aligned} \tag{6.3.10}$$

The probability distributions here are also functions of the (hyper)parameter set Θ, making the overall probability L of observing the data (*likelihood*) depend on those as well. Maximizing the likelihood yields *optimal* parameter values in view of the existing data. Once these values are known, together with some measure of variance accounting for uncertainty in their estimation, simulation-based inferences on infection spread can be properly drawn.

It is worth working out here likelihood (6.3.10) for our familiar three-farm model. For concreteness, we assume that the farm $k = 0$ starts off as infected at time $\tau_0^{I|S} = 0$ (original infector); system transitions are further assumed to occur at

times (part of the data) satisfying $0 < \tau_1^{I|S} < \tau_2^{I|S} < \tau_0^{R|I} < \tau_1^{R|I} < \tau_2^{R|I}$. Writing $t_k^{R|I} = \tau_k^{R|I} - \tau_k^{I|S}$ for the infection period of farm k, we then obtain the likelihood

$$\begin{aligned} L\left(\theta \left| x^0, \ldots, x^T \right.\right) &= (1-p_{10})^{\tau_1^{I|S}-1} p_{10} \\ &\times (1-p_{20})^{\tau_2^{I|S}-1} (1-p_{21})^{\tau_2^{I|S}-\tau_1^{I|S}-1} [1-(1-p_{20})(1-p_{21})] \\ &\times f_{T^{R|I}}(t_0^{R|I}) f_{T^{R|I}}(t_1^{R|I}) f_{T^{R|I}}(t_2^{R|I}). \end{aligned} \tag{6.3.11}$$

The first product component aggregates the escape probability for farm $k = 1$, and the second one is its capture probability. The next three components do the same for farm $k = 2$, and the final three quantify the probability of the given infectious periods. These terms depend, of course, on the problem parameters $\theta = (\theta_{I|S}, \theta_{R|I})$.

6.3.3.3 Infection Times

The parameter inference based on the likelihood of Eq. (6.3.10) presupposes that individual transition times $\tau^{I|S}$ and $\tau^{R|I}$ are known. According to our description of the data collection, however, culling times $\tau^{R|I}$ are available, but infection times $\tau^{I|S}$ may not be. The normative way out is to *marginalize* the likelihood and work with the probability that a given farm is culled on a specific day $\tau^{R|I}$. That probability is obtained by summing the probabilities of *all* $(\tau^{R|I}, \tau^{I|S})$-pairs with $\tau^{I|S}$ *unknown*, similarly to the construction that yielded $f_{T^{R|I}}$ from Eq. (6.3.9). Contrary to the situation there, however, an analytic formula is here out of reach because the unknown infection times $\tau^{I|S}$ affect and are affected by the FOI through Eq. (6.3.6). In practice, one introduces individual infection times into the estimation scheme as model parameters. Due to the high dimensionality of the resulting problem, estimation requires advanced computational tools well beyond the scope of this chapter. Quantifying the certainty with which infection times can be estimated and understanding their effect on the primary estimation problem for the *epidemiological* parameters would be a welcome mathematical contribution to the field.

6.3.4 Genetic Module

The framework above calibrates an infection model using farm locations and available temporal data. Increasingly frequently, such information is supplemented by *genetic* data and, specifically, by RNA sequences of virus isolates collected at infected farms. Incorporating such information into the framework developed so far is not trivial, as one cannot simply define a "genetic infectious process." Indeed, unlike geographic closeness, genetic similarity is *not* a driver of infection but *evidence* of it and must be incorporated as such into the model.

The ramifications of this remark are easy to unravel. First off, genetic similarity is a pairwise measure linking infected farms to potential, individual infector farms, whereas our infection module aggregated individual infector potencies into an FOI. To account for this, we need to modify the module to accommodate causal infection relations (*infection trees*), which resolve infection history at the individual level. We do this below by adding a genetic module providing evidence of such causality. The new setting effectively amounts to a *Bayesian network*, which includes our original model as a marginal over infection trees.

6.3.4.1 Parameter Inference

To ease the reader into the new framework, we revisit our MLE-based parameter inference scheme of Sect. 6.3.3 that produced optimal parameter values. Bayes law enables us to go one step beyond producing specific parameter values and view parameters as r.v.s having a *posterior probability distribution*

$$f_{\Theta|T} = \frac{1}{f_T} f_{T|\Theta}\, f_\Theta. \tag{6.3.12}$$

Here, f_T is a normalization constant. In this setting, both the parameter values Θ and temporal data T are seen as (multi-dimensional) r.v.s and their specific instantiations written as θ and t. The *prior distribution* $f_\Theta(\theta)$ on parameters encodes our knowledge (or assumptions) on what these values may be, whereas $f_{T|\Theta}(t|\theta)$ is the model-specific *likelihood* of observing the temporal data t given specific parameter values θ—here, it is the marginalization of Eq. (6.3.10) over the unknown infection times. Combined as in Eq. (6.3.12), these yields the conditional distribution $f_{\Theta|T}(\cdot|t)$ for the parameter values given the temporal data t. Note that our earlier MLE scheme corresponds to maximizing the posterior after postulating a *flat* prior—that is, after assuming all parameter values to be equally likely *a priori*.

This Bayesian framework is important here because it allows us to assign specific infectors to infected farms—that is, to introduce *infection trees* Y. Crucially, these shall enable us to exploit genetic similarities between infected farms and (potential) infectors. We treat Y as a r.v. and model, below, the probability $f_{Y|T,\Theta}(y|t,\theta)$ of a *specific* infection tree y given temporal data t and hyperparameters θ. From that probability, we can pass to a joint probability distribution of infection trees *and* hyperparameters through the laws of probability,

$$f_{Y,\Theta|T} = \frac{1}{f_T} f_{Y|T,\Theta}\, f_{T|\Theta}\, f_\Theta. \tag{6.3.13}$$

Here again, f_T is a normalization constant. The simpler model (6.3.12) can be recovered from this one by marginalizing over all infection trees Y.

The final module is designed to assimilate genetic data into our parameter inference scheme. As we discussed above, genetics do not drive infection spread

but provide evidence on the (im)probability of infection tree branches. To assimilate genetic data, we extend Eq. (6.3.13) through

$$f_{Y,\Theta|T,S} = \frac{1}{f_{T,S}} f_{S|Y,T,\Theta}\, f_{Y|T,\Theta}\, f_{T|\Theta}\, f_{\Theta}. \tag{6.3.14}$$

The new, genetic module $f_{S|Y,T,\Theta}(s|y,t,\theta)$ is the probability of observing the RNA sequences s reported in the data, under specific parameter values θ, temporal data t, and a (postulated) infection tree y. Parameters values are then distributed according to the marginal

$$\begin{aligned} f_{\Theta|T,S}(\theta|t,s) &= \sum_{y} f_{Y,\Theta|T,S}(y,\theta|t,s) \\ &= \frac{f_{T|\Theta}(t|\theta) f_{\Theta}(\theta)}{f_{T,S}(t,s)} \sum_{y} f_{S|Y,T,\Theta}(s|y,t,\theta)\, f_{Y|T,\Theta}(y|t,\theta). \end{aligned} \tag{6.3.15}$$

Here also, estimation of the model parameters Θ can proceed by maximizing this posterior; however, this does not cover the uncertainty inherent in parameter estimation (see Appendix). More sensibly, one can sample that distribution and use the resulting parameter values to simulate the stochastic epidemic model developed in Sect. 6.3.2. Repeatedly sampling and simulating yields outcome statistics that factor in both model and parameter uncertainty.

6.3.4.2 Infection Trees

An *infection tree* in our setting is a r.v. in the form of a $(K+1)\times(K+1)$ binary matrix Y, with (k,k') element $Y_{kk'} = 1$ if farm k' infected farm k and $Y_{kk'} = 0$ otherwise. Rows corresponding to the original infector ($k=0$) or to farms that remained uninfected throughout are identically zero. We allow here multiple *infectors*, so that the remaining rows can have an arbitrary number of units; see the discussion below. We also work with *rooted* trees, meaning that the farm that caused the outbreak (original infector) is assumed to be known with certainty. For example, the probable infection trees for the three-farm model with the settings leading to Eq. (6.3.11) are

$$y_1 = \begin{bmatrix} 0\,0\,0 \\ 1\,0\,0 \\ 1\,0\,0 \end{bmatrix}, \quad y_2 = \begin{bmatrix} 0\,0\,0 \\ 1\,0\,0 \\ 0\,1\,0 \end{bmatrix}, \quad y_3 = \begin{bmatrix} 0\,0\,0 \\ 1\,0\,0 \\ 1\,1\,0 \end{bmatrix}. \tag{6.3.16}$$

This ensemble expresses that farm $k=1$ was necessarily infected by the original infector $k=0$, whereas farm $k=2$ may have been infected by either (or both) of them; recall that $0 < \tau_1^{I|S} < \tau_2^{I|S} < \tau_0^{R|I} < \tau_1^{R|I} < \tau_2^{R|I}$ by earlier assumptions.

In our work above, the probability of a specific time trajectory was built from daily, individual farm transitions; recall Eq. (6.3.10). The salient component of that scheme was Eq. (6.3.6), which expressed the infection pressure exerted on farm k by aggregating contributions from *potential* infectors. The *actual* infector(s) in that model remained unknown due to aggregation. In the current setting, a (postulated) infection tree y yields the set

$$\mathscr{I}(k|y) = \{k' \mid y_{kk'} = 1\}, \tag{6.3.17}$$

which identifies explicitly the infectors of k. With this information in hand, we can build an infection module from individual transitions but with transition probability

$$f_{X_k^{t+1}|X^t}(I|x) = \left(\prod_{k' \in \mathscr{I}_t \setminus \mathscr{I}(k|y)} (1 - p_{kk'})\right)\left(\prod_{k'' \in \mathscr{I}(k|y)} p_{kk''}\right). \tag{6.3.18}$$

(Here again, $x_k^t = S$, by assumption.) Aggregation occurs separately for infectors and non-infectors, and trees where infections antedate the appearance of infectors are assigned zero probability. The escape probability is $f_{X_k^{t+1}|X^t}(S|x) = \prod_{k' \in \mathscr{I}_t} (1 - p_{kk'})$, so it complements the probabilities above by virtue of the identity

$$\sum_{I=0}^{|\mathscr{I}_t|} \sum_{k'_1,\ldots,k'_I} \left(\prod_{k' \notin \{k'_1,\ldots,k'_I\}} (1 - p_{kk'})\right)\left(\prod_{i=1}^{I} p_{kk'_i}\right) = 1.$$

The inner sum here ranges over I-element subsets of $\mathscr{I}_t$, with each summand corresponding to the probability of simultaneous infection by a specific subset of I infectors; the case $I = 0$ is the escape probability. A consequence of this is that FOI-based infection probabilities may be recovered by summing tree-based ones, as long as farms are allowed to have multiple infectors. Naturally, this corresponds to an immense number of admissible trees, with multiple-infector ones being higher order in the already small pairwise daily infection probabilities. In practice, these are extremely improbable and effectively discarded by tree-sampling algorithms. An alternative is to include those minuscule higher order terms in the escape probability, see again our discussion in Sect. 6.3.5.

Revisiting our three-farm system with event times as previously assumed, we find, for example, for infection tree y_1 the probability

$$\begin{aligned} f_{Y|T,\Theta}(y_1|t,\theta) = {} & (1 - p_{10})^{\tau_1^{I|S}-1} p_{10} \\ & \times (1 - p_{20})^{\tau_2^{I|S}-1} p_{20}(1 - p_{21})^{\tau_2^{I|S}-\tau_1^{I|S}} \\ & \times f_{T^{R|I}}(t_0^{R|I}) f_{T^{R|I}}(t_1^{R|I}) f_{T^{R|I}}(t_2^{R|I}). \end{aligned} \tag{6.3.19}$$

Effectively, the aggregated capture probability by *either* farm in Eq. (6.3.11),

$$1-(1-p_{20})(1-p_{21}) = p_{20}(1-p_{21}) + (1-p_{20})p_{21} + p_{20}p_{21},$$

has been replaced here by $p_{20}(1-p_{21})$. This term models capture by $k'=0$ and escape from $k''=1$, which is precisely the scenario encoded in y_1.

6.3.4.3 Genetics

The tree-based model above increases model complexity substantially but does not necessarily lead to more accurate predictions; see the discussion in Sect. 6.3.5. To harvest its potential, we must assimilate *evidence* of infection branches; this is precisely where genetic data come into play. The significant element in genetic data is that each infected farm is assigned the *RNA sequence* of the virus strain isolated in it. Such a sequence is specifically modeled as a "word" and formed by collating a number L of "letters" (*base pairs*) drawn from the alphabet

$$\mathscr{P} = \left\{ \begin{matrix} g & a & u & c \\ | & | & | & | \\ c & u & a & g \end{matrix} \right\}.$$

These pairs are composed of the *RNA bases* g(uanine), a(denine), u(racil), and c(ytosine). Each base pairs with a *unique* counterpart; hence, a double-stranded sequence can be recovered from a single, consistently selected (i.e., top or bottom) strand. On account of this, we work below with single-stranded sequences—that is, words of length L from the single-base alphabet $\mathscr{B} = \{g, a, u, c\}$.

We build $f_{S|Y,T,\Theta}$ using infection events—i.e., infection tree branches—through

$$\begin{aligned} f_{S|Y,T,\Theta}(s|y,t,\theta) &= \prod_{k\in\cup_\tau \mathscr{I}_\tau} f_{S_k|Y,T,\Theta}(s_k|y,t,\theta) \\ &= \prod_{k\in\cup_\tau \mathscr{I}_\tau} \prod_{\ell=1}^{L} f_{S_{k\ell}|Y,T,\Theta}(s_{k\ell}|y,t,\theta). \end{aligned} \tag{6.3.20}$$

Here, $s = (s_1, \ldots, s_K)$ are the observed RNA sequences and we have written that of the kth farm as $s_k = s_{k1}\ldots s_{kL}$; each $s_{k\ell} \in \mathscr{B}$ here is a single base, so that s_k is a string of such bases. The outer product runs over all farms infected during the outbreak and the inner over RNA loci, so this model assumes that loci mutate independently of each other. The elemental probability $f_{S_{k\ell}|Y,T,\Theta}$ corresponds to observing a specific ℓth base in the RNA sequence of the kth farm and demands additional modeling.

We envision infection as transmission of viral strains between farms with the possibility of *mutation*, so that the RNA sequence observed at an infected farm is a mutant of those in its infectors. From a modeling perspective, the

infection tree y serves to identify the "origins" of s_k through the farm's infector(s); cf. Definition 6.3.17. In view of that, we must compare the ℓth base of the infected farm to the homologous bases of its infector(s),

$$f_{S_{k\ell}|Y,T,\Theta}(s_{k\ell}|y,t,\theta) = f_{S_{k\ell}|S_{\mathscr{I}(k|y)\,\ell}}\left(s_{k\ell}|\cup_{k'\in\mathscr{I}(k|y)}\{s_{k'\ell}\}\right). \tag{6.3.21}$$

The standard model for base mutations is [21], which treats them as occurring with time- and loci-invariant probabilities,

$$f_{S_{k\ell}|S_{k'\ell}}(b|b') = P_{bb'}, \quad \text{with} \quad P = \begin{bmatrix} p_w & p_u & p_v/2 & p_v/2 \\ p_u & p_w & p_v/2 & p_v/2 \\ p_v/2 & p_v/2 & p_w & p_u \\ p_v/2 & p_v/2 & p_u & p_w \end{bmatrix}. \tag{6.3.22}$$

Here, bases are ordered as in $\mathscr{B}$ and P is a Markov matrix, hence $p_w = 1 - p_u - p_v$. This simple model divides bases into *purines* (g, a) and *pyrimidines* (t, c) and specifies transition probabilities p_u and p_v within and between these families (*transitions* and *transversions*). If farm k has a single infector $k' = \mathscr{I}(k|y)$, then this model can be applied verbatim. In the case of multiple infectors, one must formulate another scheme, for example, compare to the infected a randomly sampled (or the most probable) infector. Since multi-infector trees are improbable, any reasonable choice should suffice provided one can demonstrate that specific choices do not affect end results crucially. We stress, at this point, that working with genetic sequences demands developing a certain understanding of specifics, so that one feels at ease with such terms as *quasispecies*, *consensus sequence*, and *conserved regions*.

Finally, we return one last time to our three-farm model as we left it in Eq. (6.3.19). We write once again $s_k = (s_{k1}, \ldots, s_{kL})$, for $k = 1, 2, 3$, and recall that farm $k = 0$ infects both $k = 1$ and $k = 2$ according to y_1. Next, we write u_k, v_k, and $w_k = L - u_k - v_k$ for the number of transitions, transversions, and conservations between the RNA sequences of farm $k = 1, 2$ and $k = 0$. It now follows from Eqs. (6.3.20)–(6.3.22) that

$$f_{S|Y,T,\Theta}(s|y_1,t,\theta) = \left(p_u^{u_1}\,p_v^{v_1}\,p_w^{w_1}\right)\,\left(p_u^{u_2}\,p_v^{v_2}\,p_w^{w_2}\right) = p_u^{u_1+u_2}\,p_v^{v_1+v_2}\,p_w^{w_1+w_2}. \tag{6.3.23}$$

Substituting into Eq. (6.3.14) from this formula and from earlier results, we obtain the posterior $f_{Y,\Theta|T,S}(y_1,\theta|t,s)$ as needed.

6.3.5 Discussion

In our work above, we extended a classic, discrete SIR model to accommodate genetic evidence. Our construction of the inferential framework relied on elementary probability laws, but the differentiation between *mechanism* and *evidence* was a crucial element in devising the data assimilation scheme (6.3.14) and (6.3.15). The

model itself involved numerous modeling assumptions, among which Eq. (6.3.6) for the FOI. This affected the framework strongly through the likelihood (6.3.10) and encapsulated our assumption that I-farms challenge S-farms *independently*. In formulating it, we effectively modeled daily infection risk through a series of independent, non-identical, biased coin tosses (*Bernoulli trials*): an infection occurs if *at least one* succeeds. *Independence* is intuitive and mathematically appealing but opens the door to *multiple* infectors (multiple successes), which then propagates into infection trees and sequence matching; recall our discussion of Eqs. (6.3.18) and (6.3.22). To avoid this, one can impose a *single infector* by excluding multiple success events from the event space, but that destroys challenge independence; we invite the reader to verify this for our three-farm model and ponder on the implied dichotomy.

6.3.5.1 Effect of Infection Time Uncertainty

The major complication in the inferential framework based on Eq. (6.3.10) is lack of knowledge of infection times. The essence of the problem is that the resolution of parameters affecting the distribution of two random variables, T_1 and T_2, must be based on observations of their sum, $T = T_1 + T_2$. In an idealized scenario, where $T_i \sim f_i(t) = \theta_i \mathrm{e}^{-\theta_i t}$ with $\theta_1 \neq \theta_2$, one has

$$T \sim f(t) = \frac{\theta_1 \theta_2}{\theta_2 - \theta_1} (\mathrm{e}^{-\theta_1 t} - \mathrm{e}^{-\theta_2 t}).$$

Given observations $t_1, \ldots, t_N$ for T, the likelihood then reads

$$L(\theta | t_1, \ldots, t_N) = \left(\frac{\theta_1 \theta_2}{\theta_2 - \theta_1} \right)^N \prod_{n=1}^{N} \left(\mathrm{e}^{-\theta_1 t_n} - \mathrm{e}^{-\theta_2 t_n} \right).$$

Exponential sums are difficult to fit accurately [32] and hence, although an optimal parameter set is identifiable by MLE *in principle*, the confidence region around it may be so wide as to render it useless *in practice*. How these considerations play out in a network setting where interactions are described by multiple parameters is unclear. A more thorough mathematical investigation, starting from simple, idealized network models would be a welcome contribution.

6.3.5.2 Effect of Infection Tree Uncertainty

We incorporated genetics into the framework in two steps, first by passing to infection trees in Eq. (6.3.13) and then by assimilating genetic data through scheme (6.3.14). Model (6.3.13) only uses temporal data, so it can in principle be implemented to yield causal information. In The Netherlands, where outbreak data shows poultry farms forming dense clusters with multiple infections per

day, one cannot expect that temporal model to be able to resolve infection trees; $f_{Y,\Theta|T}$ would have a wide support over many trees. Tree inference is also nuanced mathematically, as the number of parameters to resolve—the infection branches—grows with the number of infection events. In a classical setting, estimators return a fixed number of parameter values so estimator variance is reduced by additional data; the sampling distribution converges in probability. This is not the case here, meaning that additional cases do not necessarily mitigate uncertainty. To obtain more accurate estimates, one must use an *independent* information channel—that is where genetic data enter the game. This should be contrasted to our approach in Sect. 6.2, where this issue was circumvented by *aggregating* individuals into a fixed number of homogeneous compartments. It would be interesting to see this examined in a detailed mathematical manner in this or a similar context.

6.3.5.3 Effect of Mutation Rate Uncertainty

In a similar vein, good prior estimates on mutation rates $p = (p_u, p_v, p_w)$ must be available, since assessing genetic similarity in a dataset using a model *and* inferring model parameters using that dataset is evidently circular. This can be seen in a simplified model, where we fix a sequence s_0 and use Eq. (6.3.22) repeatedly on each base (with known, fixed P) to generate a *linear chain* of sequences $s^* = (s_1, \ldots, s_K)$. Let us assume that the *set* $\{s_1, \ldots, s_K\}$ is given and our task is to *order* it—that is, to infer the *chain* $(s_1, \ldots, s_K)$. The likelihood of a given chain $s = (s_{i_1}, \ldots, s_{i_K})$ can be approximated, for $KL \gg 1$, by the bivariate Gaussian [14]

$$L(s|p) = \frac{\mathrm{e}^{-H_0(p|\hat{p})}}{2\pi\sqrt{\det \Sigma(\hat{p})}}, \quad H_0(p|\hat{p}) = \frac{1}{2}(p-\hat{p})^{\mathrm{T}}\Sigma^{-1}(\hat{p})(p-\hat{p}). \tag{6.3.24}$$

Here, $\hat{p} = (u, v, w)/KL$ are the *empirical* mutation rates for s, inferred from the total number of transitions (u), transversions (v), and conservations (w) it exhibits. The covariance matrix $\Sigma(\hat{p})$ is the inverse Hessian at $\hat{p}$ of the function

$$H(p|\hat{p}) = -KL\big(\hat{p}_u \ln(p_u) + \hat{p}_v \ln(p_v/2) + p_w \ln(\hat{p}_w)\big) \tag{6.3.25}$$

around its maximum $\hat{p} = p$. Only chains with $\hat{p} - p = \mathcal{O}(1/(KL))$ are probable; the rest have exponentially small probabilities. Since rearrangements of s^* generate (a subset of) a square lattice on $(\hat{p}_s, \hat{p}_v)$-plane with step size $1/(KL)$, only an $\mathcal{O}(1)$ number of sequences among $K!$ possible rearrangements are probable. As a consequence, the likelihood of a candidate chain s is effectively determined by how close its empirical rates $\hat{p}$ are to the actual rates p in a Mahalanobis-like distance; this demonstrates the crucial role of information on the actual mutation rates. Here also, an exhaustive mathematical study of such problems would be of definite interest.

6.4 A Case of Bad Data

Although our work up to now focused on transmission mechanisms at various levels of detail, modeling work in the health sciences is far from limited to disease transmission. In this section, we consider a very different, application-oriented question that arose during a particular study of the *population dynamics of immune cells*. The subject matter is inference of the *turnover rate* of certain cell types, which at first glance should not be radically different from our earlier work as it is an inferential task. Indeed, data can be fitted to standard ODE models that describe well the dynamics of the large and (presumably) homogeneous cell populations *in vivo*. As will become apparent below, though, the data at our disposal do *not* fit those models at all. It will be our task in this section to seek the root cause and develop a remedy for this situation. Interestingly, and although the population model is ODE-based, our work below owes more to *data-driven statistical modeling* and bona fide detective work than to DEs. Here, more than elsewhere, we encourage a hands-on approach to the problem; to that end, we have made datasets and code available [14].

6.4.1 Data and Question

6.4.1.1 Experiment

Before diving into the data, it is necessary to give a short overview of the experimental protocol that generated it. As mentioned earlier, the experiment in question was designed to measure cellular turnover rates—that is, rates at which immune cell populations renew themselves. Under physiological conditions, the primary cellular processes—cell division, death, and migration between tissues—balance each other out and the population maintains its size. Because of this, assessing that rate requires that one disentangles processes that add to/subtract from the population, i.e., that one differentiates between "new" and "old" cells. This problem has a fascinating history intertwining experimental design with theory—see [34] as a starting point—with the main idea being to count newly generated cells by *labeling* them. In the past, labels were radioactive isotopes and counting involved Geiger counters. At present, the label of choice is *deuterium* (^{2}H), which replaces hydrogen atoms in *de novo* synthesized *adenine deoxyribose* (dR) molecules (*moieties*) within cellular DNA. *Labeled* and *unlabeled* moieties (written dR* and dR) differ by molecular weight, which creates a window of opportunity for their experimental differentiation.

The data at our disposal was collected from a group of young goats according to a strict experimental protocol. Build-up of dR* in cellular DNA was effectuated by administering heavy water (^2H_2O) over an initial *uplabeling* time period $[0, \tau]$. The administration period was followed by a *washout* stage during which no deuterium was administered in any form, so that dR* in DNA decreased due to cell death. A model for label proliferation was developed in [31], with build-up and

loss proportional to ^2H_2O levels in the body and to the labeled population size, respectively. The resulting linear ODEs were integrated analytically to yield the *enrichment ratio* E—that is, the proportion of labeled adenosine moieties,

$$E(t) = h(t, t), \ t \leq \tau; \quad E(t) = h(t, \tau), \ t \geq \tau, \tag{6.4.1}$$

where we have defined the auxiliary function

$$h(x, y) = \frac{cpf}{\delta - d}\left[\frac{\delta}{d}\left(\mathrm{e}^{dy} - 1\right)\mathrm{e}^{-dx} - \left(\mathrm{e}^{\delta y} - 1\right)\mathrm{e}^{-\delta x} + \frac{\beta}{f}\left(\mathrm{e}^{-dx} - \mathrm{e}^{-\delta x}\right)\right].$$

Values for the parameters δ, β, and f are determined by procedures we shall not cover, so fitting model (6.4.1) to a time series is a matter of estimating the turnover rate d and the additional parameter cp. Figure 6.2a shows a representative time series $E(t_1), \ldots, E(t_N)$ for this setup derived from an older experiment, together with the corresponding least squares fit. Note carefully that both the data and the fit exhibit clear uplabeling and washout stages.

6.4.1.2 Question

Although the data in Fig. 6.2a trace well the labeling curve described by (6.4.1), the data from our experiment mostly do not. Figure 6.2b shows a particular time series that deviates remarkably from the expected trend. Our task is to identify the root cause of that behavior and, if possible, rectify things. To achieve this, we must investigate the *pipeline* turning cellular counts into time series (one per cell type, see [14]). We describe the procedure briefly below, using hindsight to eliminate much of the inessential complexity.

Label Administration Deuterium was administered by mixing heavy water into milk with a concentration of 2–3%; the goats had no other drinking sources. That concentration was kept constant throughout the uplabeling period, which started on day zero. The washout period lasted from week 4 to 14, during which time the goats only received ordinary drinking water.

Sample Collection Biological material was collected on a weekly basis from blood, bone marrow, and various organs and lymph nodes; see [14]. Each goat was sampled *once* in time but yielded samples from several organs; as a result, distinct points on an enrichment time series necessarily correspond to different animals. The sole exception to this were blood samples, which were collected weekly from the same goats. Each sample was marked clearly with the goat and organ from which it was collected, so mix-ups are improbable.

Cellular Extraction Single cell suspensions were made from the collected material, independently per sample. Cells were stained with fluorescent antibodies directed against different cellular markers, fixed and kept overnight at 4 °C. Those

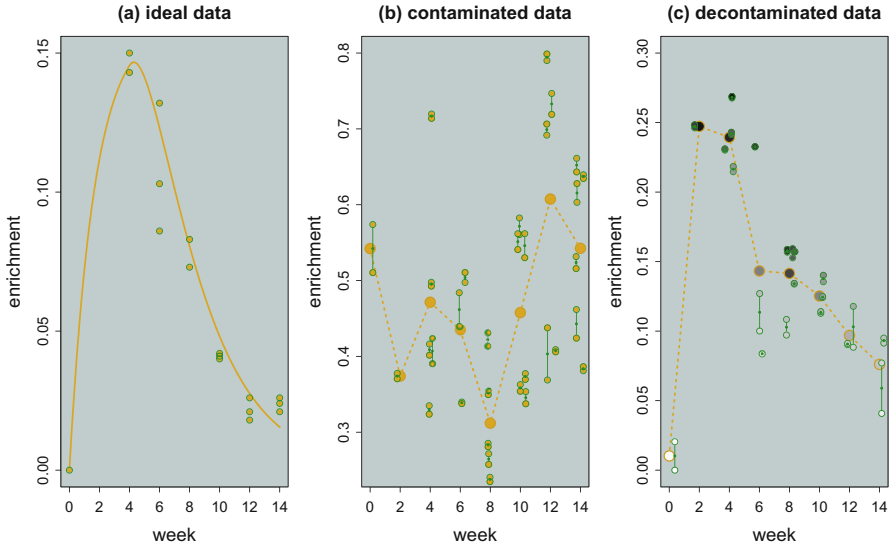


Fig. 6.2 Enrichment data from two distinct datasets. Panel (**a**): A dataset (golden/green points) and its best fit to model (6.4.1) (solid golden curve), demonstrating the characteristic behavior described in the main text: enrichment increases towards saturation while heavy water is administered (up to week 4) and then decreases exponentially. Multiple points in the same week indicate multiple independent samples; with the exception of week 6, these yield similar enrichment levels. Panel (**b**): A much less well-behaved dataset. Samples here were measured twice, yielding pairs of data points joined by a vertical line with the average marked on it (solid green point). Samples correspond to weeks $0, 2, 4, \ldots, 14$, but we have perturbed their abscissae by a fractional amount to enhance visibility. Also shown (solid golden points) are data averages per week; these outline a trend (golden dotted line) that deviates markedly from the characteristic behavior in panel (**a**). Panel (**c**): The dataset of panel (**b**) after decontamination (cf. Sect. 6.4.4). Data points and their weekly averages are shaded by the proportion of contaminant $1 - a_s$ in the original profiles, see Eq. (6.4.6); as the shade darkens, the proportion decreases from 100% to 0%. Variation between duplicates and distinct samples in the same week is reduced appreciably, and the trend also matches the characteristic behavior much better

cells underwent fluorescent-activated cell sorting (FACS) the day after, and only the cells of interest were retained per sample.

DNA Purification DNA was extracted from the retained cells of each sample using a standard DNA extraction kit according to manufacturer guidelines. Through this final preparatory step, each cellular sample (i.e., goat and organ) yielded a single DNA sample.

Enrichment Quantification Labeled and unlabeled moieties in each sample were counted *twice* using gas chromatography/mass spectroscopy (GC/MS). Specifically, samples from the same organ were run through GC/MS independently but *in tandem*, yielding (double) counts M^* and M of labeled and unlabeled molecules per sample. The *enrichment* value for each sample was obtained as the ratio

$E = M^*/(M + M^*)$, for each of the duplicate runs, and assigned to the time t at which the biological material was collected.

Time Series For each organ, a time series $E(t_1), \ldots, E(t_N)$ collecting enrichment levels from different goats was produced; recall Fig. 6.2b.

The glitch can be anywhere in this process, starting with the obvious but least specific candidate: *biological variability*. Indeed, since the enrichment time series for each organ collates data from different animals, *some* degree of variability must be expected; this is evident already in the well-behaved data of Fig. 6.2a. However, this causative candidate is overly pessimistic, if not outright facile, and it also fails to explain the high deuteration levels deep in the washout period. Since rectification potential decreases as one goes deeper into the pipeline, we work backwards starting from GC/MS quantification of DNA samples. Identifying putative problems in that step presupposes an understanding of how GC/MS works; we cover basics below.

6.4.2 Enrichment Quantification

The function of the GC/MS unit is to *count* accurately the dR and dR* molecules in a sample by *separating* them from all other molecules. It achieves that by means of a highly specific, two-tier molecular identification process that combines elements of chromatography and mass spectroscopy. Concretely, a sample undergoes a first separation into its constituents by being passed through a chromatographic column. Different molecules traverse that column with different speeds and emerge from it at *different times*; that is the first differentiation level. Of course, exit times vary stochastically among chemically identical molecules, so distinct molecular species can still *interfere* with dR and dR* moieties. Our samples contain *thousands* of distinct DNA fragments, so interference is inevitable and compromises specificity. To enhance it, the unit ionizes emerging molecules and passes them through a mass spectrometer, so that they are separated by their *mass-to-charge (m/z) value*. Ionized molecules with distinct m/z values are then counted individually. By combining these two stages, we *tag* each molecule by its detection time (t) *and* m/z value (r). This double-tagging increases differentiation by reducing drastically the interference between dR or dR* molecules and other substances.

Importantly, GC/MS reports a list of detection times and m/z values per molecule and *not* summary molecular counts M and M^*. In practice, time is quantized into successive intervals $I_1, \ldots, I_T$ lasting time δt (a few milliseconds), whereas m/z values are quantized naturally into $r_1, \ldots, r_K$. The output data assumes a matrix form C, with c_{tk} the number of molecules detected in time interval I_t at m/z value r_k. Here, we focus on dR and dR* molecules with known m/z-values r_k and r_{k^*}, so we consider time series (*profiles*)

$$g = \{c_{tk} \,|\, t = 1, \ldots, T\} \text{ for dR}, \quad g^* = \{c_{tk^*} \,|\, t = 1, \ldots, T\} \text{ for dR}^*. \tag{6.4.2}$$

The total numbers M and M^* of dR and dR* molecules are obtained by summing molecular counts over time,

$$M = \sum_{t=t_{\min}}^{t_{\max}} c_{tk}, \quad M^* = \sum_{t=t^*_{\min}}^{t^*_{\max}} c_{tk^*}; \qquad E = \frac{M^*}{M + M^*}. \tag{6.4.3}$$

Following the literature, we will use the shorthand AUC (*area under the curve*) for the profile counts M and M^* above. The time intervals

$$\mathscr{T} = \{t_{\min}, \ldots, t_{\max}\} \subset \{1 \ldots, T\} \text{ and } \mathscr{T}^* = \{t^*_{\min}, \ldots, t^*_{\max}\} \subset \{1 \ldots, T\} \tag{6.4.4}$$

are chosen to engulf the profiles of interest with minimal interference from other molecular species at the same m/z value; see Sect. 6.4.3.

It is important for our work below to note that normalization of the profiles (g, g^*) in Eq. (6.4.2) turns them into pmfs $f : \mathscr{T} \to \mathbb{R}^+$ and $f^* : \mathscr{T}^* \to \mathbb{R}^+$ for the molecular detection time,

$$f(t) = \frac{1}{M} g(t), \quad f^*(t) = \frac{1}{M^*} g^*(t). \tag{6.4.5}$$

Figure 6.3a,b shows a pair of (normalized) profiles derived from a commercially available pure (*control*) sample used for calibration. Such samples have superbly low noise-to-signal ratios and serve in what follows as (statistical) models for our sample-derived (*biological*) samples; their characteristic bimodality is due to the cis/trans isomerization of adenosine [8]. Biological samples are more susceptible to noise, as they contain numerous DNA fragments interfering with dR and dR* even after double selection; examples—some extreme—are shown in Fig. 6.3c–h.

6.4.3 Data Exploration

As evident from Eq. (6.4.3), the enrichment E of a sample is controlled by the count ratio M^*/M of labeled over unlabeled molecules. Errors in estimating that ratio (e.g., due to interfering molecules) are directly passed to the enrichment value, so the estimation of M^* and M is a focal point in our root cause analysis.

6.4.3.1 Integration Windows

To count dR and dR* molecules without interference, one must adhere to carefully controlled lab procedures and set tight *integration windows* $\mathscr{T}$, $\mathscr{T}^*$. In our analysis, these windows are modeled after the corresponding windows for the *control* profile(s) shown in Fig. 6.3a, b. First, we define a generously broad, crude time window that contains the entire control distribution f (respectively, f^*); all profiles

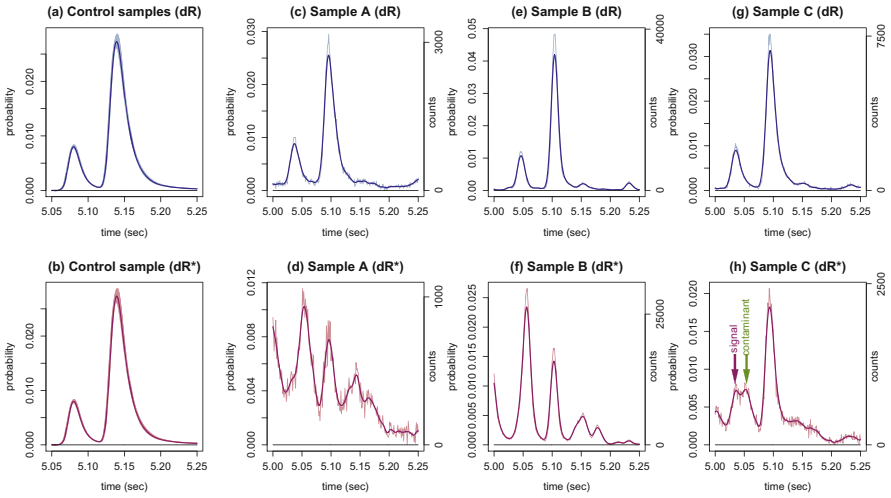


Fig. 6.3 Unlabeled (dR) and labeled (dR*) profiles for various samples. The right axis reports molecular counts per time interval (lasting dt $= 10^{-3}$ s). The left axis tabulates the pmf obtained by normalizing each profile by its AUC. (**a**–**b**) Averaged pmfs and 95% confidence intervals (CIs) derived from four control samples. The dR, dR* profiles and CIs are effectively identical; molecular counts (in the millions) are sample-specific and not reported for these averaged profiles. (**c**–**h**) dR (light blue) and dR* (pink) profiles from three biological samples corresponding to different animals and organs. The primary features of the unlabeled profiles are quite stable. Labeled profiles, instead, exhibit secondary peaks, background noise, and highly variable left modes. Profile mollification (dark blue/red) with a zero-mean Gaussian kernel ($\sigma = 3$dt) ameliorates the stochastic variation induced by low counts. In (**h**), the mollification reveals a contaminant, evident in the bimodality of the left mode; see main text for details

in Fig. 6.3 are plotted over such crude windows. Then, we define a *tighter* time window that holds 95% of the distribution's AUC. More specifically, we separate the bimodal distribution into left/right modes by means of the intervening local minimum (*valley*), compute the AUC of each mode, and crop the interval supporting each mode to retain 95% of that mode's AUC.

These (labeled/unlabeled) *control-derived* time windows are used to crop each (labeled/unlabeled) *sample-derived* profile in panels c–h after *alignment*. Indeed, it can be seen in Fig. 6.3 that distinct profiles are *shifted* in time due to machine specifics immaterial for our analysis. Estimating and removing that time shift is a classic *signal registration* problem in 1D with various possible solutions—for example, using a stable *profile feature* (right peak location, midpoint of the full width at half maximum (FWHM), or other) or a statistical approach (maximizing cross-correlation or similar). Automating this registration procedure increases pipeline throughput and eliminates human error but, concurrently, allows profiles that would have been flagged by visual inspection to pass muster; cf. panels d and f. Below, we examine the origin of such aberrant profiles and their effect on enrichment ratios.

6.4.3.2 Profile Stability

The control profiles in Fig. 6.3a, b were derived by averaging multiple (normalized) runs of distinct control samples, so we were able to also plot confidence intervals. Our first observation is that those intervals are narrow, meaning that *control* profiles are *very stable*: labeled and unlabeled profiles are well-defined and only subject to small variations. This should indeed be the case, as the shape of the detection time distribution has robust (chemical) origins and care was taken to prepare and measure samples in a controlled environment.

Next to the control profiles in panels a and b, Fig. 6.3 shows pairs of labeled and unlabeled profiles for several animals and organs. Evidently, the *unlabeled* profiles are also rather stable, cf. panels c, e, and h. Note, however, that individual profiles exhibit small random and systematic noise in the form of short-scale variations and spurious tail peaks. These effects are much more pronounced for the *labeled* profiles in panels d, f, and h, where irregularity is apparent not only in the form of additional peaks but, also, in the variable relative heights of the dominant ones. Given our earlier assertion that bimodality is a robust feature, this variability appears highly peculiar; we attempt to assess its statistical significance below.

6.4.3.3 Quantifying Profile Aberration

Figure 6.4 quantifies the AUC held by the left distribution mode for a specific cell type (i.e., time series). An analysis of *control* dR and dR* profiles shows that their left mode holds, on average, $\mu = 24\%$ of the total AUC with standard deviation $\sigma = 1\%$. The individual deviations from μ of the eight control samples are plotted in Fig. 6.4 (diamonds) in units of σ, both for the unlabeled (blue) and labeled (red) profiles. Plainly, all control samples have roughly the same left-mode AUC (μ) within a couple of standard deviations.

Repeating this analysis for the unlabeled *biological* profiles, we see that their left-mode AUC is *slightly but systematically larger* than μ by a few σ; see the 25 blue dots in the same figure. The deviations of the *labeled* profiles, on the other hand, are also positive and systematic but much more extreme. This implies that *the left-mode AUC in biological dR* profiles is significantly higher than in their pure counterparts*. A systematic deviation in the tens of standard deviations looks suspiciously close to a smoking gun, so we take a closer look below.

6.4.3.4 Root Cause of the Aberration

The nature of the problem becomes evident at close inspection of labeled profiles, such as that shown in Fig. 6.3h. The left mode here is visibly heavier than its pure counterpart in panel b and (its mollified, noise-attenuated version) features a small kink indicating bimodality. Since sufficiently separated unimodal distributions add up to bimodal ones, this finding indicates an *additional* distribution overlapping

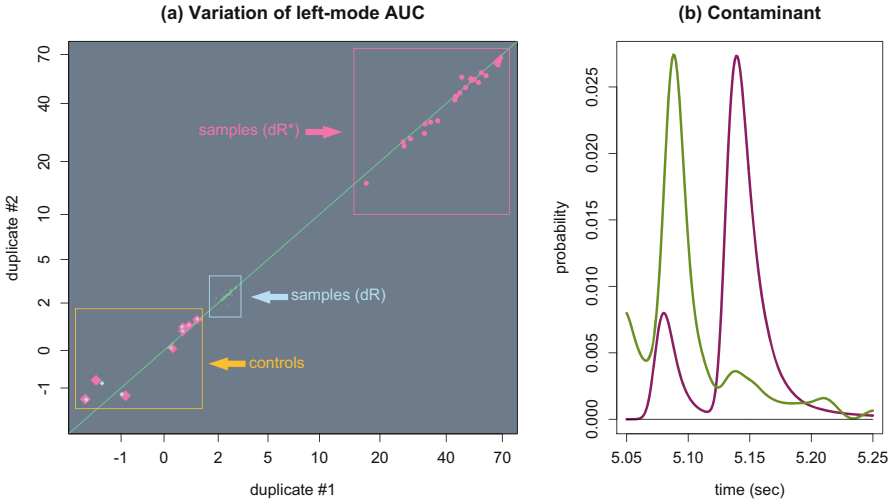


Fig. 6.4 (**a**) Scatter plot of left-mode AUC proportions for controls and biological samples measured *in duplo* (note the nonlinear scale). Each point represents duplicate measurements of one sample (blue/red: dR/dR*). The proportion statistics of controls (diamonds) are $(\mu, \sigma) \approx (0.24, 0.01)$, both for dR and dR* profiles, and all proportions are standardized so that a value n denotes the proportion $\mu + n\sigma$. Control proportions are very stable; recall Fig. 6.3a, b. Unlabeled profiles of biological samples slightly exceed μ, but labeled ones deviate by tens of standard deviations. This irregularity is systematic, seen in that duplicate variation (deviation from diagonal) is uniformly small. (**b**) Contaminant profile (green) and average (labeled) control profile (red) for the organ in Fig. 6.3a, b. Note the contaminant position relative to the control profile, which explains strong interference with left-mode AUC

with the left profile mode and elevating its AUC. Combined with earlier findings, this becomes damning evidence of *contamination*. We postulate, specifically, that irregular labeled profiles are *composite* and made up of a stable bimodal distribution (seen in control profiles) and a (still elusive) *contaminant* distribution. This hypothesis explains the elevated left-mode AUC values of such profiles and is corroborated by secondary observations, such as left modes being heavy-tailed or shifting closer to right modes. Evidently, the contaminant distribution is caused by molecules having the same m/z-value and similar detection times as dR*. Their strong influence is due to the low dR* counts (in the mere thousands, per time interval), and the sole imaginable remedy is algorithmic removal of the contaminant profile (*decontamination*).

We must remark here that root cause identification is hardly ever as linear as presented above. Instead, it is riddled with contradictory observations, false leads, and omnipresent human biases that lend it the allure of detective-like work. Modelers are brought in to furnish their distinctive mental and mathematical models but, also, their disciplinary expertise and focus on facts. Solid mathematical and scientific principles and the ability to stay the course are pivotal in that mission.

6.4.4 Data Modeling and Decontamination

We saw above that the observed deviations in enrichment values are likely due to the data being contaminated; our task in this section is to decontaminate the labeled profiles and rectify enrichment ratios accordingly. To that end, we consider a profile g^* corresponding to a biological sample and its associated distribution f^* in Eq. (6.4.5). We hypothesized that f^* is the sum of a stable bimodal distribution and a contaminant distribution; we write this as

$$f^*(t) = a_s f_s^*(t) + (1 - a_s) f_c^*(t), \quad \text{where } f^*, f_s^*, f_c^* : \mathscr{T}^* \to \mathbb{R}^+. \tag{6.4.6}$$

Here, f^* represents the observed distribution (data), f_s^* and f_c^* are blueprints (*statistical models*) of the *signal* and *contaminant* distributions, and the weight $0 \leq a_s \leq 1$ quantifies the signal content of the measured profile. This is plainly a mixture model, similar to the one in Eq. (6.2.9). Under Eq. (6.4.6), the *rectified* (i.e., decontaminated) distribution is $\bar{f}^*(t) = a_s f_s^*(t)$ that corresponds to the profile $\bar{g}^* = a_s M^* f_s^*$. As a result, the rectified enrichment ratio is

$$\bar{E} = \frac{a_s M^*}{M + a_s M^*} < E. \tag{6.4.7}$$

This result should be compared to Eq. (6.4.3). We have not rectified the profile g, as evidence points to insignificant contamination levels for unlabeled profiles; recall our earlier discussion. Through Eq. (6.4.7), rectification reduces to estimating the weight (i.e., signal content) a_s.

6.4.4.1 Signal and Contaminant Distributions

To decompose a measured distribution f^* as in Eq. (6.4.6), one must first specify signal and contaminant distributions. Previously, we considered controls to be practically contaminant-free and declared statistically significant deviations from them as evidence of contamination. On account of that, we use the average of all control distributions as f_s^*; cf. Fig. 6.3. Modeling f_c^* is more challenging, as we have no access to "pure contaminant/signal-free" samples. By the same token, however, profiles with *extreme* left-mode AUC must be superbly contaminant-rich; hence, we approximate the contaminant profile f_c^* by their average. Figure 6.4b shows the signal and contaminant profiles for the cell type considered in Fig. 6.4a.

It should be evident that our contaminant model f_c^* is a limiting step in our analysis, certainly more so than our signal model f_s^*. Indeed, as the secondary peaks in Fig. 6.4b suggest, even highly contaminated profiles hold signal from dR* molecules (middle peak) and from secondary contaminants (rightmost peak). It is important to remember this below, where we discuss data fitting schemes in detail.

6.4.4.2 Data Fitting for the Mixture Model

The distributions f_s^* and f_c^* can be tabulated for each cell type, provided that we have access to *both* controls and exceedingly contaminated samples. If this is the case, then a_s is estimated by decomposing the distribution f^* of each sample as in Eq. (6.4.6). All distributions in that equation are meant as *idealized statistical models*, but, in reality, we only have finite-sample approximations (*histograms*). It is therefore improbable that Eq. (6.4.6) will hold exactly for *any* value of a_s, and we have to resort once again to data fitting.

Data are typically fit to mixture models using expectation–maximization (EM) algorithms, which are engineered to maximize the likelihood corresponding to Eq. (6.4.6); see [3] for an overview and available computational resources. In the spirit of this chapter, we present here an intuitive approach enabled by the inherent discreteness of our setup. Recalling from Eq. (6.4.4) that time in the integration windows is *binned*, we rewrite Eq. (6.4.6) as

$$\begin{bmatrix} f_s^*(t^*_{\min}) - f_c^*(t^*_{\min}) \\ \vdots \\ f_s^*(t^*_{\max}) - f_c^*(t^*_{\max}) \end{bmatrix} a_s = \begin{bmatrix} f^*(t^*_{\min}) - f_c^*(t^*_{\min}) \\ \vdots \\ f^*(t^*_{\max}) - f_c^*(t^*_{\max}) \end{bmatrix}. \tag{6.4.8}$$

Since the two vectors are known from our tabulation of f^*, f_s^*, and f_c^*, this is an *overdetermined linear system* in one unknown (a_s) and can thus only be solved in an *approximate, optimal* sense. A normative choice is to use least squares (ordinary or otherwise), but another approach can be developed by summing both sides of Eq. (6.4.6) to obtain the relation

$$\sum_{t=t_1}^{t_2} f^*(t) = a_s \sum_{t=t_1}^{t_2} f_s^*(t) + (1 - a_s) \sum_{t=t_1}^{t_2} f_c^*(t), \quad \text{with } t_1, t_2 \text{ arbitrary.}$$

This is readily solved to yield an explicit estimate $\hat{a}_s$ for a_s,

$$\hat{a}_s = \frac{\sum_{t=t_1}^{t_2} \left(f^*(t) - f_c^*(t) \right)}{\sum_{t=t_1}^{t_2} \left(f_s^*(t) - f_c^*(t) \right)}; \quad t^*_{\min} \le t_1 < t_2 \le t^*_{\max}. \tag{6.4.9}$$

This simple scheme lumps molecular counts into left/right sums, i.e., only uses two bins, and appeals primarily on account of the uncertainty surrounding the contaminant profile. Applied to the data shown in Fig. 6.2b, it yields the visibly improved dataset of Fig. 6.2c. Here, we used the valley and right endpoint locations as natural choices for t_1 and t_2; see also our discussion at the end of this section.

This concludes our task of decontaminating the data and reevaluating enrichment ratios. Below, we highlight certain topics that are important for algorithmic implementation and add necessary mathematical nuance to the discussion.

6.4.4.3 Stochastic Noise Models

We rooted our short discussion above in the need to match data and model but avoided specifying an indicator of "goodness of fit." Here, as before, *optimality* presupposes an understanding of *closeness* between the measured and modeled distributions which, in turn, points to some measure of *distance* between them; recall our discussion in the context of Eq. (6.2.9). One typically works with the L^p norm $\|f^*(\cdot|a_s) - f^*\|_p$ between the model $f^*(\cdot|a_s)$ (right member of Eq. (6.4.6)) and the data f^* (left member); the case $p = 2$ corresponds to the aforementioned least squares. It is conceptually important to realize that such notions of distance arise from *stochastic models* for our system and encapsulate our mechanistic understanding of it. We illustrate this point directly below by modeling the ensemble of individual molecular detection times.

In particular, we model detection times as i.i.d. r.v.s. distributed according to $f^*(\cdot|a_s^*)$, for some *true* value a_s^* subject to estimation. Equation (6.4.6) then states that the data f^* is *sampled* from $f^*(\cdot|a_s^*)$—that is, the histogram of those individual detection times approximates the actual distribution. Informally, the identity

$$f^* - f^*(\cdot|a_s) = \left(f^* - f^*(\cdot|a_s^*)\right) + \left(f^*(\cdot|a_s^*) - f^*(\cdot|a_s)\right)$$

states that the difference between the observed data and a postulated model distribution may be attributed to sampling stochasticity, for one part, and to systematic deviation for the rest. A fitting algorithm differentiates between these two error terms and aspires to remove the latter. Doing that necessitates a *model* of sampling stochasticity, which we now develop for our problem. In the framework here, the probability that $f^*(\cdot|a_s)$ generates the measured profile $g^* = M^* f^*$ is given by a multinomial distribution,

$$L(a_s|g^*) = \text{prob}(g^*|a_s) = M^*! \prod_{t=t_{\min}^*}^{t_{\max}^*} \frac{\left(f^*(t|a_s)\right)^{g^*(t)}}{g^*(t)!}. \tag{6.4.10}$$

Here also, the MLE is $\hat{a}_s = \arg\max L(a_s|g^*)$. A precise investigation of this stochastic model is not in place here, but we note that an approximation similar to Eq. (6.3.24) is possible in the regime $\min g^* \gg 1$. The log-likelihood is then approximated by a bilinear form defined by some (covariance) matrix Σ—that is, by a weighted L^2 norm; the interested reader can work out the details. This application of the CLT motivates the choice and, in fact, general applicability of least squares. The careful reader might also want to take stock of the similarity between Eq. (6.4.10) and Sect. 6.3.4 dealing with genetic mutations. There, bases performed i.i.d. trials that determined their mutated state, while here labeled molecules perform i.i.d. trials determining their detection time; both yield multinomial distributions differing only in specifics. This is how stochastic, mechanistic modeling on the microscopic level informs our macroscopic optimization problem.

6.4.4.4 Profile Alignment and Cropping

The first, simple but significant application of signal processing to our problem was the profile mollification in Fig. 6.3; that process step attenuated noise and enabled the visual identification of the contaminant. The biggest algorithmic headache, however, is not background noise but that profiles *drift* in time for reasons pertaining to GC/MS operating principles.

Drifting is already evident from control samples quantified *before* and *after* a series of GC/MS runs (*start/end controls*), in that the resulting distributions are noticeably shifted and possibly stretched or compressed—their clocks are linearly related; see [14] for a demonstration. Biological profiles processed by the machine between start/end controls also appear to follow their own clocks. Making profiles share the same clock, i.e., *aligning* them, is a (possibly non-rigid) registration problem that must be solved to fix integration windows—recall Sect. 6.4.3—and to estimate parameters through the detailed scheme (6.4.8)—we implicitly assumed there that f^*, f_s^*, and f_c^* were aligned *pointwise*.

Given the uncertainty surrounding the contaminant profile, a disproportionate investment of time and effort in solving that problem would be penny wise and pound foolish. The lumped scheme of Eq. (6.4.9) may not arise from a specifically enunciated noise model, but it is rather robust as it only demands profile matching at two time points. For our work here, we chose t_1 to be a prominent profile *feature*—the valley separating the two modes—and t_2 the right profile cut-off. This choice is informed by common sense, as opposed to mathematical proof, in that the valley/right peak locations are more robust to the presence of the contaminant than features in the left mode. As we saw, decontamination using that scheme had a remarkable effect on the dataset, cf. Fig. 6.2c.

Another significant hindrance to data fitting is the presence of *additional* profile peaks due to secondary contaminants; these are clearly seen in Fig. 6.3d–f. If such contaminants fall within the cropping interval $\mathcal{T}^*$, they can bias estimation even after profile alignment. Although an obvious cure would be to add more components to the mixture model in Eq. (6.4.6), this would necessitate expanding the integration window with the danger of including even more spurious peaks. Here, we took the opposite approach and chose to *exclude* secondary peaks from the molecular counts by *cropping* profiles to intervals *even smaller* than those dictated by the 95% AUC rule. Naturally, this entails that part of the signal is also lost; to preserve the ratio M/M^*, care must be taken to crop labeled and unlabeled profiles proportionally. This way of working was implemented in producing Fig. 6.2; its relative crudeness certainly explains part of the residual between model and data.

Finally, we reiterate here that the accuracy of our denoising is also limited by the quality of the mixture model (6.4.6) and its components. As stated repeatedly, our signal model f_s^* is practically unassailable and supported by superb post-alignment stability of the control profiles. The contaminant model f_c^*, on the other hand, is rather crude; in view of the two-bin scheme (6.4.9), one is left wondering whether modeling f_c^* as being *fully* contained in the left bin would not be an equally viable assumption. The accuracy of Eq. (6.4.6) can also be called into question, as it leaves

out various noise sources. The most obvious among them is uniform background noise, which can be modeled by a uniformly distributed mixture component; the interested mathematician can examine whether uniform noise and contaminant can be merged. A more complete model would also account for error propagation by including confidence intervals, which could then be used to weigh datasets during curve fitting. This level of detail, however, is irrelevant to the aim of this section.

6.4.5 Discussion

We saw above an entertaining application of elementary signal processing techniques to data decontamination. A key message in this work is that a modeler must be actively engaged in data collection and analysis, if not already in experimental design; recall the earlier quote on the autopsy of dead experiments. Sadly, data analysis and modeling are still performed *after* the fact much more often than before it, at least in the experience of this mathematical practitioner. Why this is so is a thorny question with no clear answer. For example, although interdisciplinary projects have substantial trouble raising funds [7], successful large-scale studies in the health sciences nowadays often include modeling components anyhow. In the author's front-line experience, and not to put too fine a point on it, the blame frequently lies with the modelers themselves, in that they fail to make their presence felt, communicate their added value, or establish a clear role. Aspiring interdisciplinary modelers should not forget that non-mathematical practitioners may be mathematically limited, if not outright semiliterate [33], for reasons that probably lie in the origins of biology as a descriptive science. It is up to us, then, to express ourselves in a language they understand; this demands a certain willingness to learn that language and, in the process, much of the content it enunciates. The converse will often not hold. This may sound asymmetric or outright unfair, but there is a reason why we examined here the application of mathematics to the health sciences and not the converse.

Our second key message is that an autopsy of a "failed" experiment can have substantial merit, particularly if data recollection is out of scope or simply impossible. The health sciences are famously conservative when it comes to accuracy and precision, to the point of producing genuinely interesting work on linear regression to this day [27, 29]. This is of course with good reason, as the enterprise they represent is superbly data-driven and deals with matters of life and death. However, where a scientist may see an experiment that must be repeated with stricter controls, a mathematician (and, inescapably, an engineer) may see a realistic situation where noise obscures information; the challenge is to separate the wheat from the chaff. This may appear precarious, but in reality there is little choice to be had. Modern societies mine incomplete, unstructured, noisy data increasingly often, e.g., in the context of IoT technologies that affect the way we collect, interpret, and react to biosignals (through wearables, labs-on-a-chip, or even self-driving cars). Such a practical spirit may be ill-fitted to absolute *understanding* but is far

more conducive to *navigating* the world we live in. As these trends play out and evolve, data processing algorithms will be increasingly applied in real time and in the open; the opportunities presented to adventurous mathematicians will intensify accordingly.

We close this chapter with a few words on the day-to-day business of developing mathematical solutions for practical problems. Earlier admonitions in this chapter focused on co-developing a *lingua franca* with practitioners and sticking to the problems on display; both indeed remain page one in the manual. However, a problem that is clear right off the bat is a rare occurrence in the real world, so much of modeling factually turns out to be problem definition. Interpreting and tackling ill-defined problems is a superbly fluid business, and a creative mathematical mind is sure to find many tangents to go off on. Earlier philippics against focusing on problems of purely mathematical interest aside, to *not* go off on a tangent is to forgo one's mathematical *identity*; this is a grave danger for mathematicians *working as such* extramurally. This is so because "mathematicians are good for ideas" [27] but, incidentally, also for much more, so they are routinely mistaken for programmers, data analysts, generalists, jacks of all trades, and magicians. It is up to modelers to establish their identity and role in the interdisciplinary projects they serve and, when necessary, to expand work boundaries with the prospect of including certain "tangents" in them. Cross-scientific mutualism is often the very saber of creation, and mathematics and applications have a long-standing tradition of feeding one another with extradisciplinary solutions or broad "tangents." (Certain topics in this chapter also came about this way—for example, existence, uniqueness and asymptotics, genetic chains, and stochastic noise models.) Scratching the mathematical itch too often may be professionally irresponsible, but not indulging it at all is criminally defeatist; this (somewhat schizophrenic) *tension* is a central theme in extramural mathematics. Although this work mode may not suit everybody, it has its own charms, valor, and value: "*The ability to juggle symbols as the pure mathematician does without regard to the immediate meaning of the symbols is but half of being a mathematician. The other half is the ability to apply the mathematics to the real world*" [18].

Appendix: A Short Primer on Parameter Estimation

The fundamental belief underpinning any modeling endeavor is that *system measurements* can be approximately generated by a specific *model*. In general terms, inference uses such measurements to mitigate uncertainty present in the underlying model. In this short appendix, we assume a well-defined *class* of candidate models that differ only in particulars; our task is to locate among them the one that *best fits* the available measurements (*data*). Here, these models share a common functional form containing finitely many parameters, so we speak of a *parametric family* and *parametric inference*. Lifting the uncertainty surrounding the parameter *values* is the inferential task par excellence.

Parameter values can be inferred in various ways joined by a common thread. Typically, unknown values are obtained as solutions to an *optimization* problem involving the model class and available data; in the problems treated here, that data is model outputs such as values of the dependent variables. For a deterministic model, a reasonable minimal requirement for an estimator would seemingly be *self-consistency*: given data generated by simulating a model with specific parameter values, a self-consistent estimator would return those precise parameter values, i.e., invert the simulation. Imposing that condition is reasonable, as long as distinct parameter values yield well-defined, distinct data (*parameter identifiability* [24]). However, the models treated in this chapter are *probabilistic*: specific parameter settings only have a certain *probability* to generate specific data. This makes the correspondence between parameter values and data both one-to-many and many-to-one, and it necessitates rethinking what can be reasonably expected from an estimator.

To address this problem, we start with univariate r.v.s $X_1, \ldots, X_N$ defined on a common sample space Ω and having distributions $f_{X_1}, \ldots, f_{X_N}$. We then write $X = (X_1, \ldots, X_N) : \Omega \to \mathscr{X}$ for the multivariate r.v. collecting them, and we recognize $\mathscr{X} \subset \mathbb{R}^N$ as the space where data resides. This *data space* is equipped with an induced joint probability distribution $f_X : \mathscr{X} \to \overline{\mathbb{R}}_+$, and each point $x = (x_1, \ldots, x_N)$ in it corresponds to a full set of system measurements. In general, this joint distribution does *not* follow trivially from the marginals $f_{X_1}, \ldots, f_{X_N}$; determining it may be a sizable part of the modeling process and a closed-form expression outside reach, if the problem does not possess additional structure. A favorable case occurs when $X_1, \ldots, X_N$ are pairwise *independent*, as f_X then has the product decomposition $f_X(x) = \prod_{n=1}^N f_{X_n}(x_n)$; another, trivial case occurs when r.v. components are algebraically constrained. Often, neither is true and modeling f_X is nontrivial. As a concrete example, the reader should derive the sampling distribution of $X = \sum_{n=1}^N X_n/N$ (sample mean) corresponding to i.i.d. Gaussian r.v.s $X_1, \ldots, X_N$.

We now assume that f_X depends on a set of parameters $\Theta = (\theta_1, \ldots, \theta_M) \in \Delta$ and write $f_{X|\Theta}(\cdot|\theta)$ to reflect this. The parameter values θ are the subject of *inference*, i.e., of *mapping data to parameter values* by means of an *estimator* $\hat{\Theta} : \mathscr{X} \to \Delta$. This function will *unambiguously* (i.e., *deterministically*) map specific data to specific parameter values without recourse to the parameter values that *generated* the data. It is in this sense that parameter estimation reverse-engineers data generation. To proceed intelligently with estimator design, we note that parameter values generate data *probabilistically*—by sampling $f_{X|\Theta}(\cdot|\theta)$—but $\hat{\Theta}$ maps these to parameter estimates *deterministically*. The combination of sampling and estimation is therefore probabilistic in nature, meaning that a *fixed* set of parameter values generates different data and thus gives rise to *various* estimates of those values. In fact, the composite map $\hat{\Theta} \circ X : \Omega \to \Delta$ is a *transformed* version of X and hence automatically an r.v. in its own right. Indeed, any measurable set U in parameter space Δ is assigned the measure of its pre-image $\hat{\Theta}^{-1}(U)$ in data space $\mathscr{X}$ which, in turn, inherits that of $X^{-1}(\hat{\Theta}^{-1}(U))$ in sample space Ω.

Being a r.v., the estimator is distributed according to some *sampling distribution* $f_{\hat{\Theta}|\Theta}$ that depends on the unknown parameters values. This observation suggests adapting the deterministic notion of self-consistency to that of an *unbiased estimator*, which amounts to demanding that

$$\int_{\Delta} \hat{\theta}\, f_{\hat{\Theta}|\Theta}(\hat{\theta}|\theta)\, \mathrm{d}\hat{\theta} \;=\; \int_{\mathscr{X}} \hat{\Theta}(x)\, f_{X|\Theta}(x|\theta)\, \mathrm{d}x \;=\; \theta, \quad \theta \in \Delta. \tag{6.4.11}$$

If this condition holds, then the *expected* parameter estimates match the true parameter values, i.e., the estimator is *correct on average* although individual estimates inevitably deviate from the truth. That deviation can be quantified (again on average) using the variance of $f_{\hat{\Theta}|\Theta}$, which one would like to keep as low as possible; note that *some* variance is inevitable, see the Cramér–Rao bound [11]. These notions of estimator bias and variance permeate estimation theory fundamentally. For example, the aforementioned variance bound links to information theory and geometry [1], whereas modern machine learning work often involves biased estimators that trade off accuracy for precision.

In our work in this chapter, we employed the *likelihood* $L(\theta|x) = f_{X|\Theta}(x|\theta)$ with which parameter values $\theta \in \Delta$ generate given data $x \in \mathscr{X}$. We specifically used the *maximum likelihood estimator* (MLE),

$$\hat{\Theta}(x) = \arg\max_{\theta} L(\theta|x) = \arg\max_{\theta} f_{X|\Theta}(x|\theta), \quad x \in \mathscr{X}. \tag{6.4.12}$$

In words, the estimate for the parameter value generating given data is the value maximizing the *probability* (likelihood) of generating that data. The evident circularity in this statement manifests that sampling and inference run contrary to each other. Note that neither existence nor uniqueness of the MLE is automatic (nor universal) and that the MLE is often biased. However, if $X_1, \ldots, X_N$ are i.i.d. and $N \to \infty$, then $f_{\hat{\Theta}|\Theta}(\cdot|\theta)$ is an *approximate Gaussian* centered at θ by the *central limit theorem* (CLT). For more detailed introductions to parameter inference at two different levels, we refer the reader to [10, 25].

Acknowledgements The work in Sect. 6.3 was initiated and supervised by Gert-Jan Boender and Thomas Hagenaars (Bacteriology and Epidemiology, Wageningen University and Research). The work in Sect. 6.4 was initiated by and done in collaboration with Rob de Boer (Theoretical Biology and Bioinformatics, Utrecht University), José Borghans (University Medical Center Utrecht), Ad Koets, and Lars Ravesloot (Bacteriology and Epidemiology, Wageningen University and Research). The author thanks them dearly for opening up a world of scientific opportunity and scholarship to him.

References

1. Amari, S., Nagaoka, H.: Methods of Information Geometry. American Mathematical Society, Providence (2000). ISBN: 0-8218-0531-2
2. Barto, A.G.: Discrete and continuous models. Int. J. Gen. Syst. **4**(3), 163–177 (1978). https://doi.org/10.1080/03081077808960681
3. Benaglia, T., Chauveau, D., Hunter, D.R., et al.: mixtools: an R package for analyzing mixture models. J. Stat. Softw. **32**(6) (2010). https://doi.org/10.18637/jss.v032.i06
4. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, Berlin (2006). ISBN: 0-387-31073-8
5. Boender, G.J., Hagenaars, T.J., Bouma, A., et al.: Risk maps for the spread of highly pathogenic avian influenza in poultry. PLoS Comput. Biol. **3**(4), 704–712 (2007). https://doi.org/10.1371/journal.pcbi.0030071
6. Box, G.E.P.: Science and statistics. J. Amer. Stat. Assoc. **71**(356), 791–799 (1976). https://doi.org/10.1080/01621459.1976.10480949
7. Bromham, L., Dinnage, R., Hua, X.: Interdisciplinary research has consistently lower funding success. Nature **534**(7609) (2016). https://doi.org/10.1038/nature18315
8. Busch, R., Neese, R.A., Awada, M., et al.: Measurement of cell proliferation by heavy water labeling. Nat. Prot. **2**(12), 3045–3057 (2007). https://doi.org/10.1038/nprot.2007.420
9. Council of the European Communities: Council directive 2005/94/ec of 20 December 2005 on community measures for the control of avian influenza and repealing directive 92/40/eec. Off. J. Eur. Union **49**, L10/16–65 (2006). ISSN: 1725-2555
10. Cox, D.R.: Principles of Statistical Inference. Cambridge University Press, Cambridge (2006). ISBN: 978-0-521-86673-6
11. Cramér, H.: Mathematical Methods of Statistics. Princeton University Press, Princeton (1946)
12. Dorado-García, A., Smid, J.H., van Pelt, W., et al.: Molecular relatedness of ESBL/AmpC-producing Escherichia coli from humans, animals, food and the environment: a pooled analysis. J. Antimicrob. Chemother. **73**(2), 339–347 (2018). https://doi.org/10.1093/jac/dkx397
13. Fisher, R.A.: Presidential address. Sankhyā Ind. J. Stat. **4**(1), 14–17 (1938)
14. GitHub repository. https://github.com/azagaris
15. Gutenkunst, R.N., Waterfall, J.J., Casey, F.P., et al.: Universally sloppy parameter sensitivities in systems biology models. PLoS Comp. Biol. **3**, 1871–1878 (2007). https://doi.org/10.1371/journal.pcbi.0030189
16. Hald, T., Wegener, H.C.: Quantitative assessment of the sources of human salmonellosis attributable to pork. In: Proceedings of the 3rd ISECSP, pp. 200–205 (1999)
17. Hald, T., Vose, D., Wegener, H.C., et al.: A Bayesian approach to quantify the contribution of animal–food sources to human salmonellosis. Risk Anal. **24**, 255–269 (2004). https://doi.org/10.1111/j.0272-4332.2004.00427.x
18. Hamming, R.W.: Toward a lean and lively calculus: report of the conference/workshop to develop curriculum and teaching methods for calculus at the college level. Am. Math. Mon. **95**(5), 466–471 (1988). https://doi.org/10.1080/00029890.1988.11972034
19. Karch, H., Denamur, E., Dobrindt, U., et al.: The enemy within us: lessons from the 2011 European Escherichia coli O104:H4 outbreak. EMBO Mol. Med. **4**, 841–848 (2012). https://doi.org/10.1002/emmm.201201662
20. Kermack, W.O., McKendrick, A.G.: A contribution to the mathematical theory of epidemics. Proc. R. Soc. A **115**, 700–721 (1927). https://doi.org/10.1098/rspa.1927.0118
21. Kimura, M.: Estimation of evolutionary distances between homologous nucleotide distances. Proc. Natl. Acad. Sci. **78**, 454–458 (1981)
22. Kullback, S., Leibler, R.A.: On information and sufficiency. Ann. Math. Stat. **22**(1), 79–86 (1951). https://doi.org/10.1214/aoms/1177729694
23. Pearl, J.: Causality: Models, Reasoning and Inference. Cambridge University Press, New York (2000). ISBN: 978-0521895606

24. Raue, A., Kreutz, C., Maiwald, T., et al.: Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. Bioinformatics **25**, 1923–1929 (2009). https://doi.org/10.1093/bioinformatics/btp358
25. Schervish, M.J.: Theory of Statistics. Springer, New York (1995). ISBN: 978-1-4612-8708-7
26. Snow, J.: On the Mode of Communication of Cholera. John Churchill, London (1855)
27. Sorg, L.: Forward-looking panel tackles issues of the Mathematics of Planet Earth. SIAM News Blog (2016)
28. Stegeman, A., Bouma, A., Elbers, A.R.W., et al.: Avian Influenza A Virus (H7N7) epidemic in The Netherlands in 2003: course of the epidemic and effectiveness of control measures. J. Infect. Dis. **190**(12), 2088–2095 (2004). https://doi.org/10.1086/425583
29. Tan, C.Y., Iglewicz, B.: Measurement-methods comparisons and linear statistical relationship. Technometrics **41**(3), 192–201 (1999). https://doi.org/10.1080/00401706.1999.10485668
30. Tufte, E.R.: Visual Explanations: Images and Quantities, Evidence and Narrative. Graphics Press, Cheshire (1997). ISBN: 978-0961392123
31. Vrisekoop, N., den Braber, I., de Boer, A.B., et al.: Sparse production but preferential incorporation of recently produced naïve T cells in the human peripheral pool. Proc. Natl. Acad. Sci. **105**(16), 6115–6120 (2008). https://doi.org/10.1073/pnas.0709713105
32. Waterfall, J.J., Casey, F.P., Gutenkunst, R.N., et al.: Sloppy-model universality class and the Vandermonde matrix. Phys. Rev. Lett. **97**, 150601 (2006). https://doi.org/10.1103/PhysRevLett.97.150601
33. Wilson, E.O.: Letters to a Young Scientists. Liveright, New York (2003). ISBN: 978-0871403858
34. Zilversmit, D.B., Entenman, C., Fishler, M.C.: On the calculation of "turnover time" and "turnover rate" from experiments involving the use of labeling agents. J. Gen. Physiol. **26**(3), 325–331 (1943)

Part III
Ecology and Evolution

Chapter 7
Multistability in Ecosystems: Concerns and Opportunities for Ecosystem Function in Variable Environments

Ehud Meron, Yair Mau, and Yuval R. Zelnik

Abstract Ecosystems are highly nonlinear dissipative systems characterized by multiplicity of stable and unstable states. Two major concerns are associated with multistable ecosystems in variable environments. The first is related to the increased likelihood of extreme climate events at regional scales, such as droughts, floods, and heat waves, that may result in abrupt transitions to malfunctioning ecosystem states. The second concern is related to the dominant role played by humans in shaping and transforming the ecology of the Earth, and to the detrimental effects that such transformations often have. Using mathematical models of dryland ecosystems as a case study, we discuss recent advances that shed new light on these concerns. We first argue that state transitions can be gradual or incomplete rather than abrupt, providing opportunities for prevention and recovery. We further argue that analyzing the unstable states that exist along with the stable ones, identifying their existence ranges and their stable and unstable manifolds, can help to devise human intervention forms that direct ecosystems towards desired functional ecosystem states, without impairing ecosystem function. We conclude by presenting open problems and delineating further research directions.

Keywords Dryland ecosystems · Vegetation patterns · Multistability · Front dynamics · Abrupt and gradual state transitions · Human intervention

E. Meron (✉)
Blaustein Institutes for Desert Research and Physics Department, Ben-Gurion University of the Negev, Beersheba, Israel
e-mail: ehud@bgu.ac.il

Y. Mau
Department of Soil and Water Sciences, Robert H. Smith Faculty of Agriculture, Food and Environment, The Hebrew University of Jerusalem, Rehovot, Israel
e-mail: yair.mau@mail.huji.ac.il

Y. R. Zelnik
Centre for Biodiversity Theory and Modelling, Theoretical and Experimental Ecology Station, CNRS, Moulis, France
e-mail: yuval.zelnik@sete.cnrs.fr

H. G. Kaper, F. S. Roberts (eds.), *Mathematics of Planet Earth*, Mathematics of Planet Earth 5, https://doi.org/10.1007/978-3-030-22044-0_7

7.1 Introduction

Ecosystems are highly nonlinear dissipative systems involving various positive feedbacks between biotic and abiotic factors [52, 60, 81]. The stabilizing effects that these feedbacks have on ecosystem states result in multiplicity of stable states in wide ranges of environmental conditions [62]. These states often include spatially periodic patterns and localized structures, in addition to spatially uniform states [59, 60]. Ecosystems, however, seldom have the time span to converge to stable asymptotic states [35]; rather, their dynamics are interrupted by natural drivers, such as droughts, fires, floods, or forest pest outbreaks, and by human intervention motivated by various functional needs, including ecosystem services, land-use changes, and restoration of degraded ecosystems.

The varying conditions that ecosystems are subjected to, natural and human driven, can induce transitions to malfunctioning states by driving ecosystems across basin boundaries, or across thresholds where stable functioning states are destabilized or disappear. These state transitions, or "regime shifts," can be abrupt [71, 72], but are not necessarily so—they can also proceed gradually through the propagation of degradation fronts as model studies predict [3, 76, 92, 93]. Abrupt transitions involving large decline in ecosystem function are of high concern because of the projections for increased climate variability at regional scales [21, 51]. This concern is reflected by an intensive current effort to devise early-warning signals for impending abrupt transitions [41, 70]. The conditions under which state transitions are expected to be gradual rather than abrupt, and thereby provide opportunities for prevention or recovery, are far less understood.

Varying conditions can also affect the multiple *unstable states* that exist along with the stable states, changing their stable and unstable manifolds or their very existence. Understanding these states, whether they are spatially uniform, periodic, or localized, is essential for studying transient ecosystem dynamics in general [35], and transient dynamics induced by human intervention in particular. Unlike natural drivers of ecosystem change, which are erratic and unpredictable, human intervention is generally planned and controlled, and yet is often detrimental to the ecosystem in question [15, 16, 66]. Studying unstable states holds much promise for devising human intervention forms that direct ecosystem dynamics towards desired self-organized functional states. This can be achieved by identifying the growing eigenmodes associated with unstable states and studying the dynamics in the phase space they span. This approach, which puts ecosystems on tracks of self-organization towards desired ecosystem states from the start, has hardly been pursued.

Out of all contexts of ecological multistability, dryland ecosystems stand out as an excellent case study for closing the knowledge gaps mentioned above. In addition to the variety of research problems that drylands pose, related to the escalating concerns about desertification and biodiversity loss [1, 17], they show striking phenomena of vegetation pattern formation (Fig. 7.1) [14, 22, 23, 86], and they are describable by mathematical models that capture remarkably well a wide

Fig. 7.1 Areal photographs of nearly periodic vegetation patterns in nature: (**a**) a spot pattern in Zambia [4], (**b**) a stripe pattern in Niger [86], (**c**) a gap ("fairy circle") pattern in Namibia (courtesy of S. Getzin). From [59]

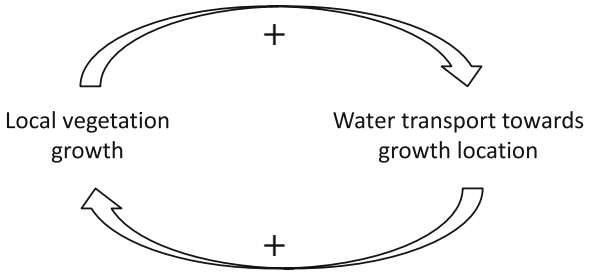


Fig. 7.2 Schematic illustration of the general positive feedback that drives vegetation pattern formation in water-limited systems. While accelerating vegetation growth in existing patches, these processes inhibit the growth in the patch surroundings, thereby favoring vegetation pattern formation. From [59]

range of observed phenomena [4, 59, 60], including multistability of uniform states, periodic patterns, and localized structures [23, 79, 92, 93].

The formation of large-scale vegetation patterns in drylands has been attributed to small-scale positive feedbacks between local vegetation growth and water transport towards the growth location, as Fig. 7.2 illustrates. Depending on the water transport mechanism, different feedbacks have been distinguished [59, 60]: (1) overland water flow induced by increased infiltration of surface water into the soil in areas of denser vegetation (infiltration feedback); (2) water conduction by laterally extended roots, the size of which increases with above-ground biomass (root-augmentation feedback); and (3) soil-water diffusion induced by strong local uptake at the vegetation-growth location and the soil-water gradients it forms (soil-water diffusion feedback). The infiltration feedback is strong in landscapes where bare soil tends to form physical or biological crusts that significantly reduce the infiltration rate relative to areas covered by vegetation [18, 23, 68]. The root-augmentation feedback is strong for plant species with high root-to-shoot ratios and laterally spread root systems [2, 24, 26]. The soil-water diffusion feedback is strong for plants with high root-to-shoot ratios and confined root systems, and for sandy soils with high hydraulic conductivities [9, 42, 93]. While these feedbacks promote local vegetation growth by drawing water from the adjacent areas of sparser vegetation, they inhibit the growth of the sparser vegetation [49, 58, 67]. This favors

nonuniform vegetation growth, or the growth of spatially periodic modes, which leads to vegetation patterns. Another pattern-forming feedback is associated with water advection, such as overland water flow on a slope [73, 75, 78] or fog advection by wind [5]. The interception of advected water by vegetation patches produces a shadowing effect on water transport in the slope or wind directions that leads to banded vegetation.

Several mathematical models have been proposed to describe vegetation pattern formation in drylands [24, 43, 48, 68, 77, 89]. These models represent a continuum modeling approach [systems of partial differential equations (PDEs)] in which the lowest level of description refers to small-scale processes rather than to individual plants, as in individual-based models [12, 13, 29]. The advantage of the continuum approach in the present context is that it lends itself to the powerful methods of pattern-formation theory. Indeed, considerable progress has been made in understanding the variety of uniform, periodic, and localized ecosystem states along the rainfall gradient, using pattern formation tools such as linear stability analysis of uniform states [26, 42, 73] and of patterned states [78, 79], derivation of amplitude (normal-form) equations [28, 87], and computation of bifurcation diagrams using numerical continuation methods [11, 76, 92, 93].

Despite the progress that has been made so far using PDE models of dryland ecosystems, many ecologically significant questions that are addressable with such models are still open or unstudied. In this paper we present and discuss open questions related to the two concerns described earlier: state transitions to malfunctioning ecosystem states and detrimental effects of human intervention.

Outline of the Chapter In Sect. 7.2, we discuss two dryland ecosystems—grasslands in western Namibia and northwestern Australia—which show striking pattern-formation phenomena and provide excellent opportunities to confront model predictions with empirical data. Since the two ecosystems feature different pattern-forming feedbacks, they are also described by different PDE models. We then use these models to address problems related to state transitions in Sect. 7.3 and to the effects of human intervention in Sect. 7.4, and describe some of the progress that has been made. We conclude by delineating directions for future research in Sect. 7.5.

7.2 The Namibian and Australian Grassland Ecosystems

Empirical testing of vegetation pattern-formation phenomena in controlled laboratory experiments is generally impractical because of the long time scales of plant growth. Remote-sensing observations provide a good alternative in fairly homogeneous and undisturbed areas, especially when the spatial scales involved are large enough to be detectable by satellite images. The availability of satellite images that go backward in time along with long-term future observations provide probes for pattern dynamics too. While vegetation pattern formation has been observed worldwide [14], two grassland ecosystems stand out in meeting the conditions of spatial homogeneity, lack of disturbances, and large spatial scales.

These are the so-called fairy circles of western Namibia [22, 40] and the recently discovered fairy circles of northwestern Australia [23]. Fairy circles are circular gaps of barren soil in grasslands that show large-scale order as Fig. 7.1c shows. The average gap diameters, 6 m in Namibia and 4 m in Australia, are large enough to be easily detectable in satellite images. The fairy circles of Namibia and of Australia show similar biomass patterns, but differ in their soil properties. In the Namibian ecosystem the soil is sandy and thus characterized by high infiltration rates of rainfall into the soil and by high hydraulic conductivities. In contrast, the top-soil layer in the Australian ecosystem is a hardly permeable claypan that generates overland water flow (runoff). As a consequence, different pattern-forming feedbacks are expected to generate the fairy-circle patterns in the two ecosystems: the soil-water diffusion feedback in Namibia and the infiltration feedback in Australia [23]. Since the plant species in both ecosystems have confined root systems, the root-augmentation feedback appears to be less significant.

In what follows, we consider the vegetation model introduced by Gilad et al. [24, 26, 60], which, unlike other models, captures all three feedbacks. The model consists of integral-partial differential equations for the areal densities of above-ground biomass $B(\mathbf{r}, t)$, soil water $W(\mathbf{r}, t)$, and overland water $H(\mathbf{r}, t)$, all in units of [kg/m^2], where $\mathbf{r} = (x, y)$ [m] represents the spatial coordinates in the plane, and t [y] represents time. Depending on the dominant feedback at work, different model simplifications can be made [60]. The confined roots in both ecosystems can be used to simplify the integral terms in the general model to algebraic terms, assuming highly localized, delta-function root kernels [42]. The resulting system of three partial differential equations (PDEs) can be further simplified in studying the Namibian ecosystem, because of the high infiltration rate of sandy soil, which prevents runoff. In that case, the system of three PDEs can be reduced to a pair of PDEs for the biomass and soil-water variables [93]. The PDEs that describe the Australian and Namibian ecosystems, in dimensional forms, are as follows:

Australian Ecosystem

$$\begin{aligned} \partial_t B &= G_B B\,(1 - B/K) - MB + D_B \nabla^2 B\,, \\ \partial_t W &= IH - L_W W - G_W W + D_W \nabla^2 W\,, \\ \partial_t H &= P - IH - L_H H - \nabla \cdot \mathbf{J}\,, \end{aligned} \tag{7.2.1}$$

where $\mathbf{J} = -2D_H H \nabla(H + Z)$ is the overland water flux, which depends on the ground topography function, $Z = Z(x, y)$, assumed to be independent of time (no erosion or deposition processes).

Namibian Ecosystem

$$\begin{aligned} \partial_t B &= G_B B\,(1 - B/K) - MB + D_B \nabla^2 B\,, \\ \partial_t W &= P - L_W W - G_W W + D_W \nabla^2 W\,. \end{aligned} \tag{7.2.2}$$

In Eqs. (7.2.1) and (7.2.2), $\nabla^2 = \partial_x^2 + \partial_y^2$ is the Laplacian in the plane, and

$$L_W = \frac{N_W}{1 + R_W B/K}, \quad L_H = \frac{N_H}{1 + R_H B/K}, \quad I = A\frac{B + Qf}{B + Q}, \tag{7.2.3}$$

$$G_B = \Lambda W(1 + EB)^2, \quad G_W = \Gamma B(1 + EB)^2, \tag{7.2.4}$$

are, respectively, the rates of soil-water evaporation, overland water evaporation, infiltration, biomass growth, and water uptake. The quantity P [mm/y] represents the precipitation rate, K [kg/m^2] represents late-growth species-specific biomass constraints, such as stem strength for woody vegetation or maximal attainable biomass in the life cycle of annuals, and E [m^2/kg] represents the root-to-shoot ratio. In obtaining Eq. (7.2.2) we assumed a flat or mildly sloped terrains that do not induce overland water flow. We note that the specific biomass dependence of G_B and G_W in Eq. (7.2.4) follows from a root architecture described by a Gaussian root kernel in the original model [60]. Other choices of root distributions can lead to different forms for G_B and G_W. Information about the remainder of the parameters and about non-dimensional forms of the model equations can be found in Refs. [23, 59, 60]. Although we refer here to two particular ecosystems involving herbaceous vegetation, the models are more general and applicable to woody vegetation as well.

Out of the three pattern-forming feedbacks, the Namibian ecosystem model captures only the soil-water diffusion feedback. The strength of this feedback is controlled by the root-to-shoot ratio E and by the soil-water diffusivity D_W; increasing any of these parameters strengthens the feedback, as it acts to increase soil-water diffusion towards vegetation patches. The Australian ecosystem model captures in addition the infiltration feedback; a strong feedback is obtained with sharp infiltration contrast $f \ll 1$ (see I in Eq. (7.2.3)) and large runoff transport coefficient D_H, as both act to speed up overland water flow. The two feedbacks suggest different spatial distributions of soil-water with respect to biomass: anti-phase distributions (maxima of biomass coincide with minima of soil water) in the case of the soil-water diffusion feedback, and in-phase distributions in the case of the infiltration feedback. A linear stability analysis of the uniform vegetation state indeed confirms these expectations [42]. In the Namibian ecosystem, where the soil-water diffusion feedback appears to be the dominant one, the distributions are expected to be anti-phase. A recent empirical study indeed supports this expectation [8]. An additional support for the soil-water diffusion feedback comes from another recent study according to which lateral water transport in the soil occurs over distances as large as 7.5 m, which is consistent with the typical length scale associated with the fairy circles [9]. In the Australian ecosystem the infiltration feedback is the dominant one, because of the claypan top layer that forms a hardly permeable soil crust. As overland water infiltrates mostly in vegetation patches, the biomass and soil-water distributions are likely to be in-phase [23]. An additional biomass-water feedback captured by the model equations for both ecosystems is

associated with reduced evaporation in vegetation patches, hereafter the "shading feedback." This is a positive but non-pattern-forming feedback because it does not involve water transport. Yet, it plays an important role in inducing multiple stable states as we discuss below.

It should be noted that an alternative explanation of the fairy-circle phenomenon has been proposed, according to which the circles represent foraging areas of termite nests [40, 83]. The termite hypothesis, however, does not explain the fairy circles of Australia, where termite nests were found to be uncorrelated to the circles [23]. The correlations that have been found between rainfall patterns and fairy-circle dynamics in Namibia [22, 93], the occurrence of Namibian fairy circles within a narrow rainfall range between the 70 and 120 mm/y isohyets [22], and observations of fairy circles with no termite colonies also in Namibia [65], pose additional challenges to the termite hypothesis.

The general Gilad et al. model [26] and its two simplified versions (7.2.1) and (7.2.2) show a universal sequence of basic vegetation states along the rainfall gradient as Fig. 7.3 illustrates [28, 50, 59, 68, 89]: bare soil, hexagonal spot pattern, stripe pattern, hexagonal gap pattern, and uniform vegetation. The emergence of gap patterns from uniform vegetation and the morphology changes that these patterns go through as rainfall decreases, first to stripe patterns and then to spot patterns, represent a population-level mechanism to cope with water stress. By self-organizing in spatial patterns the vegetation benefits not only from direct rainfall, but also from water transport towards vegetation patches from the surrounding bare-soil patches. In the Namibian ecosystem water is transported mainly by soil-water diffusion [9, 65, 93], whereas in the Australian ecosystem the transport is mainly through overland water flow [23]. With further rainfall decrease the water-contributing bare-soil areas should increase in size to compensate for the lower rainfall, which drives the two morphological changes mentioned above. Both the Namibian and Australian ecosystems show strikingly regular gap patterns (Fig. 7.1c). Statistical analyses of these patterns, including the calculation of pair-correlation functions, show a dominating hexagonal order, where each gap is surrounded on average by six equidistant gaps, as the models predict [22, 23].

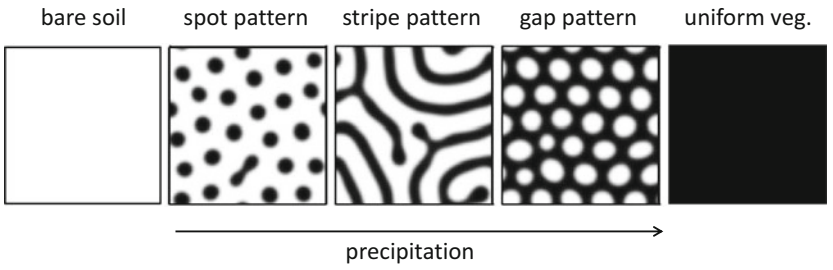


Fig. 7.3 The five basic vegetation states along the rainfall gradient as obtained by model simulations; uniform vegetation, hexagonal gap pattern, stripe pattern, hexagonal spot pattern, and bare soil. From [59]

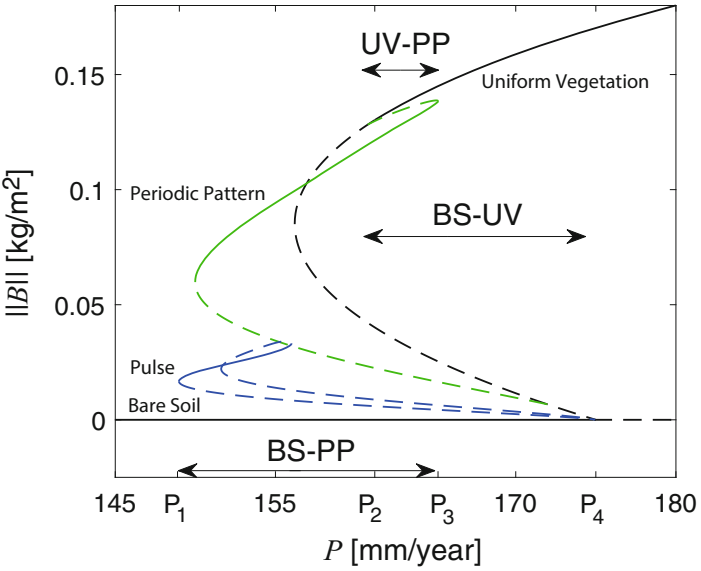


Fig. 7.4 Bifurcation diagram showing different types of bistability ranges along the rainfall gradient. The diagram shows four solutions of the Namibian ecosystem model (7.2.2) in 1d, representing bare soil, uniform vegetation, periodic vegetation pattern, and a single vegetation patch (pulse). The vertical axis is the L^2 norm of the biomass variable, while the horizontal axis represents the precipitation rate. Solid (dashed) lines represent stable (unstable) solutions. The thresholds P_4, P_3, P_2, P_1 denote, respectively, the uniform instability of bare soil, the appearance of periodic patterns in a saddle-node bifurcation, the nonuniform instability of uniform vegetation, and the disappearance of isolated vegetation patches, represented by a pulse solution, in a saddle-node bifurcation. The horizontal double arrows represent three types of bistability ranges: (1) bare soil and uniform vegetation, BS-UV ($P_2 < P < P_4$), (2) uniform vegetation and periodic patterns, UV-PP ($P_2 < P < P_3$), and (3) bare soil and periodic patterns, BS-PP ($P_1 < P < P_3$). The latter range includes periodic patterns made of weakly interacting pulses. Note that the UV-PP bistability range is, in fact, a tristability range as the bare-soil solution is also stable

The model Eqs. (7.2.1) and (7.2.2) also predict several forms of multiple stable states along the precipitation axis, associated with the positive biomass-water feedbacks that the equations capture. Figure 7.4 shows a bifurcation diagram that illustrates three types of bistability ranges. The simplest form is bistability of bare soil and (spatially) uniform vegetation (BS-UV in Fig. 7.4), which results from a uniform (zero wavenumber) imperfect pitchfork bifurcation [59, 82] of bare soil to uniform vegetation ($P = P_4$ in Fig. 7.4). Note that the negative-biomass solution is discarded as it does not represent a physical state. This bistability range can be realized with high evaporation rates in bare soil, which stabilize the bare-soil solution up to precipitation values where uniform vegetation is also a stable solution. In the Namibian ecosystem model these conditions can be realized when the soil-water evaporation is fast relative to water uptake and vegetation growth. In the model for the Australian ecosystem the condition for bare-soil stabilization can be realized with a high evaporation rate of overland water relative to the infiltration rate [23].

Such a high evaporation rate is consistent with observed surface temperatures in bare soil, which can be as high as 75 °C [23].

Besides bistability of uniform states, two main bistability forms that involve uniform and patterned states, are possible: bistability of uniform vegetation and periodic patterns—hexagonal gap patterns in two spatial dimensions (2d) (UV-PP in Fig. 7.4), and bistability of bare soil and periodic patterns—hexagonal spot patterns in 2d (BS-PP in Fig. 7.4). These bistability forms are obtainable with strong pattern-forming feedbacks that lead to subcritical nonuniform (finite-wavenumber) instabilities of uniform vegetation [59], and act to stabilize the patterned states once they are formed. In the Australian ecosystem model (7.2.1) this is the infiltration feedback (small f), while in the Namibian ecosystem model (7.2.2) it is the soil-water diffusion feedback (large E and D_W/D_B). A third bistability form involving uniform and patterned states is possible under conditions of weak shading feedback (small R_W and R_H in Eq. (7.2.3)) in addition to a strong pattern-forming feedback. In that case, the instability of bare soil to uniform vegetation is supercritical, resulting in a stability range of uniform low-biomass vegetation as Fig. 7.10 shows.

Conditions that give rise to both bistability of bare soil and uniform vegetation and bistability of uniform vegetation and periodic patterns result in a tristability range of uniform vegetation, periodic patterns, and bare soil (see Fig. 7.4). Indeed, aerial images of fairy circles in Australia reveal mixtures of nearly periodic gap patterns and large bare-soil areas, suggesting the possible stability of both states [23]. In the next section we discuss possible implications of these multistability forms to state transitions.

7.3 Abrupt vs. Gradual State Transitions

Underlying the view of regime shifts as abrupt state transitions is the presumption that these transitions are global, encompassing the whole system. This view does not take into account the spatial confinement of most disturbances, such as clear-cutting, grazing, fires, or infestation, which often induce local state transitions, rather than global transitions. The dynamics that follow local state transitions crucially depend on the transition zones that separate the two alternative stable states. These zones are fronts whose structures and dynamics have thoroughly been studied in various pattern formation contexts [10, 32, 36, 59, 63]. Depending on the dynamics of a single front, on the interactions between adjacent fronts, and on instabilities that fronts may go through, different asymptotic states can result.

We begin by discussing the simpler case of bistability of uniform states. In the context of dryland ecosystems, bistability of two uniform states can be realized in precipitation ranges where both uniform vegetation and bare soil are stable states (see Fig. 7.4). In this case fronts generically propagate. A particular control-parameter value may exist for which the front is stationary, often called the Maxwell point, but any deviation from this value results in front motion [59]. A simple example can illustrate these general results. Consider the equation

$$\partial_t u = \lambda u + \alpha u^2 - u^3 + \partial_x^2 u \,. \tag{7.3.1}$$

For $\alpha > 0$ the zero solution goes through a subcritical pitchfork bifurcation at $\lambda = 0$ that results in bistability range, $-\alpha^2/4 < \lambda < 0$, of the zero solution and the nonzero solution, $u_+ = \alpha/2 + \sqrt{(\alpha/2)^2 + \lambda} > 0$. Within this range propagating front solutions that are biasymptotic to the two states exist, e.g., $u \to 0$ as $x \to -\infty$ and $u \to u_+$ as $x \to \infty$. The front speed, c, of a propagating front is uniquely determined by the parameters λ and α [59] and can be calculated by considering constant-speed fronts. Inserting $u(x,t) = u(x - ct)$ into Eq. (7.3.1) we obtain

$$\frac{d^2u}{dz^2} + c\frac{du}{dz} - \frac{dV}{du} = 0 \,, \tag{7.3.2}$$

where $z = x - ct$ and

$$V = -\frac{\lambda}{2}u^2 - \frac{\alpha}{3}u^3 + \frac{1}{4}u^4 + V_0 \tag{7.3.3}$$

is a double-well potential with minima $V_0 = V(0)$ and $V_+ = V(u_+)$ at the zero and nonzero solutions. Multiplying Eq. (7.3.2) by du/dz and integrating we find

$$c \propto \int_{-\infty}^{\infty} \frac{dV}{du}\frac{du}{dz}dz = \int_0^{u_+} \frac{dV}{du}du = V_+ - V_0 \,. \tag{7.3.4}$$

The Maxwell point corresponds to the value $\lambda = \lambda_M = -2\alpha^2/9$ at which $V_+ = V_0$, i.e., to a stationary front. Clearly, any deviation from the Maxwell point results in wells of different depth and, consequently, in front motion.

A consequence of the generic property of front propagation is that domains of one stable state embedded in the second stable state either shrink or expand. In the course of time the fronts that bound these domains approach one another, their tails begin to overlap and the fronts interact [36]. When these interactions are attractive, as in the particular example given by Eq. (7.3.1), expanding domains coalesce into bigger ones by front annihilation. In that case, a disturbance that results in an expanding domain of a given state will eventually lead to a global transition to this state, but in a gradual manner—by front propagation. Note that such a transition can take place anywhere from the Maxwell point to the edge of the bistability range or the tipping point where an abrupt global transition occurs [3]. Global transitions of this kind are also possible with weak repulsive front interactions, but when the interactions are strong enough the fronts may come to a stop rather than annihilate [27, 32]. In the case of Eq. (7.2.2), repulsive front interactions can prevent a global transition from uniform vegetation to bare soil, as Fig. 7.5 illustrates. Repulsive interactions result from reduced competition for water in diminishing vegetation domains. When these interactions are strong enough the asymptotic state is not uniform, but rather a spatial pattern. That pattern consists of large bare-soil domains separated by vegetation stripes, and reflects a partial regime shift [59, 91].

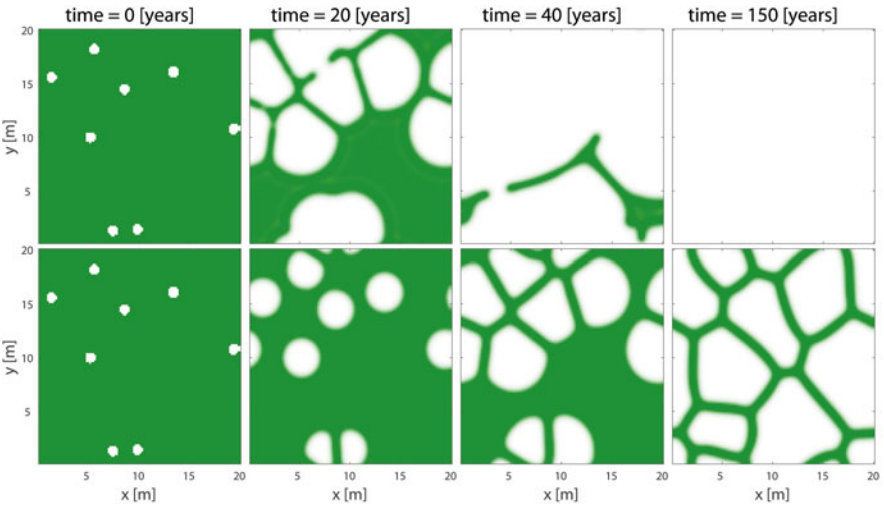


Fig. 7.5 Global vs. partial gradual regime shifts. Shown are the dynamics of a locally disturbed uniform vegetation state, obtained by solving Eq. (7.2.2) numerically in 2d at two precipitation values that are sufficiently below the Maxwell point, where small bare-soil domains expand into the surrounding vegetation areas. Top row: $P = 165$ [mm/y], far below the Maxwell point, where fast expansion of initially small bare-soil domains eventually leads to a global shift to uniform bare soil. Bottom row: $P = 170$ [mm/y], closer to the Maxwell point, where bare-soil domains expand more slowly and repulsive front interactions result in a partial regime shift to large bare-soil domains separated by narrow vegetation stripes. See Ref. [91] for additional information. From [91]

Fronts in bistable systems may go through two general types of instabilities, transverse and longitudinal, as Fig. 7.6 illustrates for an activator-inhibitor type system [59]. Transverse instabilities involve front-structure changes along the front line [27, 30], such as curvature modulations (Fig. 7.6a). By contrast, longitudinal instabilities involve changes normal to the front, e.g., a change in the position of an inhibitor front relative to an activator front. A good example is the so-called Nonequilibrium Ising-Bloch (NIB) bifurcation [7, 32]. In a bistable system with an inversion symmetry, this is a pitchfork front bifurcation in which a stationary (Ising) front is destabilized to a pair of counter-propagating (Bloch) fronts [7, 31, 32, 39, 61] (Fig. 7.6b). Although front instabilities are local processes, occurring in the confined front zone, their influence usually extends to the entire system. A transverse instability results in the growth of fingers that split at their tips into new fingers, which grow and tip-split again until the entire system is filled up with a stationary labyrinthine pattern (Fig. 7.6a) [27, 30]. Note that this process requires repulsive front interactions to prevent the coalescence of adjacent fingers into larger domains. A longitudinal instability, such as the NIB bifurcation, can result in counter-propagating front segments that develop in the course of time into space-filling spiral waves [31, 33, 53]. In both types of front instability the asymptotic state is a spatial pattern, either stationary or time dependent, rather than an alternative uniform state.

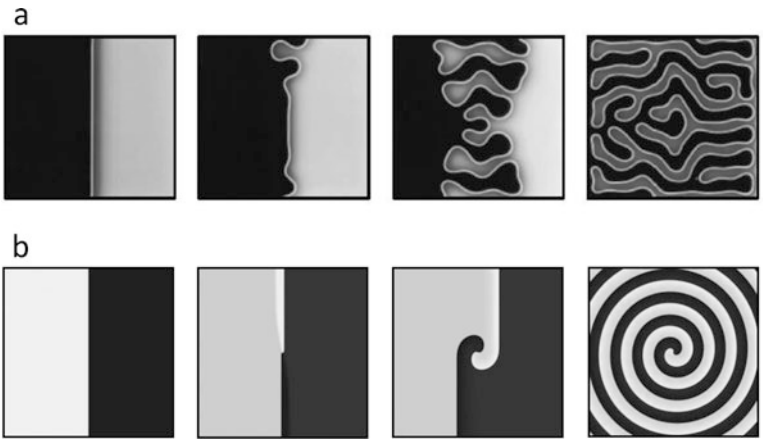


Fig. 7.6 Local front instabilities can lead to global patterns. Shown are snapshots of numerical simulations of a FitzHugh–Nagumo model (time proceeds from left to right) that illustrate transverse (**a**) and longitudinal (**b**) front instabilities, and the asymptotic patterns they lead to—labyrinthine pattern and spiral wave, respectively. Further details in Ref. [30]. From [59]

Front instabilities may well be found in models of dryland vegetation as these are also activator-inhibitor systems showing bistability of uniform states (UV and BS), where biomass is the activator and lack of soil-water—the inhibitor [44]. Front dynamics in bistability of uniform vegetation and bare soil have received little attention so far [20, 76]. Further studies are needed to test the relevance of incomplete regime shifts, driven by repulsive front interactions and front instabilities, to real ecosystems, such as the Namibian and the Australian ecosystems discussed in Sect. 7.2.

More attention has been devoted to the two bistability forms of uniform and patterned states: uniform vegetation and periodic gap patterns, and periodic spot patterns and bare soil [6, 11, 74, 79, 80, 92–95]. According to pattern-formation theory, and in contrast to bistability of uniform states, when one of the alternative states is a periodic pattern, fronts can be stationary or pinned in a *range* of the control parameter [64]. In this range alternative-state domains can remain fixed in size, neither expanding nor retracting, forming a multitude of stable hybrid states. The latter can be spatially localized, representing single alternative-state domains of different sizes, or spatially extended, corresponding to various combinations of localized domains. In a bifurcation diagram, such as that shown in Fig. 7.7 for the Namibian ecosystem model, localized hybrid states appear as solution branches that snake back and forth as the sizes of the domains they represent change, a behavior that has been termed "homoclinic snaking" [45, 46]. The snaking solution branches occupy a subrange of the bistability range—the snaking range. Thus, three front types can be distinguished in a bistability range of uniform and patterned states: a stationary pinned front within the snaking range and two fronts moving in opposite directions on either side of this range. Local disturbances within a snaking range

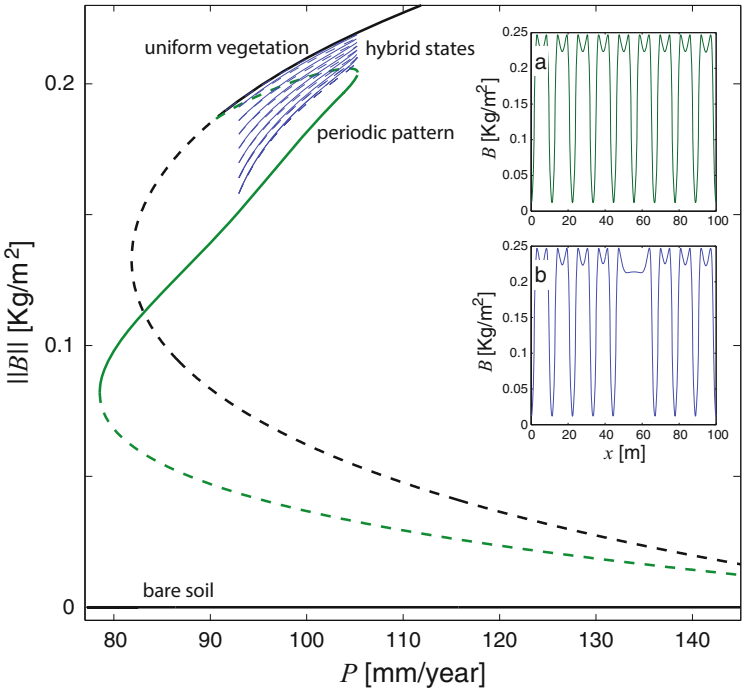


Fig. 7.7 Bifurcation diagram for the Namibian ecosystem model (7.2.2) in 1d. The vertical axis is the L^2 norm of the biomass variable. Solid (dashed) lines represent stable (unstable) solutions. The diagram shows a precipitation (P) range where both uniform vegetation and periodic vegetation pattern are stable. Within this range there exists a subrange of hybrid states. The insets show spatial profiles of a 1d periodic gap pattern (**a**) and of a hybrid state consisting of a periodic gap pattern with a missing gap, i.e., with the smallest domain of the alternative uniform vegetation state (**b**). From [59]

should have little effect as the fronts are stationary and initial alternative-state domains quickly converge to nearby hybrid states. Local disturbances outside the snaking range, but still inside the bistability range, result in gradual shifts. Gradual shifts may also occur within the snaking range when the system is subjected to environmental fluctuations. Such fluctuations, if strong enough, can kick the system temporarily outside the snaking range, where fronts do propagate, and thereby induce hybrid-state transitions that gradually shift the system towards the alternative stable state [3, 93]. The wider the snaking range the more resilient the system is to environmental fluctuations and local disturbances. Identifying the biotic and abiotic parameters that control the width of the snaking range relative to the bistability range is therefore a highly significant unstudied problem.

Hybrid states are likely to exist in the Namibian ecosystem, as the satellite images in Fig. 7.8 suggest [93]. Finding empirical evidence for hybrid-state transitions and gradual regime shifts is more intricate. The closest evidence comes from studies of fairy-circle "birth" and "death" events [85], which have been interpreted recently

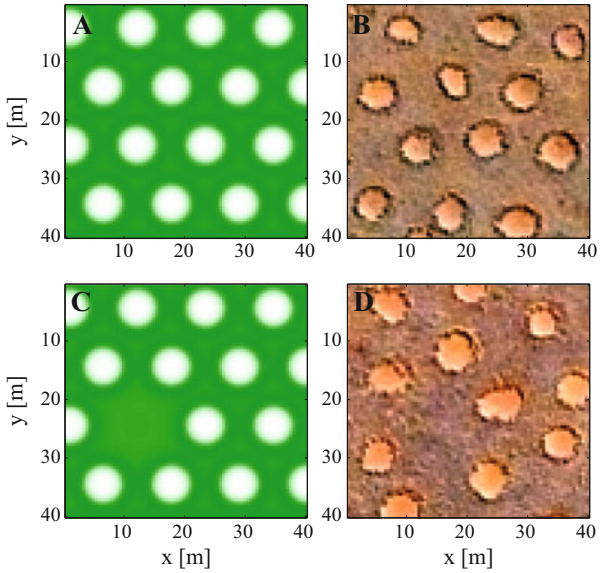


Fig. 7.8 Patterned states in the Namibian ecosystem. Hexagonal gap pattern obtained by integrating the Namibian ecosystem model (7.2.2) in 2d (**a**), and a satellite image showing a similar nearly hexagonal pattern of fairy circles in western Namibia (**b**). A hybrid state consisting of one missing gap obtained from the model (**c**), and a similar hybrid state obtained from a satellite image (**d**). Adapted from [93]

as hybrid-state transitions induced by rainfall variability [93]. Figure 7.9a shows satellite images that span a period of 10 years and demonstrate the birth of a fairy circle after a severe drought in 2007. Figure 7.9b shows a simulation of the Namibian ecosystem model, using the first satellite image in 2004 as an initial condition with a precipitation value within the snaking range, and mimicking the 2007 drought by a precipitation downshift that takes the system outside the snaking range for a period of 1 year. As the simulation snapshots show, a new gap has appeared after a 10-year period, exactly at the same location where the actual fairy circle has appeared. The simulated temporal escape from the snaking range that was needed to induce the formation of the new gap supports the view of fairy-circle birth events as hybrid-state transitions. A series of droughts can result in a cascade of hybrid-state transitions and a gradual shift [93], but empirical evidence for such a cascade has not been reported yet.

Homoclinic snaking can also be found in a bistability range of low-biomass uniform vegetation and periodic spot pattern, as Fig. 7.10 shows [11], implying the feasibility of hybrid-state transitions and gradual shifts in fluctuating environments. However, when the periodic-pattern solution branch extends to the stability range of the bare-soil state, homoclinic snaking breaks down in what appears to be a Belyakov–Devaney transition [38]. In that case shifts from periodic patterns to bare soil, or desertification, are found to be abrupt [95]. While most model studies predict

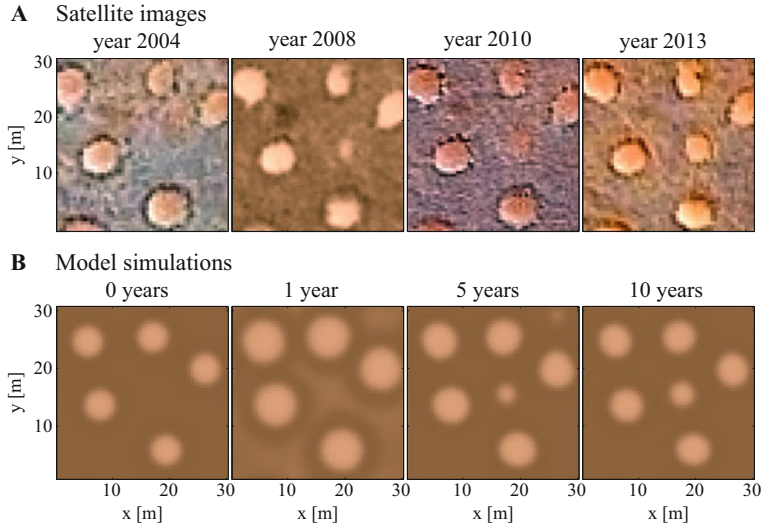


Fig. 7.9 Birth of fairy circles as a hybrid-state transition. (**a**) Satellite images showing the birth of a new fairy circle. (**b**) Snapshots of model simulations [Eq. (7.2.2)], using initial conditions derived from the 2004 satellite image (first snapshot from left) and mimicking a 1-year drought (second snapshot from left), which show similar fairy-circle birth dynamics. Adapted from [93]

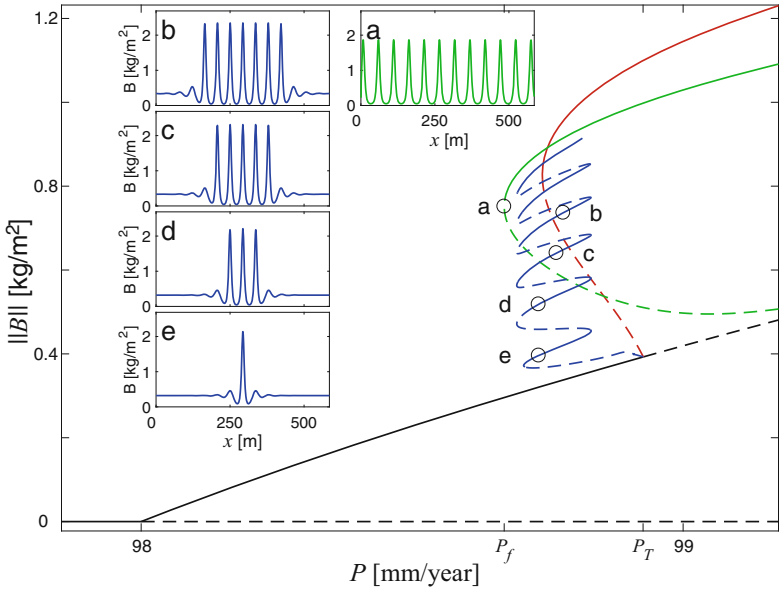


Fig. 7.10 Bifurcation diagram for the Australian ecosystem model (7.2.1) in 1d with weak shading feedback (small R_W and R_H in (7.2.3)). The diagram shows a small bistability range, $P_f < P < P_T$, of low-biomass uniform vegetation (black line) and a periodic pattern (green line) and a subrange of homoclinic snaking. The inset (**a**) shows the periodic solution while the insets (**b–e**) show localized pattern solutions or hybrid states

wide bistability ranges of spot patterns and bare soil, and thus the likelihood of abrupt desertification, these studies have been confined to a single species. Quite often dryland landscapes consist of woody and herbaceous species, forming a bistability range of woody spot patterns and uniform herbaceous vegetation [25]. Studies of two-species models do predict homoclinic snaking [47], suggesting that the degradation of woody spot patterns may be gradual rather than abrupt.

In the tristability range of bare soil, periodic patterns, and uniform vegetation [see Fig. 7.7] many front types are expected to coexist, pinned, or moving; fronts separating domains of uniform vegetation and bare soil, domains of uniform vegetation and periodic patterns, and domains of periodic patterns and bare soil. The dynamics, interactions, and stability properties of these front solutions, and the implications for regime shifts have hardly been studied [96].

7.4 Human Intervention Along Unstable Eigenmodes

The dominant role played by humans in shaping and transforming the ecology of the Earth is well recognized [19]. More than three-quarters of the terrestrial biosphere have already been transformed into anthropogenic biomes by human populations and this trend is intensifying. A major question that arises in this regard is how to intervene in ecosystem dynamics so as to achieve the intervention goal without harming ecosystem function. While this question appears to be overwhelmingly hard in many contexts of human intervention, it may be tractable for selected contexts that are simple enough to be modeled mathematically and yet ecologically significant. In the following we focus on a specific example of such a context, vegetation restoration in fluctuating environments, and use it to illustrate a general approach to human intervention that highlights the roles of unstable states.

A common restoration practice is water harvesting by spatially periodic ground modulations, such as parallel micro-catchments, that capture overland water flow and along which vegetation is planted [88]. This is a spatial resonance problem where a system that tends to form a periodic pattern with a preferred wave-number k_0 is forced to follow an external template with a different wave-number k_f. The question we wish to address here is the following: given a stripe-like template of ground modulations in the x direction, characterized by a wave-vector $\mathbf{k}_f = (k_f, 0)$, what should be the vegetation-planting pattern in order to achieve the restoration goal of establishing a bio-productive state that remains functional in a fluctuating environment?

We address this question using Eq. (7.2.1), modified to include a periodically modulated infiltration rate to mimic periodic soil-crust removal, a lighter and more cost-effective intervention form than micro-catchments [56],

$$I(B) = A\frac{B + Qf}{B + Q}, \quad f = f_0\left[1 + \frac{\gamma_f}{2}\left(1 + \cos\left(k_f x\right)\right)\right], \tag{7.4.1}$$

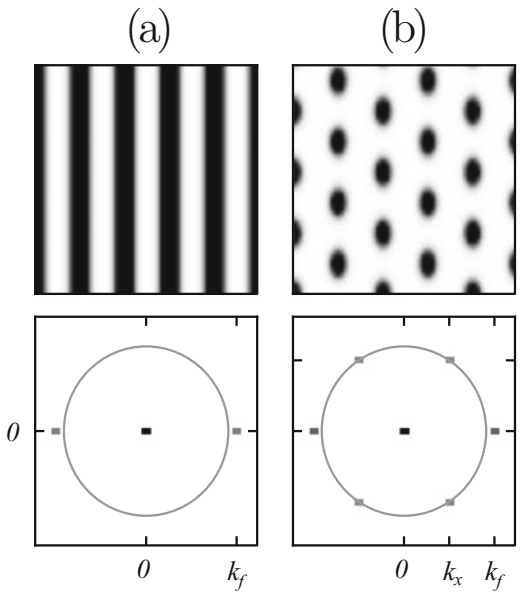


Fig. 7.11 Resonant responses to stripe-like ground modulations with wave-vector $\mathbf{k}_f = (k_f, 0)$. Top panels: a resonant stripe pattern (**a**) and a resonant rhombic pattern (**b**) in the x (horizontal) and y plane, obtained by numerical integration of Eqs. (7.2.1) and (7.4.1). Dark shades denote high biomass. Bottom panels: the corresponding Fourier transforms (in absolute value), where darker dots denote higher absolute values, and the circle $|\mathbf{k}| = k_0$. The peaks at $\pm\mathbf{k}_f$ and the absence of peaks on the circle of radius k_0 in (**a**) indicate that the stripe pattern is in 1:1 resonance with the forcing. The four peaks on the circle of radius k_0 in (**b**) represent the two oblique modes, $\mathbf{k}_\mp = (-k_x, \mp k_y)$, and their complex conjugates, $-\mathbf{k}_\mp$. The value $k_x = k_f/2$ indicates that the rhombic pattern is 2:1 resonance with the forcing. From [56]

where $f_0 \ll 1$ is the infiltration contrast of an unmodulated soil and γ_f represents the modulation strength. According to this form the infiltration rate in a densely vegetated area is high, $I \approx A$, because the biomass density there is significantly higher than Q, a species-dependent reference value representing an over 50% increase of the infiltration rate, whereas in bare soil it is much lower, $I = Af_0$ in unmodulated bare soil and $I = (1 + \gamma_f)Af_0$ in bare soil with removed crust. Figure 7.11 shows two types of resonant patterns and the absolute values of their Fourier transforms: (a) a stripe pattern that locks to the forcing in a 1:1 resonance (vegetation stripe at each ground modulation), (b) a rhombic pattern that locks to the forcing in a 2:1 resonance (vegetation spot at every second ground modulation). The Fourier transforms show the basic modes that constitute these patterns; a stripe mode, $\mathbf{k} = \mathbf{k}_f = (k_f, 0)$, in the case of a stripe pattern (and its conjugate mode $(-k_f, 0)$), and three modes in the case of a rhombic pattern, a stripe mode, $\mathbf{k}_f = (k_f, 0)$, and two oblique modes, $\mathbf{k}_\pm = (k_x, \pm k_y)$, where $k_x = k_f/2$ and k_y are such that the total wave-number k is equal to the preferred wave-number k_0, i.e., satisfies $k^2 = k_x^2 + k_y^2 = k_0^2$. Note that the three wave-vectors $\mathbf{k}_f$ and $\mathbf{k}_\pm$ satisfy the

resonance relation, $\mathbf{k}_f + \mathbf{k}_+ + \mathbf{k}_- = 0$, which drives the simultaneous growth of the three modes, and that in the case of an exact resonance ($k_f = k_0$) the rhombic pattern becomes a hexagonal pattern consisting of three wave-vectors 120° apart. Numerical studies of Eqs. (7.2.1) and (7.4.1) reveal a bistability range of resonant stripe and rhombic patterns and that rhombic patterns extend to lower precipitation values than stripe patterns [56].

Both stripe and rhombic patterns represent productive states, and establishing any one of them would satisfy the restoration goal. The larger area of vegetation coverage in the case of stripe patterns does not necessarily imply higher total biomass or productivity, because of the larger water-contributing areas in the case of rhombic patterns, which results in higher biomass densities in the vegetation patches. The remaining question is which of the two patterns is more resilient to droughts, and thus better functioning in fluctuating environments? Figure 7.12 shows the response of resonant stripe patterns to a moderate precipitation downshift, which results in convergence to a rhombic pattern (top row), and to a stronger downshift, which results in collapse to bare soil (bottom row) despite the existence of stable rhombic patterns. The same numerical experiment conducted with an initial rhombic pattern results in no significant pattern change. These results suggest that stripe patterns are less resilient to droughts than rhombic patterns.

In order to understand the mechanism of collapse to bare soil let us focus on the amplitudes A, a_+, a_- of the stripe and the two oblique modes, respectively, in terms of which the state variables $U = (B, W, H)$ can be approximated as

$$U(\mathbf{x}, t) \approx U_0 + U_1\, A\, e^{ik_f x} + U_2\, a_+\, e^{i\mathbf{k}_- \cdot \mathbf{r}} + U_3\, a_-\, e^{i\mathbf{k}_+ \cdot \mathbf{r}} + \text{c.c.}\,, \tag{7.4.2}$$

where $U_0,\ U_1,\ U_2,\ U_3$ are constant vectors, and we assumed proximity to the bare-soil instability and weak ground modulations. Equations for the amplitudes $A, a_+,$

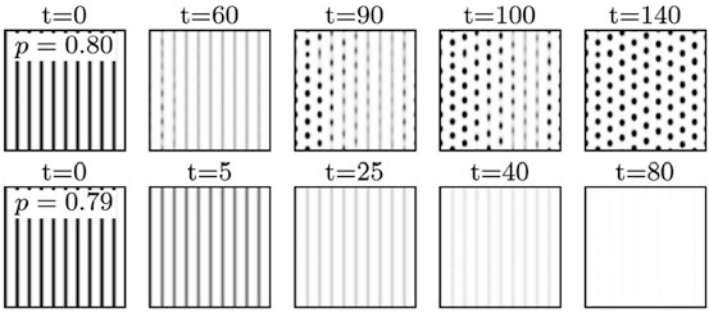


Fig. 7.12 Resilience of stripe patterns to droughts. Numerical simulations of Eqs. (7.2.1) and (7.4.1) showing the response of a resonant stripe pattern, obtained at a precipitation value within the bistability range of stripe and rhombic patterns, to precipitation downshifts of different strength to a range where rhombic patterns are still stable. The response to a moderate downshift results in quick convergence to a rhombic pattern (top row), while the response to a stronger downshift results in collapse to bare soil. Adapted from [56]

a_- have been derived for a simple pattern-formation model, the Swift–Hohenberg equation with parametric spatial forcing [54–56]. Derivation of amplitude equations for the vegetation model (7.2.1) and (7.4.1) is harder and has not been done yet. However, because of the universal character of amplitude equations we may expect their general form to apply to the restoration problem as well. Indeed, the amplitude equations can produce a bifurcation diagram similar to that found by numerical integration of the vegetation model [56], interpreting the bifurcation parameter as the precipitation rate P. In the following we use these amplitude equations to study the response to precipitation downshifts. A schematic form of the diagram obtained from the amplitude equations is shown in Fig. 7.13.

Consider two precipitation downshifts of different strengths, applied to stable stripe patterns as the arrows in Fig. 7.13 indicate: a moderate downshift to P_2 where stripe solutions exist but are unstable (green arrow), and a stronger downshift to P_1 where stripe solutions do not exist (red arrow). Note that both downshifts take the system to a precipitation range where rhombic patterns are still stable solutions and significantly far from the saddle-node bifurcation at which they disappear. Figure 7.14 shows the phase planes spanned by the amplitude moduli $\rho_S = |A|$ and $\rho_R = |a_+| = |a_-|$ at P_2 and P_1, where we took advantage of the symmetry between the two oblique modes in the precipitation range we consider. Shown in

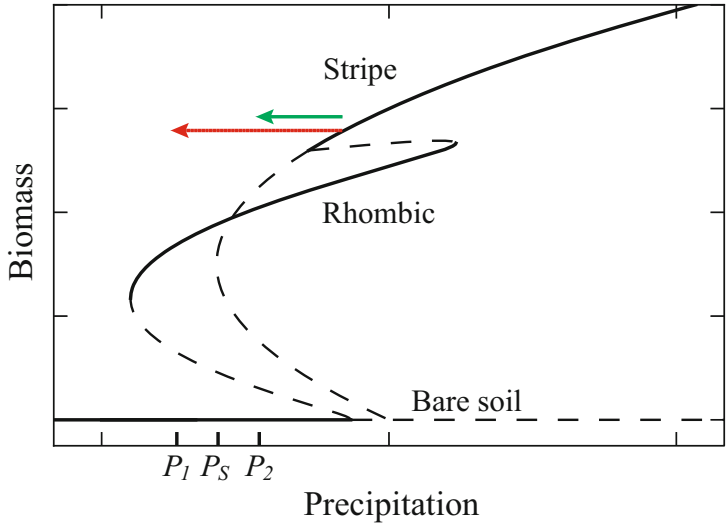


Fig. 7.13 A schematic bifurcation diagram for vegetation restoration. The solution branches describe bare soil, stripe pattern, and rhombic pattern, where solid (dashed) lines denote stable (unstable) solutions. The vertical axis represents the L^2 norm of the biomass expressed in terms of the modes' amplitudes, $\sqrt{|A|^2 + |a_+|^2 + |a_-|^2}$. The precipitation value P_S denotes the disappearance of unstable stripe solutions in a saddle-node bifurcation. The green and red arrows represent precipitation downshifts from the stability range of stripe patterns to precipitation values $P_2 > P_S$ and $P_1 < P_S$, respectively. Adapted from [56]

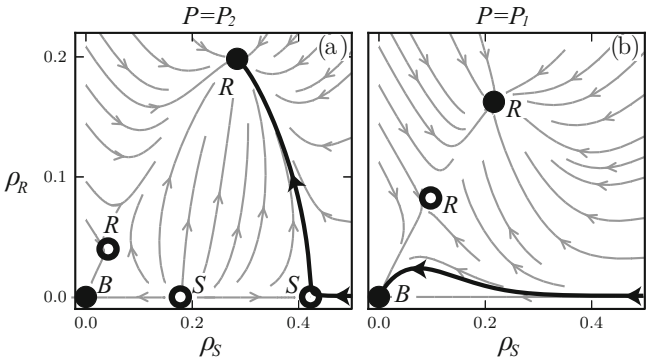


Fig. 7.14 Phase-space dynamics in the plan spanned by $\rho_S = |B|$ and $\rho_R = |a_\pm|$ at (**a**) $P = P_2$, where a pair of unstable stripe solutions exists, (**b**) at $P = P_1$, where the unstable stripe solutions no longer exist (see Fig. 7.13). The solid (hollow) circles denote stable (unstable) stationary states. The labels B, S, and R denote the bare-soil state, stripe patterns, and rhombic patterns, respectively, and the arrows denote the vector field of the amplitude equations. The responses of a stable resonant stripe pattern to precipitation downshifts are shown by the thick black phase portraits: (**a**) A moderate shift to a range where unstable stripe solutions still exist, results in a smooth transition to a rhombic pattern. (**b**) A stronger shift to a range where stripe solutions no longer exist, results in a collapse to bare soil. Adapted from [56]

these phase planes are the stationary uniform and patterned states (fixed points) that exist at the respective precipitation value, and their stability properties. Also shown in Fig. 7.14 are phase trajectories (black lines) of numerical solutions of the amplitude equations, starting with stripe solutions that were computed at a higher P within their range of stability. A moderate downshift to P_2 results in a smooth transition to a rhombic pattern as Fig. 7.14a shows. The unstable large-amplitude stripe solution plays a crucial role in this response; its unstable manifold, which represents the growth of the two oblique eigenmodes, acts as a barrier for the flow in phase space and prevents convergence to the stable bare-soil solution. By contrast, a stronger downshift to P_1 results in collapse to bare soil, as the stripe solution and its unstable manifold no longer exists to constrain the phase-space flow.

This analysis shows that the common and intuitive restoration practice in a 1:1 stripe pattern, where the planting pattern coincides with the ground modulation pattern, can result in a productive state but suffers from poor resilience to precipitation downshifts (droughts). By contrast, restoration in a rhombic pattern, which initiates the growth of the oblique modes, results in a productive and resilient state. More generally, these results suggest to disentangle the planting pattern from the ground modulation pattern and determine the former by identifying the growing (unstable) eigenmodes and analyzing the phase space they span. By focusing on the dynamical constraints that unstable states impose through their stable and unstable manifolds, judicious choices of planting patterns that result in functional ecosystem states can be made.

The phase-space information may also be used in managing ecosystems that have already been restored. The resilience of restored stripes can possibly be improved by spatial periodic biomass modulations, obtained by trimming or by managed grazing. This is because of the unstable rhombic solution whose stable manifold separates the basins of attraction of the bare-soil state and the stable rhombic pattern. Periodic biomass modulations will place the system in the attraction basin of the rhombic pattern by creating projections along the oblique eigenmodes.

These considerations can be generalized and applied to other contexts of human intervention in ecosystem dynamics, besides restoration, such as range management, regime-shift control, agroecology, and others. While they appear to rely on the availability of faithful mathematical models, empirical data analysis may prove to be a possible alternative when mathematical models are absent, such as the extraction of eigenmodes and phase-space elements from spatial Fourier transforms of satellite images. We note that the growing eigenmodes need not be spatially extended like the stripe and oblique modes in the restoration example. Contexts that involve localized structures, such as fronts in gradual regime shifts, can give rise to localized eigenmodes associated with translation symmetry [59] and possible front instabilities [34, 37, 91].

7.5 Conclusion

While pattern-formation phenomena in dryland ecosystems have been the subject of many theoretical and empirical studies [4, 13, 14, 57–59, 67, 84], many fewer studies have addressed the implications of pattern formation to ecosystem function in variable and disturbed environments [60], where state transitions may take place [41, 69], and in ecosystems subjected to human intervention. In Sect. 7.3 we considered several cases of bistable ecosystems, distinguishing between bistability of two uniform states and bistability of uniform and patterned states. In each case we discussed the implications of front dynamics to transitions from functional ecosystems states to less functional or dysfunctional states, emphasizing three aspects of front dynamics: single-front motion, front interactions, and front instabilities. The latter two aspects have received little attention even in the simplest bistability case of two uniform states [20]. The availability of fairly realistic models that are still simple enough to be mathematically tractable, such as the Namibian ecosystem model, should motivate additional studies. Because of the disparate length scales associated with biomass fronts (short) and water fronts (long) singular-perturbation methods may prove very useful in such studies [30, 36]. The relevance of homoclinic snaking in bistability ranges of uniform and patterned states to dryland vegetation has been demonstrated in several studies [11, 93, 94], including two-species models [47], but the physical and ecological factors that affect front pinning and determine the size of the homoclinic snaking range have remained unexplored.

Another intriguing and unstudied question is related to the similarity of pattern-formation phenomena in bistability ranges of uniform and patterned states and

bistability ranges of two uniform states. An example of such a phenomenon is a single gap of bare soil in otherwise uniform vegetation. Such a gap can be realized in a bistability range of uniform vegetation and periodic pattern as a hybrid state, but it can also be realized in a bistability range of uniform vegetation and bare soil as an outcome of repulsive front interactions. The capacity to determine which mechanism is at work in a given realization of a single gap is highly significant. In the former case rainfall fluctuations may drive the system outside the snaking range and induce a gradual shift to the less productive periodic gap pattern, which amounts to moderate desertification. In the latter case rainfall fluctuations may drive the system below the Maxwell point and induce a gradual shift to the unproductive bare-soil state, which amounts to severe desertification.

Although the significance of unstable states in ecosystem dynamics has already been stressed [35], the roles these states may play in planning human intervention have remained largely unexplored. An example of a significant problem that can be studied using an approach similar to that described for the restoration of degraded landscapes is range management in drought-prone ecosystems. Consider, for example, managing grazing in uniform grasslands. The disappearance of unstable uniform-vegetation solutions at low precipitation rates may induce collapse to bare soil rather than convergence to a periodic pattern, very much like the disappearance of unstable stripe solutions in the restoration problem. This suggests the management of grazing in spatial patterns, in order to locate the system in the basin of attraction of the periodic solution. Such management may result not only in the achievement of an ecosystem service—feeding livestock—but also in improved resilience to droughts.

Finally, the difficulty to conduct controlled laboratory experiments calls for the development of advanced data-analysis methods for remote sensing observations, geared to test model predictions of structural and dynamical fairy-circle characteristics, such as large-scale hexagonal order disrupted by penta-hepta defects [63] and hybrid-state transitions [22, 90, 93].

Acknowledgements Some of the results described here have been reported in earlier publications with additional colleagues, including Golan Bel, Stephan Getzin, Aric Hagberg, Lev Haim, Omer Tzuk, and Hezi Yizhaq. We gratefully acknowledge their contributions. The research leading to the results described in this chapter received funding from the Israel Science Foundation Grant 305/13.

References

1. Adeel, Z., Safriel, U., Niemeijer, D., et al.: Ecosystems and human well-being: Desertification synthesis. Technical Report of the Millennium Ecosystem Assessment, World Resources Institute, Washington, D.C. (2005)
2. Barbier, N., Couteron, P., Lefever, R., et al.: Spatial decoupling of facilitation and competition at the origin of gapped vegetation patterns. Ecology **89**, 1521–1531 (2008)
3. Bel, G., Hagberg, A., Meron, E.: Gradual regime shifts in spatially extended ecosystems. Theor. Ecol. **5**, 591–604 (2012)

4. Borgogno, F., D'Odorico, P., Laio, F., et al.: Mathematical models of vegetation pattern formation in ecohydrology. Rev. Geophys. **47**, RG1005 (2009)
5. Borthagaray, A.I., Fuentes, M.A., Marquet, P.A.: Vegetation pattern formation in a fog-dependent ecosystem. J. Theor. Biol. **265**(1), 18–26 (2010)
6. Chen, Y., Kolokolnikov, T., Tzou, J., et al.: Patterned vegetation, tipping points, and the rate of climate change. Eur. J. Appl. Math. **26**, 945–958 (2015)
7. Coullet, P., Lega, J., Houchmanzadeh, B., et al.: Breaking chirality in nonequilibrium system. Phys. Rev. Lett. **65**, 1352 (1990)
8. Cramer, M.D., Barger, N.N.: Are Namibian fairy circles the consequence of self-organizing spatial vegetation patterning? PloS One **8**(8), e70,876 (2013)
9. Cramer, M.D., Barger, N.N., Tschinkel, W.R.: Edaphic properties enable facilitative and competitive interactions resulting in fairy circle formation. Ecography **40**, 1210–1220 (2017)
10. Cross, M.C., Greenside, H.: Pattern Formation and Dynamics in Nonequilibrium Systems. Cambridge University Press, Cambridge (2009)
11. Dawes, J.H.P., Williams, J.L.M.: Localised pattern formation in a model for dryland vegetation. J. Math. Biol. **73**, 1–28 (2015)
12. DeAngelis, D.L., Gross, L.J. (eds.): Individual-Based Models and Approaches on Ecology: Concepts and Models. Chapman and Hall, New York (1992)
13. DeAngelis, D.L., Yurek, S.: Spatially explicit modeling in ecology: A review. Ecosystems **20**(2), 284–300 (2017). https://doi.org/10.1007/s10021-016-0066-z
14. Deblauwe, V., Barbier, N., Couteron, P., et al.: The global biogeography of semi-arid periodic vegetation patterns. Glob. Ecol. Biogeogr. **17**, 715–723 (2008)
15. Dirzo, R., Young, H.S., Galetti, M., et al.: Defaunation in the anthropocene. Science **345**, 401–406 (2014)
16. D'Odorico, P., Bhattachan, A., Davis, K.F., et al.: Global desertification: Drivers and feedbacks. Adv. Water Resour. **51**, 326–344 (2013)
17. Duraiappah, A.K., Naeem, S.: Ecosystems and human well-being: biodiversity synthesis. Technical Report of the Millennium Ecosystem Assessment, World Resources Institute, Washington, DC. (2005)
18. Eldridge, D.J., Zaady, E., Shachak, M.: Infiltration through three contrasting biological soil crusts in patterned landscapes in the Negev, Israel. J. Stat. Phys. **148**, 723–739 (2012)
19. Ellis, E.C.: Ecology in an anthropogenic biosphere. Ecol. Monogr. **85**, 287–331 (2015)
20. Fernandez-Oto, C., Tlidi, M., Escaff, D., et al.: Strong interaction between plants induces circular barren patches: fairy circles. Phil. Trans. R. Soc. A **372**(2027), 20140009 (2014)
21. Field, C.B., Barros, V., Stocker, T.F., et al.: Managing the risks of extreme events and disasters to advance climate change adaptation: a special report of the Intergovernmental Panel on Climate Change. Technical Report, Cambridge University Press, Cambridge, UK, and New York, NY (2013)
22. Getzin, S., Wiegand, K., Wiegand, T., et al.: Adopting a spatially explicit perspective to study the mysterious fairy circles of Namibia. Ecography **38**, 1–11 (2015)
23. Getzin, S., Yizhaq, H., Bell, B., et al.: Discovery of fairy circles in Australia supports self-organization theory. Proc. Natl. Acad. Sci. **113**(13), 3551–3556 (2016)
24. Gilad, E., Von Hardenberg, J., Provenzale, A., et al.: Ecosystem engineers: from pattern formation to habitat creation. Phys. Rev. Lett. **93**, 098105 (2004)
25. Gilad, E., Shachak, M., Meron, E.: Dynamics and spatial organization of plant communities in water limited systems. Theor. Popul. Biol. **72**, 214–230 (2007)
26. Gilad, E., Von Hardenberg, J., Provenzale, A., et al.: A mathematical model for plants as ecosystem engineers. J. Theor. Biol. **244**, 680 (2007)
27. Goldstein, R.E., Muraki, D.J., Petrich, D.M.: Interface proliferation and the growth of labyrinths in a reaction-diffusion system. Phys. Rev. E **53**, 3933–3957 (1996)
28. Gowda, K., Riecke, H., Silber, M.: Transitions between patterned states in vegetation models for semiarid ecosystems. Phys. Rev. E **89**, 022,701 (2014)
29. Grimm, V., Railsback, S.F.: Individual-based Modeling and Ecology. Princeton University Press, Princeton (2005)

30. Hagberg, A., Meron, E.: Complex patterns in reaction diffusion systems: a tale of two front instabilities. Chaos **4**, 477–484 (1994)
31. Hagberg, A., Meron, E.: From labyrinthine patterns to spiral turbulence. Phys. Rev. Lett. **72**, 2494–2497 (1994)
32. Hagberg, A., Meron, E.: Pattern formation in non-gradient reaction diffusion systems: the effects of front bifurcations. Nonlinearity **7**, 805–835 (1994)
33. Hagberg, A., Meron, E.: The dynamics of curved fronts: beyond geometry. Phys. Rev. Lett. **78**, 1166–1169 (1997)
34. Hagberg, A., Meron, E., Rubinstein, I., et al.: Controlling domain patterns far from equilibrium. Phys. Rev. Lett. **76**, 427–430 (1996)
35. Hastings, A.: The key to long-term ecological understanding? Trends Ecol. Evol. **19**, 39–45 (2004)
36. van Heijster, P., Doelman, A., Kaper, T.J., et al.: Front interactions in a three-component system. SIAM J. Appl. Dyn. Syst. **9**(2), 292–332 (2010)
37. Hilker, F.M., Lewis, M.A., Seno, H., et al.: Pathogens can slow down or reverse invasion fronts of their hosts. Biol. Invasions **7**(5), 817–832 (2005)
38. Homburg, A.J., Sandstede, B.: Homoclinic and heteroclinic bifurcations in vector fields. In: Handbook of Dynamical Systems, vol. 3, pp. 379–524. Elsevier, Amsterdam (2010)
39. Ikeda, H., Mimura, M., Nishiura, Y.: Global bifurcation phenomena of travelling wave solutions for some bistable reaction-diffusion systems. Nonlinear Anal. Theory Methods Appl. **13**, 507–526 (1989)
40. Juergens, N.: The biological underpinnings of Namib Desert fairy circles. Science **339**(6127), 1618–1621 (2013)
41. Kéfi, S., Vishwesha, G., Brock, W.A.: Early warning signals of ecological transitions: methods for spatial patterns. Plos One **9**, e92097 (2014)
42. Kinast, S., Zelnik, Y.R., Bel, G., et al.: Interplay between turing mechanisms can increase pattern diversity. Phys. Rev. Lett. **112**, 078701 (2014)
43. Klausmeier, C.A.: Regular and irregular patterns in semiarid vegetation. Science **284**, 1826–1828 (1999)
44. Kletter, A.Y., von Hardenberg, J., Meron, E.: Ostwald ripening in dryland vegetation. Commun. Pure Appl. Anal. **11**, 261–273 (2012)
45. Knobloch, E.: Spatially localized structures in dissipative systems: open problems. Nonlinearity **21**, T45 (2008)
46. Knobloch, E.: Spatial localization in dissipative systems. Ann. Rev. Condens. Matter Phys. **6**(1), 325–359 (2015)
47. Kyriazopoulos, P., Jonathan, N., Meron, E.: Species coexistence by front pinning. Ecol. Complex. **20**, 271–281 (2014)
48. Lefever, R., Lejeune, O.: On the origin of tiger bush. Bull. Math. Biol. **59**, 263–294 (1997)
49. Lejeune, O., Couteron, P., Lefever, R.: Short range co-operativity competing with long range inhibition explains vegetation patterns. Acta Oecol. **20**(3), 171–183 (1999)
50. Lejeune, O., Tlidi, M., Lefever, R.: Vegetation spots and stripes: dissipative structures in arid landscapes. Int. J. Quantum Chem. **98**, 261–271 (2004)
51. Maestre, F.T., Eldridge, D.J., Soliveres, S., et al.: Structure and functioning of dryland ecosystems in a changing world. Annu. Rev. Ecol. Evol. Syst. **47**(1), 215–237 (2016)
52. Marten, G.G.: Human Ecology - Basic Concepts for Sustainable Development. Earthscan Publications, London (2001)
53. Marts, B., Hagberg, A., Meron, E., et al.: Bloch-front turbulence in a periodically forced Belousov-Zhabotinsky reaction. Phys. Rev. Lett. **93**(108305), 1–4 (2004)
54. Mau, Y., Hagberg, A., Meron, E.: Spatial periodic forcing can displace patterns it is intended to control. Phys. Rev. Lett. **109**, 034102 (2012)
55. Mau, Y., Haim, L., Hagberg, A., et al.: Competing resonances in spatially forced pattern-forming systems. Phys. Rev. E **88**, 032,917 (2013)
56. Mau, Y., Haim, L., Meron, E.: Reversing desertification as a spatial resonance problem. Phys. Rev. E **91**, 012,903 (2015)

57. Meron, E.: Modeling dryland landscapes. Math. Model. Nat. Phenom. **6**, 163–187 (2011)
58. Meron, E.: Pattern-formation approach to modelling spatially extended ecosystems. Ecol. Model. **234**, 70–82 (2012)
59. Meron, E.: Nonlinear Physics of Ecosystems. CRC Press, Taylor & Francis Group, Boca Raton (2015)
60. Meron, E.: Pattern formation – a missing link in the study of ecosystem response to environmental changes. Math. Biosci. **271**, 1–18 (2016)
61. Mimura, M., Tohma, M.: Dynamic coexistence in a three-species competition–diffusion system. Ecol. Complex. **21**, 215–232 (2015)
62. Petraitis, P.: Multiple Stable States in Natural Ecosystems. Oxford University Press, Oxford (2013)
63. Pismen, L.: Patterns and Interfaces in Dissipative Dynamics. Springer Series in Synergetics. Springer, Berlin (2006)
64. Pomeau, Y.: Front motion, metastability and subcritical bifurcations in hydrodynamics. Phys. D **23**, 3 (1986)
65. Ravi, S., Wang, L., Kaseke, K.F., et al.: Ecohydrological interactions within "fairy circles" in the Namib Desert: revisiting the self-organization hypothesis. J. Geophys. Res. Biogeosci. **122**(2), 405–414 (2017)
66. Reynolds, J.F., Smith, D.M.S., Lambin, E.F., et al.: Global desertification: building a science for dryland development. Science **316**(5826), 847–851 (2007)
67. Rietkerk, M., van de Koppel, J.: Regular pattern formation in real ecosystems. Trends Ecol. Evol. **23**(3), 169–175 (2008)
68. Rietkerk, M., Boerlijst, M.C., van Langevelde, F., et al.: Self-organization of vegetation in arid ecosystems. Am. Nat. **160**, 524–530 (2002)
69. Rietkerk, M., Dekker, S.C., de Ruiter, P.C., van de Koppel, J.: Self-organized patchiness and catastrophic shifts in ecosystems. Science **305**, 1926–1929 (2004)
70. Scheffer, M., Bascompte, J., Brock, W.A., et al.: Early-warning signals for critical transitions. Nature **461**, 387–393 (2009)
71. Scheffer, M., Carpenter, S.R.: Catastrophic regime shifts in ecosystems: linking theory to observation. Trends Ecol. Evol. **18**, 648–656 (2003)
72. Scheffer, M., Carpenter, S., Foley, J.A., et al.: Catastrophic shifts in ecosystems. Nature **413**, 591–596 (2001)
73. Sherratt, J.A.: An analysis of vegetation stripe formation in semi-arid landscapes. J. Math. Biol. **51**(2), 183–197 (2005). https://doi.org/10.1007/s00285-005-0319-5
74. Sherratt, J.A.: Pattern solutions of the Klausmeier model for banded vegetation in semiarid environments, V: the transition from patterns to desert. SIAM J. Appl. Math. **73**, 1347–1367 (2013)
75. Sherratt, J.A.: When does colonisation of a semi-arid hillslope generate vegetation patterns? J. Math. Biol. **73**, 199–226 (2016)
76. Sherratt, J.A., Synodinos, A.D.: Vegetation patterns and desertification waves in semi-arid environments: mathematical models based on local facilitation in plants. Discrete Contin. Dynam. Systems B **17**(8), 2815–2827 (2012)
77. Shnerb, N.M., Sarah, P., Lavee, H., et al.: Reactive glass and vegetation patterns. Phys. Rev. Lett. **90**, 0381011 (2003)
78. Siero, E., Doelman, A., Eppinga, M.B., et al.: Striped pattern selection by advective reaction-diffusion systems: resilience of banded vegetation on slopes. Chaos **25**(3), 036411 (2015)
79. Siteur, K., Siero, E., Eppinga, M.B., et al.: Beyond Turing: The response of patterned ecosystems to environmental change. Ecol. Complex. **20**(0), 81–96 (2014)
80. Siteur, K., Eppinga, M.B., Doelman, A., et al.: Ecosystems off track: rate-induced critical transitions in ecological models. Oikos **125**, 1689–1699 (2016)
81. Stone, L., Weisburd, R.S.J.: Positive feedback in aquatic ecosystems. Trends Ecol. Evol. **7**, 263–267 (2016)
82. Strogatz, S.H.: Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering. Westview Press, Boulder (2001)

83. Tarnita, C., Bonachela, J.A., Sheffer, E., et al.: A theoretical foundation for multi-scale regular vegetation patterns. Nature **541**, 398–401 (2017)
84. Tongway, D.J., Valentin, C., Seghieri, J., et al. (eds.): Banded Vegetation Patterning in Arid and Semiarid Environments: Ecological Processes and Consequences for Management. Ecological Studies, vol. 149. Springer, Basel (2001)
85. Tschinkel, W.: The life cycle and life span of Namibian fairy circles. PloS One **7**(6), e38056 (2012)
86. Valentine, C., d'Herbes, J., Poesen, J.: Soil and water components of banded vegetation patterns. Catena **37**, 1–24 (1999)
87. van der Stelt, S., Doelman, A., Hek, G.M., et al.: Rise and fall of periodic patterns for a generalized Klausmeier–Gray–Scott model. J. Nonlinear Sci. **23**, 39–95 (2013)
88. Vohland, K., Barry, B.: A review of in situ rainwater harvesting (RWH) practices modifying landscape functions in African drylands. Agric. Ecosyst. Environ. **131**, 119–127 (2009)
89. von Hardenberg, J., Meron, E., Shachak, M., et al.: Diversity of vegetation patterns and desertification. Phys. Rev. Lett. **89**, 198101 (2001)
90. Wiegand, T., Kissling, W.D., Cipriotti, P.A., et al.: Extending point pattern analysis for objects of finite size and irregular shape. J. Ecol. **94**(4), 825–837 (2006)
91. Zelnik, Y.R., Meron, E.: Regime shifts by front dynamics. Ecol. Indic. **94**, 544–552 (2018). https://doi.org/10.1016/j.ecolind.2017.10.068
92. Zelnik, Y.R., Kinast, S., Yizhaq, H., et al.: Regime shifts in models of dryland vegetation. Philos. Trans. R. Soc. A **371**, 20120358 (2013)
93. Zelnik, Y.R., Meron, E., Bel, G.: Gradual regime shifts in fairy circles. Proc. Natl. Acad. Sci. **112**, 12,327–12,331 (2015)
94. Zelnik, Y.R., Meron, E., Bel, G.: Localized states qualitatively change the response of ecosystems to varying conditions and local disturbances. Ecol. Complex. **25**, 26–34 (2016)
95. Zelnik, Y.R., Uecker, H., Feudel, U., et al.: Desertification by front propagation? J. Theor. Biol. **418**, 27–35 (2017)
96. Zelnik, Y.R., Gandhi, P., Knobloch, E., et al.: Implications of tristability in pattern-forming ecosystems. Chaos Interdiscip. J. Nonlinear Sci. **28**(3), 033609 (2018)

Chapter 8
Measurement of Biodiversity: Richness and Evenness

Fred S. Roberts

Abstract Evidence about the health of ecosystems is often thought to be related to biodiversity. Traditional attempts to define biodiversity consider two components: richness—the number of species in the ecosystem—and evenness—the extent to which species are evenly distributed. This chapter studies attempts to make both concepts precise using mathematical approaches. It describes a number of evenness indices that have been widely used, studies axioms for evenness that an index could be required to satisfy, and explores which evenness indices satisfy those axioms. The chapter also considers evenness indices that "preserve" certain partial orders. The relationship between richness and evenness and attempts to derive measures of biodiversity based on both richness and evenness are explored.

Keywords Axiomatic approach · Biodiversity · Ecosystem · Evenness · Index · Measure · Partial order · Richness

8.1 Introduction

The planet is constantly changing, but the pace of change has accelerated in recent decades. Construction and deforestation change habitats. Fishing, hunting, and poaching affect the population of many species. Fossil fuel combustion leads to increasing atmospheric greenhouse gas concentrations and, in turn, changing climates that affect the ability of existing species to survive in an ecosystem while opening up the same ecosystem for new, sometimes competing, species. Commerce and transport introduce nonnative species that can impact the population of existing species. All of this has led to concerns about the health of the planet and in particular the health of important ecosystems. We need evidence about the health

F. S. Roberts (✉)
DIMACS Center, Rutgers University, Piscataway, NJ, USA
e-mail: froberts@dimacs.rutgers.edu

H. G. Kaper, F. S. Roberts (eds.), *Mathematics of Planet Earth*, Mathematics of Planet Earth 5, https://doi.org/10.1007/978-3-030-22044-0_8

of ecosystems in order to make better decisions about them and about the activities that affect them.

Evidence about the health of ecosystems is often obtained by measuring "biodiversity." An index of biodiversity allows us to set specific goals and measure progress toward them. This chapter will describe approaches to measuring biodiversity and, in particular, axiomatic approaches leading to measures or indices of biodiversity.

The 1992 Convention on Biological Diversity [49] set the goal: *By 2010, achieve a significant reduction of the current state of biodiversity loss at the global, regional, and national level* [50]. But how can we tell if we have achieved this goal? We need to be able to measure biodiversity. There is a long history of attempts to develop indices of biodiversity, raising many mathematical challenges. We will discuss some of these approaches and challenges in this chapter.

But what is biodiversity? There is a long history of trying to define it, and it is a multidimensional concept. Components of biodiversity include species diversity, genetic diversity within species, ecosystem diversity, and ecosystem services and processes. Tracing back to the work of Whittaker [51], many ecologists distinguish among alpha, beta, and gamma diversity, defined, respectively, as the diversity of a given site in a region, the difference in diversity among sites in the region, and the diversity of the region as a whole. We return to this distinction below.

The Convention on Biodiversity defines biodiversity as *the variability among living organisms from all sources including, inter alia, terrestrial, marine and other aquatic ecosystems, and the ecological complexes of which they are part; this includes diversity within species, between species and of ecosystems* [49].

The term "biodiversity" was coined by Walter Rosen during a 1986 National Forum on BioDiversity [45] and was first used in the literature in the proceedings of that meeting [53]. Since then, hundreds of papers have attempted to define it precisely. The literature is based on ideas from other disciplines that go back more than 100 years. Over the years, many indices have been proposed. However, by way of warning, there is a remarkable inconsistency in the literature of measurement of biodiversity. The same term has been used for different indices. Some papers use a measure and others its negative or its reciprocal. Some "normalize" it, so it is a measure between 0 and 1. A lower number means lower biodiversity in some indices, higher biodiversity in others. (For example, 0 could be lowest or highest.) Unfortunately, there are some papers that use names for indices without saying which version of the index they are using.

Consider a toy example. One ecosystem has six butterflies, one grasshopper and one beetle. (Of course, there are many species of butterflies, grasshoppers, beetles, etc., but we disregard that.) A second ecosystem has four butterflies and four grasshoppers. Which has more biodiversity? The first has more species. The second has its species more evenly distributed in terms of numbers. Traditional approaches consider two basic determinants of biodiversity: *richness*, or the number of species in an ecosystem; and *evenness*, or the extent to which species are equally distributed [25]. We will discuss measurement of each of these concepts and how to combine them into one measure.

However, are richness and evenness reasonable determinants of biodiversity, and are they the only ones? Among other things, these concepts assume that all species are equal, all individuals are equal (disregarding differences in size, health, etc.), and that spatial distribution is irrelevant. But are these assumptions enough to give us useful indices of biodiversity? Do we really want an ecosystem with as many leopards as zebras? Is a forest with 100 hemlock trees and 100 oak trees, well interspersed, equally as diverse as a forest with 100 hemlocks in one half and 100 oaks in another half? Moreover, some species are highly "visible" or considered centrally important for the purposes of conservation biology (e.g., lions, elephants). Other species are indicator species of the health of an ecosystem, and we may want to give their presence (or absence) higher priority. For instance, lichens respond to changes in forest structure resulting from changes in air quality or climate. Disappearance of lichens may indicate environmental stress (high levels of sulfur dioxide, nitrogen oxides, etc.). Similarly, algal species in aquatic systems may indicate organic pollution and nutrient loading (e.g., nitrogen, phosphorus) and mussels are sensitive to siltation and low dissolved oxygen in water.

Outline of the Chapter In Sect. 8.2, we discuss several factors that come into play when one tries to define a measure of the *richness* of an ecosystem. We suggest that, when details are ignored, the number of species is a good measure of richness. In Sect. 8.3, we address the second measure of biodiversity, namely *evenness*. Here, we introduce Simpson's index, the Coefficient of Variation, the Shannon–Wiener index, Pielou index, and the Gini index, with their mathematical definitions. In Sect. 8.4, we present several axioms originally formulated in an economics context by Dalton and show that the five indices discussed in Sect. 8.3 all satisfy Dalton's axioms. In Sect. 8.5, we list additional axioms that could be applied to select specific measures of the evenness of ecosystems. In the following sections we address the issue of how to compare the evenness of different ecosystems using the concept of *partial order*. In Sect. 8.6, we introduce the Lorenz partial order and the generalized Lorenz partial order and show that the Gini index and the Coefficient of Variation both reflect the general Lorenz partial order, while the other three indices (Simpson, Shannon—Wiener, and Pielou) do not. Section 8.7 highlights the (tenuous) relationship between the measures of richness and evenness. Section 8.8 is devoted to attempts to combine richness and evenness in a single measure of biodiversity through the development of a sequence of partial orders. The choice of a particular partial order from the sequence can be tested through a set of axioms, which are discussed in Sect. 8.9. In the final Sect. 8.10, we summarize our discussion and present some ideas for further investigation.

8.2 Measuring Richness

Richness S is usually interpreted as the number of different species in an ecosystem. This has some major drawbacks. It disregards the presence or absence of "important" or "indicator" species. It may depend on the sampling process to detect species,

and that sampling process could be biased, could depend on the length of time sampling is done, the intensity of the sampling process, and the size of the area sampled [4, 16, 43]. Richness defined this way also increases with the presence of species we don't want to have (e.g., invasive species) [21, 26].

Some of these problems with richness have been made precise through various mathematical models. Consider the case of the connection between time spent sampling and number of species detected. Soberon and Llorente [43] note that there is evidence that, as time spent collecting increases, the number of species identified asymptotically approaches some limit. They investigate different assumptions about the probability of detecting a new species in a given time period, given the number of species that have been detected so far.

Let $P(j, t, \Delta t)$ be the probability that one new species is detected during a time interval Δt after time t, given that the search has already identified j species at time t. Soberon and Llorente assume that P grows linearly with Δt,

$$P(j, t, \Delta t) = F(j, t)\Delta t. \tag{8.2.1}$$

The simplest assumption about $F(j, t)$ is that it is constant over time and varies linearly with the number of species already found,

$$F(j, t) = a - bj. \tag{8.2.2}$$

That is, as the species list grows, the probability of identifying a new species in the interval Δt decreases proportionally to the size of the list. This model may be appropriate in a small area or with a well-known group of species. It leads to a differential equation involving the probability $p(j, t)$ that at time t the list has exactly j species. The solution gives the expected richness $S(t)$ at time t,

$$S(t) = (a/b)[1 - e^{-bt)}]. \tag{8.2.3}$$

This solution does exhibit the asymptotic property.

Soberon and Llorente also study an exponential model,

$$F(j, t) = ae^{-bj}. \tag{8.2.4}$$

Here, as the species list grows, the probability of identifying a new species in the time interval Δt decreases exponentially with the size of the list. This assumption may be reasonable in a large area or with species relatively unknown and where the probability of finding a new species never reaches zero. The solution to the resulting differential equation depends on assumptions about the probability distribution $p(j, t)$,

Lamas et al. [20] studied the collection of butterflies at the Pakitza Biological Station at Parque Nacional Manu in Peru. They fitted the data to the linear and exponential models and a third model. All fits were good. The largest asymptote was 905 species, but they extrapolated to different numbers of species found.

The conclusion is that the answer to the question whether a linear, exponential, or other model is most suitable may depend upon the collecting experience or procedure—for example, whether the probability of finding a new species becomes dramatically more and more difficult over time.

Another line of modeling work seeks to connect the richness S to the area A being sampled. One model of the relationship is the power law,

$$S = kA^c, \tag{8.2.5}$$

where k and c are constant. The parameter k is called the species richness factor and the parameter c the species accumulation factor. This model goes back to Arrhenius in 1921 [1], is widely used, and has a large amount of theoretical and empirical support [32, 33, 44]. One downside of this model is that the intensity of sampling is a factor in the richness measured, not just the area. If you sample more intensively, clearly you should find more species.

In spite of the number of species found being connected to area sampled, sampling method, and other factors, the number of species is still considered a good measure of one part of biodiversity, namely richness.

As noted earlier, ecologists distinguish among alpha, beta, and gamma diversity. In terms of richness, alpha diversity can be interpreted as the number of species in a given site in a region, gamma diversity as the number of species in the entire region, and beta diversity between two sites as the number of species in one but not the other. However, there are many more subtle issues involved in making these concepts precise. For instance, beta diversity could be percentage similarity among various sites in a region, ratio of gamma diversity to average alpha diversity, etc. For a discussion and a variety of definitions, see [46, 47].

8.3 Measuring Evenness

Consider two toy ecosystems. The first has two snakes, two lizards, two turtles, two salamanders, two toads, two birds, two beetles, and two frogs. The second has one snake, one lizard, one turtle, one salamander, one toad, one bird, one beetle, and eight frogs. They have the same number of species, so are equally rich. The first has as even a distribution as possible, while the second is highly uneven. Clearly, richness and evenness are two different concepts.

Measures of evenness in ecology are frequently based on ideas going back in the economic literature to the early 1900s, specifically to the work of Gini [13, 14] on measure of even income or wealth distribution and to the work of Dalton [7] on measures of inequality. Other measures of biodiversity or of evenness go back to work in communication theory, in particular the work of Claude Shannon [39] on entropy in information theory. These ideas are predated in statistical mechanics by Boltzmann's work on entropy in the nineteenth century. The notion of evenness is also applied in other areas. For example, it is used to study the extent to which

we have achieved a stable degree of social or economic equity [11, 12], to study scientific collaboration [31], to study interdisciplinarity of journals [22], and many other topics.

The literature has many proposed measures of evenness. The papers [42, 48] provide surveys of popular evenness indices. Before discussing a few examples, we introduce some notation.

- S, the number of species;
- x_i, the *abundance* of species i—that is, the number of individuals or, in some cases, some measure of biomass of species i in the ecosystem;
- $\mathbf{x} = (x_1, x_2, \ldots, x_S)$, the *abundance vector*;
- a_i, species i's proportion of the population, $a_i = x_i / \sum_j x_j$;
- $\mathbf{a} = (a_1, a_2, \ldots, a_S)$;
- $f(\mathbf{x}) = f(x_1, x_2, \ldots, x_S)$, the *evenness* of the ecosystem.

Unless otherwise indicated, sums are taken over all species ($i = 1, \ldots, S$). We adopt the convention that $f(\mathbf{x})$ is low if the ecosystem is very even, high if it is very uneven. It is common in the literature to take $f(\mathbf{x})$ to be between 0 and 1.

In the example given at the beginning of this section, the first population has abundance vector $\mathbf{x} = (2, 2, 2, 2, 2, 2, 2)$ and the second has vector $\mathbf{x} = (1, 1, 1, 1, 1, 1, 8)$. Also, the first population has $\mathbf{a} = (1/7, 1/7, \ldots, 1/7)$, while the second has $\mathbf{a} = (1/14, 1/14, \ldots, 1/14, 8/14)$.

8.3.1 Simpson's Index

A well-known index of evenness is *Simpson's index* [41], given by

$$\text{Simpson's index, } \lambda(\mathbf{x}) = \sum_i {a_i}^2. \tag{8.3.1}$$

This is the probability that any two individuals drawn at random from an infinite population will belong to the same species. In our example if $f(\mathbf{x}) = \lambda(\mathbf{x})$, then for the first population, $f(\mathbf{x}) = (1/7)^2 + (1/7)^2 + \ldots + (1/7)^2 = 7/49 = 1/7 = 0.143$. For the second population, $f(\mathbf{x}) = (1/14)^2 + (1/14)^2 + \ldots + (1/14)^2 + (8/14)^2 = 0.357$. Some biologists prefer high evenness to mean more even, and so use $1 - \lambda(\mathbf{x})$ or $1/\lambda(\mathbf{x})$ instead of $\lambda(\mathbf{x})$.

8.3.2 Coefficient of Variation

Another index of evenness is the *Coefficient of Variation*, given by

$$\text{Coefficient of Variation, } V(\mathbf{x}) = \sigma/\mu, \tag{8.3.2}$$

where μ is the mean and σ the standard deviation,

$$\mu = (1/S)\sum_i x_i, \quad \sigma^2 = (1/S)\sum_i (x_i - \mu)^2. \tag{8.3.3}$$

If $f(\mathbf{x}) = V(\mathbf{x})$, $x_i = \mu$ for all i, so $f(\mathbf{x}) = 0$. For any population without a perfectly even distribution, $f(\mathbf{x}) > 0$. For example, when $\mathbf{x} = (1, 1, 1, 1, 1, 1, 8)$, $\mu = 2$, $\sigma = \sqrt{6} = 2.449$, and $V(\mathbf{x}) = 1.225$.

8.3.3 Shannon–Wiener Diversity Index

A third index of evenness, coming out of information theory, is the *Shannon–Wiener Diversity index* or, as it is called in information theory, the *Shannon Entropy*,

$$\text{Shannon–Wiener Diversity (Shannon Entropy) index, } H'(\mathbf{x}) = -\sum_i a_i \ln a_i. \tag{8.3.4}$$

This index quantifies (in expected value) the information contained in a message, in units such as bits. A fair coin has entropy of one bit. If a coin is unfair and you are asked to bet, you will have less uncertainty. The Shannon–Wiener index is maximized if each x_i is the same. We use $-H'$ so that the index will be minimized if each x_i is the same. For example, when $\mathbf{x} = (2, 2, 2, 2, 2, 2, 2)$, $-H'(\mathbf{x}) = -1.946$, while if $\mathbf{x} = (1, 1, 1, 1, 1, 1, 8)$, $-H'(\mathbf{x}) = -1.451$.

8.3.4 Pielou Index

An index derived from the Shannon–Weiner index is the *Pielou index* [29, 30], named after Canadian statistical ecologist Evelyn Chrystalla "E.C." Pielou.

$$\text{Pielou index, } J('\mathbf{x}) = H'/H'_{\max} = H'/\ln S, \tag{8.3.5}$$

where H' is the Shannon–Wiener Index and $H'_{\max}$ is the maximum value that H' attains, i.e., $\ln S$, which occurs when all x_i are equal. We will use $-J'$, so that the more even distribution of population gets the lower number. When $\mathbf{x} = (2, 2, 2, 2, 2, 2, 2)$, $-J'(\mathbf{x}) = -0.278$, while if $\mathbf{x} = (1, 1, 1, 1, 1, 1, 8)$, $-J'(\mathbf{x}) = -0.207$.

8.3.5 Gini Index

The *Gini index* was introduced in econometrics by Italian statistician Corrado Gini in 1909 [13] and 1912 [14]. It is defined as follows. Start with counting all of the absolute differences among all the counts x_i,

$$M = \sum_i \sum_j |x_i - x_j|. \tag{8.3.6}$$

The average M/S^2 is called the *mean absolute difference*. Next, normalize by dividing by the average value of x_i, i.e., by

$$A = \frac{1}{S} \sum_i x_i. \tag{8.3.7}$$

The ratio M/A is called the *relative mean absolute difference* and gives the *Gini index*,

$$\text{Gini index, } G'(\mathbf{x}) = M/(S \sum_i x_i). \tag{8.3.8}$$

The Gini index is easiest to calculate if the x_i are ordered. If we order from high to low, we get $M = \sum (S + 1 - 2i)x_i$. Dividing by $S \sum x_i$ gives the following formula for the *Gini Index*:

$$\text{Gini index, } G'(\mathbf{x}) = \frac{S+1}{S} - \frac{2}{S} \sum_i i a_i. \tag{8.3.9}$$

If the x_i are ordered from low to high, we get the negative of this expression,

$$G'(\mathbf{x}) = \frac{2}{S} \sum_i i a_i - \frac{S+1}{S}. \tag{8.3.10}$$

There are a number of variations of this in the literature. The Gini index is widely used to measure things like inequality of income or wealth distribution. For example, $G' = 0.25$ for Denmark, 0.70 for Namibia. Higher G' means things are more uneven.

If all the x_i are the same, then $a_i = 1/S$ for all i, and

$$\sum_i i a_i = \frac{1}{S} \sum_i i = \frac{1}{S}[(\tfrac{1}{2}S)(S+1)] = \tfrac{1}{2}(S+1). \tag{8.3.11}$$

Thus, $G' = 0$. In all other cases, $G' > 0$. We will give a geometric interpretation of the Gini index in Sect. 8.6.

In our example, when $\mathbf{x} = (2, 2, 2, 2, 2, 2, 2)$, $G' = 0$. If $\mathbf{x} = (1, 1, 1, 1, 1, 1, 8)$, then using the low to high version of the equation for the Gini index, we get $G' = (2/7)(1/14 + 2/14 + 3/14 + 4/14 + 5/14 + 6/14 + 56/14) - 8/7 = 3/7 = 0.429$.

8.4 Dalton's Axioms

There are many other indices of evenness that have been proposed over the years. How does one choose? One idea is to write down some general principles (axioms) that a measure of evenness should satisfy and see which of the suggested indices satisfy them. This approach has been widely used in other areas of application, the famous Arrow's axioms for a social welfare function being a prime example. In the case of Arrow, his impossibility theorem [2] shows that there is no social welfare function ("voting rule," "consensus function") that satisfies a reasonable set of axioms. By way of contrast, there are a variety of well-known examples where the given axioms uniquely determine a given function, as for example the axioms that uniquely determine a single solution called the Shapley value in a multi-player game [40], or the axioms that uniquely determine a measure of distance between preference rankings [19]. Another example is the set of axioms for a measure of trophic status of a species in a food web, for which there are multiple solutions, but one minimal one [19]. See [35] for a discussion of these and other examples.

We present some axioms originally due to Dalton [7] in the economics literature and widely discussed in the literature of biodiversity (see for example [9, 10, 36]). (For a different set of axioms, aimed at diversity more generally, see [17].)

Dalton's *Principle of Permutation Invariance* says that concentration or diversity or evenness is not a property of (names of) individual species but of a group of species considered as a whole. Precisely, this is interpreted to mean that if π is a permutation of $\{1, 2, \ldots, S\}$, then

$$f(x_{\pi(x_1)}, x_{\pi(x_2)}, \ldots, x_{\pi(x_S)}) = f(x_1, x_2, \ldots, x_S). \tag{8.4.1}$$

Dalton's *Principle of Scale Invariance* says that a measure of concentration or diversity or evenness should not be influenced by the units used. Precisely, if c is a constant, then

$$f(cx_1, cx_2, \ldots, cx_S) = f(x_1, x_2, \ldots, x_S). \tag{8.4.2}$$

It could be argued that this is not just about units. For instance, (400, 100) might be considered more evenly distributed than (4, 1), since the larger number of the second species at least creates the impression of more evenness. This might be more a matter of diversity than evenness. Also, (4, 1) is not particularly diverse, since having only one individual of the second species creates a very vulnerable situation. However, vulnerability is not a matter of evenness. All this having been

said, Dalton's Principle of Scale Invariance has been widely accepted as a reasonable requirement for a measure of evenness.

Dalton's *Transfer Principle* says that when the rich get richer and the poor get poorer, inequality rises. In other words, if you increase the population of a more abundant species and decrease the population of a less abundant species, there is less evenness. This is made precise as follows: If $x_i < x_j$ and $0 < h \leq x_i$, then

$$f(x_1, x_2, \ldots, x_{i-1}, x_i - h, x_{i+1}, \ldots, x_{j-1}, x_j + h, x_{j+1}, \ldots, x_S) > f(x_1, x_2, \ldots, x_S). \tag{8.4.3}$$

We may now ask: Which evenness indices satisfy these axioms? In particular, which of the indices we have defined above do so?

All five indices we have defined—Simpson's index, Coefficient of Variation, Shannon–Wiener index, Pielou index, and Gini index—satisfy all three of these Dalton axioms/principles. Verifying this for the first two principles is straightforward. Consider the Transfer Principle. For the Simpson index, $f = \lambda = \sum_i {a_i}^2$. Let $T = \sum_k x_k$. If $x_i < x_j$ and $h > 0$, then

$$\begin{aligned} &f(x_1, x_2, \ldots, x_{i-1}, x_i - h, x_{i+1}, \ldots, x_{j-1}, x_j + h, x_{j+1}, \ldots, x_S) \\ &- f(x_1, x_2, \ldots, x_S) = (1/T^2)(2x_j h - 2x_i h + 2h^2) > 0. \end{aligned} \tag{8.4.4}$$

Next, consider the Shannon–Wiener index (Shannon Entropy), $-H'$, and assume $S = 2$. Figure 8.1 shows that H' is maximized when $a_1 = a_2 = 0.5$; moving to the left (reducing x_1 by h and increasing x_2 by h) decreases H'. Thus, this switch increases $-H'$, as required by the Transfer Principle. Similar reasoning applies when $S > 2$.

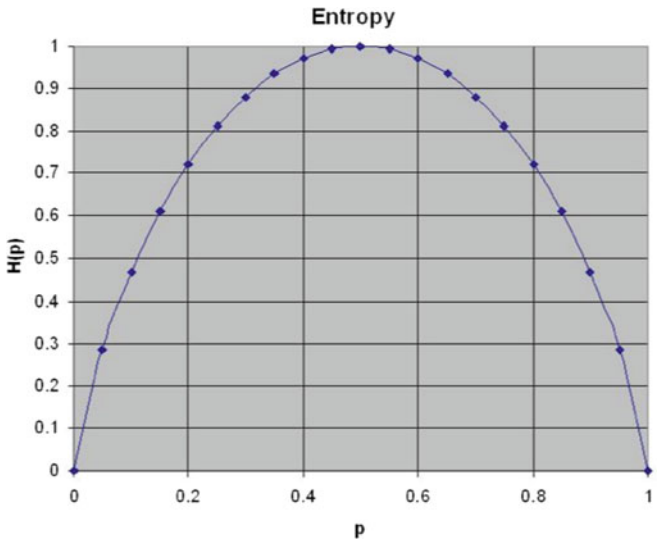


Fig. 8.1 Shanon–Wiener index H' (not $-H'$) normalized to a scale of 0 to 1

It is not hard to verify the Transfer Principle for the other three indices considered so far—Coefficient of Variation, Pielou, and Gini.

Since the five indices we have defined all satisfy the three Dalton axioms, how can we choose among these indices? One way, of course, is to add more axioms.

8.5 Additional Axioms for Measures of Evenness

Here are two new principles that we might impose on measures of evenness. First, the *If the Rich-Get-Richer Principle*. If x_i is the maximum of $\{x_1, x_2, \ldots, x_S\}$, then for every $h > 0$,

$$f(x_1, x_2, \ldots, x_{i-1}, x_i + h, x_{i+1}, \ldots, x_S) > f(x_1, x_2, \ldots, x_S). \tag{8.5.1}$$

Second, the *Principle of Nominal Increase*. If not all x_i are equal, then for every $h > 0$,

$$f(x_1 + h, x_2 + h, \ldots, x_S + h) < f(x_1, x_2, \ldots, x_S). \tag{8.5.2}$$

The latter principle implies that, if everyone receives the same nominal increase in salary, there is less inequality (f decreases) or, if the population of every species increases by the same number of individuals, then there is less inequality. Unfortunately, even these two new principles do not allow us to separate the five evenness indices we have defined.

Egghe and Rousseau [10] show that the If the Rich get Richer Principle and the Principle of Nominal Increase follow from the earlier Dalton Axioms of Permutation Invariance, Scale Invariance, and the Transfer Principle.

Here is another interesting principle: The *Replication Principle* says that, if a population is replicated and a new population consists of the old plus the new one, then the evenness should not change. For instance, this principle says that (3, 7) and (3, 7, 3, 7) should have the same evenness. There is room for discussion as to whether this principle is applicable to evenness of populations in ecosystems.

The Pielou index $-J'$ violates Replication. For example, if $\mathbf{x}$ = (8,2) and $\mathbf{y}$ = (8,8,8,8,8,8,8,8,8,8,2,2,2,2,2,2,2,2,2,2) then $-J'(\mathbf{x}) = -0.72$, $-J'(\mathbf{y}) = -0.86$. Of course, since $-H' = -J' \ln S$, Shannon–Wiener also violates Replication. The Simpson index also violates this condition; for example, if $\mathbf{x}$ = (8,2) and $\mathbf{y}$ = (8,8,2,2), we get $\lambda(\mathbf{x}) = 0.68$ and $\lambda(\mathbf{y}) = 0.34$. However, it is not hard to show that Gini and the Coefficient of Variation satisfy Replication.

Of course, we might add other axioms to further separate the various evenness indices, and also derive other indices that satisfy collections of axioms of interest. There is still much work to be done using axiomatic approaches in this way, and to identify which axioms are appropriate for which biodiversity contexts. It would certainly be interesting to seek out a set of axioms which determine a unique

measure of evenness, as in the Shapley value in game theory and the Kemeny–Snell measure of distance between preference rankings discussed above.

8.6 The Lorenz Partial Order

In this section, we take a different approach. Certain abundance vectors are clearly less evenly distributed than others. We introduce an order relation, $\prec$, on abundance vectors called the *Lorenz Partial Order* and seek out measures of evenness that reflect this partial order. (This section is heavily based on [27, 28, 38]; see also [15].)

If $\mathbf{x}$ and $\mathbf{y}$ are two abundance vectors, we will want an evenness index so that $\mathbf{x} \prec \mathbf{y}$ implies that the evenness index of $\mathbf{x}$ is less than the evenness index of $\mathbf{y}$. This means $\mathbf{x}$ is more even than $\mathbf{y}$. It seems reasonable to want $\prec$ to be a partial order,

$$\text{if } \mathbf{x} \prec \mathbf{y}, \text{ then } \mathbf{y} \not\prec \mathbf{x}, \tag{8.6.1}$$

$$\text{if } \mathbf{x} \prec \mathbf{y} \text{ and } \mathbf{y} \prec \mathbf{z}, \text{ then } \mathbf{x} \prec \mathbf{z}. \tag{8.6.2}$$

However, it could be that neither $\mathbf{x} \prec \mathbf{y}$ nor $\mathbf{y} \prec \mathbf{x}$, in which case we say that $\mathbf{x}$ and $\mathbf{y}$ are incomparable.

The literature has several widely used ways to define such partial orders, although approaches to derive this partial order axiomatically or derive it from fundamental theories about species distributions are lacking. A widely used way to define the partial order $\prec$ follows an idea developed by Max Lorenz in 1905 [23] in a study of inequality of wealth distribution. We discuss this idea next.

Assume that $x_1 \le x_2 \le \cdots \le x_S$. Let $b_j = a_1 + a_2 + \cdots + a_j$ for $j = 1, \ldots, S$, so b_j is the cumulative proportion of the population due to the first j species. The *Lorenz curve* for the abundance vector $\mathbf{x} = (x_1, x_2, \ldots, x_S)$ is the curve in 2-space that connects the $S+1$ points $(0, 0)$, $(1/S, b_1)$, $(2/S, b_2)$, $\ldots$, $(S/S, b_S)$ by straight lines.

For example, if $\mathbf{x}$ = (2,4,4,10), then $\mathbf{a}$ = (1/10,2/10,2/10,5/10) and $b_1 = 1/10, b_2 = 3/10, b_3 = 5/10, b_4 = 1$. The Lorenz curve for $\mathbf{x}$ connects the points (0,0), (1/4,1/10), (2/4,3/10), (3/4,5/10), and (1,1); see Fig. 8.2. Note that the Lorenz curve for (4,8,8,20) is the same as the Lorenz curve for (2,4,4,10); again, we use the points (0,0), (1/4,1/10), (2/4,3/10), (3/4,5/10), and (1,1). In general, $(x_1, x_2, \ldots, x_S)$ and $(cx_1, cx_2, \ldots, cx_S)$ have the same Lorenz curve.

Figure 8.3 compares the Lorenz curve for $\mathbf{x}$ = (2,4,4,10) with the Lorenz curve for $\mathbf{y}$ = (5,5,5,5), where all species are equal. The latter is the curve of perfect evenness.

It is not hard to show that the Gini index for a given abundance vector is the ratio of the area between the Lorenz curve for that vector and the line of perfect evenness and the area under the line of perfect evenness. Since the latter is 1/2, the Gini index is twice the area between the two Lorenz curves.

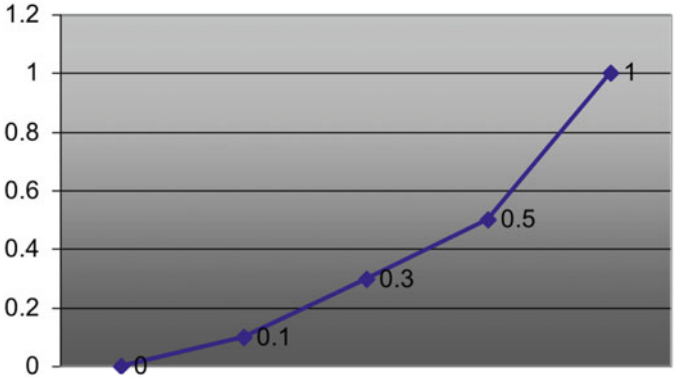


Fig. 8.2 Lorenz curve for the abundance vector (2,4,4,10)

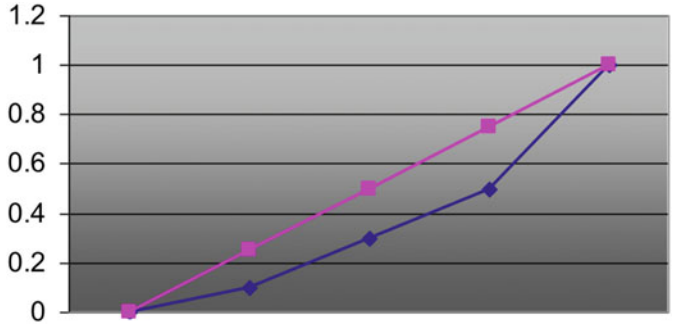


Fig. 8.3 Lorenz curves for the abundance vectors (2,4,4,10) (blue) and (5,5,5,5) (pink)

Sometimes the curve for one abundance vector **x** is strictly above the curve for another vector **y** at all points, i.e., $x_i > y_i$ for all i. If that is the case, we define $\mathbf{x} \prec \mathbf{y}$ in the Lorenz partial order and say that **x** *is more even than* **y**.

Figure 8.4 shows the curves for **x** = (5,6,20,25) and **y** = (2,3,7,44). The former lies entirely above the latter, and therefore $\mathbf{x} \prec \mathbf{y}$ and **x** is more even than **y**.

Two Lorenz curves can cross over, as in Fig. 8.5, which shows the curves for $\mathbf{x} = (5, 6, 20, 25)$ and $\mathbf{y} = (2, 12, 13, 29)$. In this case, neither $\mathbf{x} \prec \mathbf{y}$ nor $\mathbf{y} \prec \mathbf{x}$. The two abundance vectors are incomparable in the Lorenz partial order.

We can now define a *partial order* on Lorenz curves that corresponds to the partial order $\prec$ on abundance vectors: Two Lorenz curves L and L' satisfy the partial order relation $L \prec L'$ if curve L lies strictly above curve L'.

This partial order $\prec$ on curves allows us to do more than the partial order defined so far on vectors. We can now compare abundance vectors with different numbers of species. Defined this way, the order (either on abundance vectors or Lorenz curves) is called the *generalized Lorenz partial order*.

Using this generalized Lorenz partial order, we see in Fig. 8.6 that the curve for $\mathbf{x} = (5, 6, 20, 25)$ is above the curve for $\mathbf{y} = (2, 3, 7, 44, 44)$, so $\mathbf{x} \prec \mathbf{y}$, and (5,6,20,25) is more even than (2,3,7,44,44).

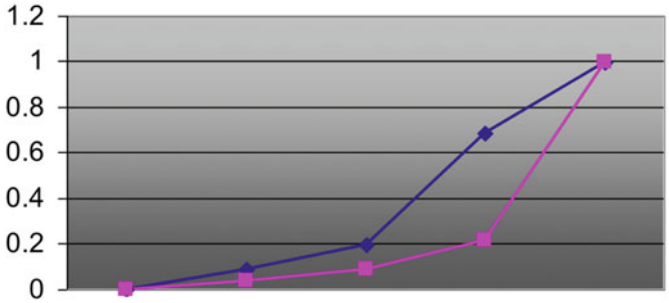


Fig. 8.4 Lorenz curves for the abundance vectors (5,6,20,25) (blue) and (2,3,7,44) (pink). The former is more even than the latter

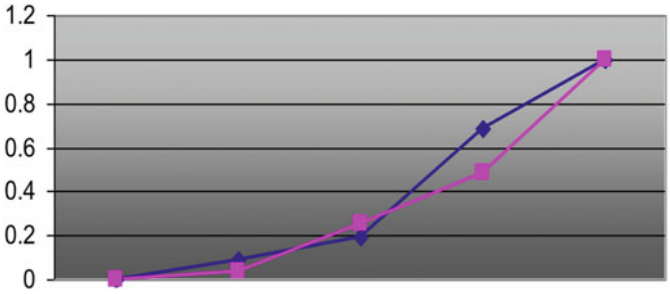


Fig. 8.5 The Lorenz curves for the abundance vectors (5,6,20,25) (blue) and (2,12,13,29) (pink) cross over. These two abundance vectors are incomparable in the Lorenz partial order

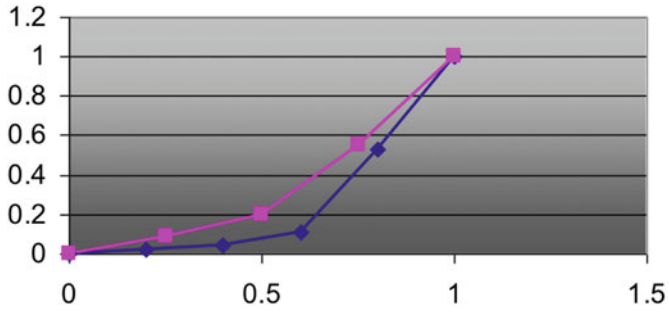


Fig. 8.6 The Lorenz curve for the vector (5,6,20,25) is above the Lorenz curve for the vector (2,3,7,44,44), so the former is more even than the latter

The next requirement we would like to place on an evenness function f is that it *reflects* the (generalized) Lorenz partial order,

$$\mathbf{x} \prec \mathbf{y} \;\Rightarrow\; f(\mathbf{x}) < f(\mathbf{y}). \tag{8.6.3}$$

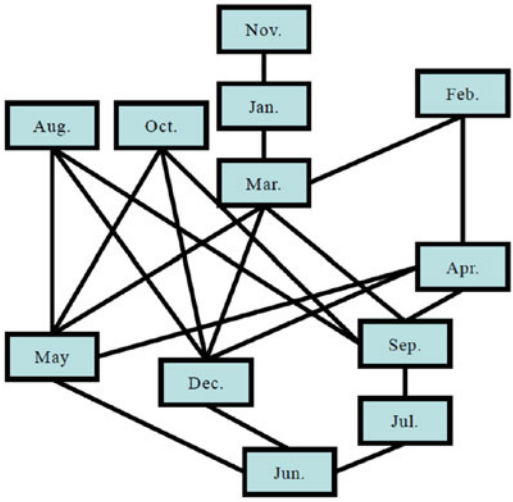


Fig. 8.7 A Lorenz partial order for Cocody Bay, Mauritius. Adapted from [27]

Table 8.1 Gini index for Cocody Bay, Mauritius, as calculated in [27]

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
0.234	0.224	0.181	0.173	0.107	0.076	0.109	0.255	0.130	0.220	0.292	0.114

All of the indices we have defined so far satisfy this condition if the number of species in $\mathbf{x}$ and $\mathbf{y}$ are the same. However, this is not the case if the number of species can differ and we are dealing with the generalized Lorenz partial order. It is not hard to show that, if the number of species in $\mathbf{x}$ and $\mathbf{y}$ can differ, then the Gini index and the Coefficient of Variation satisfy this additional requirement, but Simpson, Shannon–Wiener, and Pielou do not.

To illustrate the ideas, consider the Lorenz partial order derived for Cocody Bay, Mauritius, shown in Fig. 8.7. The partial order is computed in [27] from data in [6]. We see, for example, that Jun$\prec$Dec, Mar$\prec$Feb, and Sep$\prec$Aug. June should get the lowest evenness. Table 8.1 shows the values of the Gini index G'. It is easy to see that the Gini index "preserves" the partial order,

$$\mathbf{x} \prec \mathbf{y} \;\Rightarrow\; G'(\mathbf{x}) < G'(\mathbf{y}). \tag{8.6.4}$$

8.7 Richness vs. Evenness

We have discussed two components of biodiversity, richness and evenness. Here we ask: How are they related?

Wilsey et al. [52] and Ma [24] compared richness and evenness at grassland sites in the North American Great Plains and in agricultural fields in Finland, respectively. Wilsey et al. found a weak negative correlation between richness and

evenness, while Ma found no consistent relationship. Bock et al. [3] did a much more extensive comparison using many more species. (The earlier studies only used flowering plants.) They studied 48 plots in the grasslands and mesquite-oak savannas in the Sonoita Valley of southeastern Arizona in the USA. The study involved 150 species of flowering plants, 32 of grasshoppers, 70 of butterflies, 9 of lizards, 87 of summer birds, 92 of winter birds, and 48 of rodents. Correlations of richness and evenness were neutral to moderately negative for each group. Thus, richness alone is an incomplete representation of biodiversity, since it does not account for species evenness.

Zhang et al. [54] note that "the relationship between species richness and evenness across communities remains an unsettled issue in ecology from both theoretical and empirical perspectives. As a result, we do not know the mechanisms that could generate a relationship between species richness and evenness, and how this responds to spatial scale." Using Pielou's index J' to study sub-alpine meadow communities in the eastern Qinghai-Tibetan Plateau, they found a consistent negative correlation between S and J'.

One may ask: To what extent does richness or evenness impact long-term biodiversity? Daly et al. [8] study this question using a stochastic, spatial, individual-based model to simulate ecosystem dynamics. They study a community of four interacting bacterial species and simulate long-term system behavior. Their results show that higher initial evenness has "a small stabilizing effect on ecosystem dynamics by extending the time until the first extinction."

In the next section we discuss ways to use both concepts, richness and evenness, to define a measure of biodiversity.

8.8 Combining Richness and Evenness

Since biodiversity is more than just richness and more than just evenness, we can explore ways of combining both measures into one index. Jost [18] discusses whether it is possible to decompose biodiversity into independent richness and evenness components. We shall take a different approach, defining different partial orders that reflect richness and evenness and discussing the idea that a biodiversity measure can preserve a given partial order. One way to do that is to use the Lorenz curve L, as well as the number of species S. We will represent an abundance vector by a pair (S, L) and introduce a number of different partial orders $\prec_k$ between pairs (S, L) and (S', L'). Then we seek biodiversity functions B that reflect the partial order $\prec_k$ in question,

$$(S, L) \prec_k (S', L') \Rightarrow B(S, L) < B(S', L'). \tag{8.8.1}$$

The reader should note that, in the present discussion, a lower B value means more biodiversity, just as a lower f value means more evenness when f is an evenness

index. The ideas in the following discussion are due to Rousseau and van Hecke [37] and Rousseau et al. [38].

We introduce a sequence of partial orders, labeled $\prec_1$, $\prec_2$, etc., and identify the first partial order, $\prec_1$, with the generalized Lorenz partial order.

The generalized Lorenz partial order does not account for richness. If richness is the main factor, we could use the partial order $\prec_2$,

$$\mathbf{x} \prec_2 \mathbf{y} \text{ if } S > S' \text{ or } (S = S' \text{ and } L(\mathbf{x}) \prec L'(\mathbf{y})), \tag{8.8.2}$$

where $\prec$ is the ordinary Lorenz partial order. (This is "lexicographic ordering.")

If both the number of species and the Lorenz curve are considered, we could use $\prec_3$,

$$\mathbf{x} \prec_3 \mathbf{y} \text{ if } (S \geq S' \text{ and } L(\mathbf{x}) \prec L(\mathbf{y})). \tag{8.8.3}$$

A fourth partial order arises from an altered Lorenz curve called an *intrinsic diversity profile* or a *k-dominance curve*. List x_i in nondecreasing order. Plot the points $(0, 0), (1, b_1), (2, b_2), \ldots, (S, b_S)$. Recall that $b_j = a_1 + a_2 + \ldots + a_j$. In other words, plot the cumulative number of species on the horizontal axis and the cumulative proportion of the population on the vertical axis. The result is a k-dominance curve. Figure 8.8 shows the k-dominance curves for $\mathbf{x} = (1, 3, 4, 10)$ (pink) and $\mathbf{y} = (1, 1, 3, 3, 4, 4, 10, 10)$ (blue). If we accept the Replication principle, then these populations have the same evenness. However, $\mathbf{x}$ has fewer species, so it should have less biodiversity and therefore higher B than $\mathbf{y}$. In general, we say that

$$\mathbf{x} \prec_4 \mathbf{y} \text{ if the k-dominance curve for } \mathbf{x} \text{ lies below that for } \mathbf{y}. \tag{8.8.4}$$

In the example, $\mathbf{y} \prec_4 \mathbf{x}$.

Rousseau et al. [38] give two more partial orders, which we will not discuss here.

We want to find biodiversity measures B that have the property

$$\mathbf{x} \prec_k \mathbf{y} \Rightarrow B(x) < B(y). \tag{8.8.5}$$

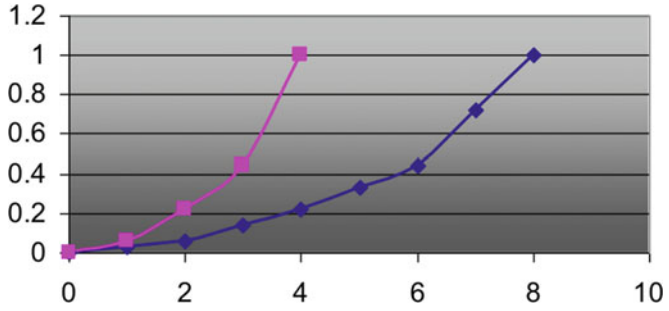


Fig. 8.8 k-dominance curves for (1,3,4,10) and (1,1,3,3,4,4,10,10)

But how do we decide which of the partial orders $\prec_k$ to use to derive a biodiversity measure? One approach is to use axioms, as we did for evenness. That is the approach we take in the next section.

8.9 Axioms to Select a Partial Order to Compare Biodiversity Using Both Richness and Evenness

Rousseau, van Hecke, Nijssen, and Bogaert have proposed a number of axioms (or principles) that can be applied to the selection of a partial order when both richness and evenness are used to compare biodiversity [38].

The *Inheritance Principle* says that, if species richness S is held fixed, then classical evenness determines the partial order to use.

The *Dual Inheritance Principle* says that, if two abundance vectors have the same (classical) Lorenz curve, then species richness determines the partial order. A special case is *Pielou's axiom*, which says that, given two communities each having complete evenness, one with S species and the other with $S + 1$ species, the latter should be more biodiverse and therefore have a lower biodiversity index.

The *Balance Property* says that there must be two abundance vectors of different species richness that are incomparable in the partial order. Thus, species richness cannot completely determine biodiversity.

The *Dual Balance Property* says that there must be two abundance vectors of different evenness (as determined by the generalized Lorenz curves) that are incomparable in the partial order. Thus, evenness cannot completely determine biodiversity.

The generalized Lorenz order $\prec_1$ violates Pielou's axiom and therefore the Dual Inheritance Principle.

The second partial order $\prec_2$ violates the Balance Property, since species richness determines the order if two populations have different species richness.

The third partial order $\prec_3$ satisfies both Inheritance principles and both Balance properties. For example, consider $\mathbf{x} = (S, L) = (1, 3, 4, 10)$ and $\mathbf{y} = (S', L') = (1, 1, 3, 3, 3, 3, 11, 11)$. Then $S < S'$, and it can be shown that $L' \prec_1 L$. It follows that $\mathbf{x}$ and $\mathbf{y}$ are incomparable in $\prec_3$, so both Balance axioms hold.

The fourth partial order $\prec_4$ also satisfies both Inheritance principles and both Balance properties.

Since both $\prec_3$ and $\prec_4$ satisfy the four axioms proposed in [38], the question is how to choose between the two of them. For every $\mathbf{x}$ and $\mathbf{y}$, $\mathbf{x} \prec_3 \mathbf{y}$ implies $\mathbf{x} \prec_4 \mathbf{y}$. Thus, $\prec_4$ allows us to make at least as many (and in fact more) comparisons than $\prec_3$. In that sense, $\prec_4$ is preferable.

We now seek a biodiversity index B that reflects $\prec_4$,

$$\mathbf{x} \prec_4 \mathbf{y} \Rightarrow B(x) < B(y). \tag{8.9.1}$$

Rousseau et al. found one index B that satisfies this condition. It is defined up to a scaling factor by the area below the k-dominance curve. They called it an *adapted*

Gini index. Much work remains to be done to assess whether this or other measures of biodiversity are useful.

8.10 Concluding Remarks

The discussion of richness, evenness, and a combination of richness and evenness shows that more research is needed to define useful measures of biodiversity that can be precisely defined and used by a wide variety of researchers. To give just one example of a problem that needs to be addressed, we note that the classical approach to evenness could be modified to incorporate weights of importance of different species such as indicator species or invasive species.

Different indices of biodiversity, even different indices of evenness, have different advantages and disadvantages. It can be important to see if different indices yield consistent conclusions. We have already observed that richness and evenness might even be negatively correlated. However, an interesting line of research could be to study families of biodiversity indices that depend upon some parameter and give conditions on the range of values of the parameter where the indices will give consistent conclusions, e.g., consistent rankings of biodiversity. This idea is discussed in [5, 34].

We have given criteria (axioms) for a measure of evenness, but not for a measure of biodiversity. As with evenness, such criteria need to be made precise, perhaps using mathematical formulations similar to those discussed in this chapter. There is already a literature on this topic; for example, see [17].

Since variations of different measures are widely used, and numbers obtained are compared, it is important to make sure that everyone is using the same definition. Many authors try to "normalize," so that measures of biodiversity (or evenness) take on values between 0 and 1. It is often assumed, going back to Dalton as one of his axioms, that $B(x_1, x_2, \ldots, x_S) = 0$ when all x_i are equal.

To make the maximum value of B equal to 1, we might make a transformation. For instance, some people use the Coefficient of Variation V in its inverse form, $1/V$. This can be normalized by using instead $(2/\pi)\arctan(1/V)$.

Ideally, a measure of biodiversity should be "intelligible"—intuitively meaningful, easy to explain, easy to understand. For example, Simpson's index has a simple probabilistic interpretation and Gini's index has a simple geometric interpretation.

The biodiversity measure should be sensitive in the sense that it reflects changes in data. However, it should also be robust in the sense that it is insensitive to small changes, especially when the data are not known to great accuracy. Both those concepts could be made precise using mathematical language.

A measure of biodiversity is applied to a particular ecosystem at a particular instant of time. A goal of biodiversity preservation is to create systems that maintain relatively stable biodiversity into the future. A good measure of biodiversity should be usable in models that help us predict that under certain conditions of an evolving ecosystem, the biodiversity will remain relatively stable.

There is no one "best" measure of biodiversity. The measure that is used should be tied to the application for which it is used. This will depend in turn on potential biases and problems in data gathering, the sampling procedure used, the area of the region in question, the goals of the study, etc.

We need to be able to understand the uncertainty in claims about (positive or negative) changes in biodiversity and to find ways to use biodiversity measures to understand how to achieve ecosystems that are sustainable and maintain stability into the future. Only by putting the measurement of biodiversity on a firm mathematical foundation can we be confident that we are capturing the true diversity in nature.

Acknowledgements Parts of this chapter (in particular, some of the Introduction, parts of the nontechnical discussion of Richness in Sect. 8.2 and Evenness in Sect. 8.3, and a portion of the concluding Sect. 8.10) were used in a book (report) *Mathematical and Statistical Challenges for Sustainability* edited by Margaret Cozzens and Fred Roberts, and in particular in Fred Roberts' contribution to the Working Group I Report on Human Well-Being and the Natural Environment, included in the aforementioned book and authored by Alejandro Adem, Michelle Bell, Margaret Cozzens, Charmaine Dean, Francesca Dominici, Avner Friedman, Fred Roberts, Steve Sain, and Abdul-Aziz Yakubu. The author thanks the National Science Foundation for its support under grant DMS-1246305 to Rutgers University.

References

1. Arrhenius, O.: Species and area. J. Ecol. **9**, 95–99 (1921)
2. Arrow, K.: Social Choice and Individual Values. Cowles Commission Monograph, vol. 12. Wiley, New York (1951). Second edition (1963)
3. Bock, C.E., Jones, Z.F., Bock, J.H.: Relationships between species richness, evenness, and abundance in a southwestern savanna. Ecology **88**, 1322–1327 (2007)
4. Boulinier, T., Nichols, J.D., Sauer, J.R., et al.: Estimating species richness: the importance of heterogeneity in species detectability. Ecology **79**, 1018–1028 (1998)
5. Buckland, S.T., Magurran, A.E., Green, R.E., et al.: Monitoring change in biodiversity through composite indices. Philos. Trans. R. Soc. B **360**, 243–254 (2005)
6. Daget, J.: Les Modèles Mathématiques en Écologie. Masson, Paris (1976)
7. Dalton, H.: The measurement of inequality of incomes. Econ. J. **30**, 348–361 (1920)
8. Daly, A.J., Baetens, J.M., De Baets, B.: The impact of initial evenness on biodiversity maintenance for a four-species in silico bacterial community. J. Theor. Biol. **387**, 189–205 (2015)
9. Egghe, L., Rousseau, R.: Elements of concentration theory. In: Egghe, L., Rousseau, R. (eds.) Informetrics, pp. 97–137. Elsevier, Amsterdam (1990)
10. Egghe, L., Rousseau, R.: Transfer principles and a classification of concentration measures. J. Am. Soc. Inf. Sci. **42**, 479–489 (1991)
11. Firebaugh, G.: Empirics of world income inequality. Am. J. Sociol. **104**, 1597–1630 (1999)
12. Firebaugh, G.: Inequality: What it is and how it is measured. In: The New Geography of Global Income Inequality. Harvard University Press, Cambridge (2003)
13. Gini, C.: Il diverso accrescimento delle classi sociali e la concentrazione della richezza. Giornale degli Economisti, serie II **38**, 27–83 (1909)
14. Gini, C.: Variabilite mutabilita, Parte II. Technical report, Univ. di Cagliari III, Studi Economico-coaguridic della Facotta di Giurisprudenza (1912)

15. Gosselin, F.: Lorenz partial order: the best known logical framework to define evenness indices. Commun. Ecol. **2**, 197–207 (2001)
16. Gotelli, N.J., Colwell, R.K.: Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecol. Lett. **4**, 379–391 (2001)
17. Grabchak, M., Marcon, E., Lang, G., et al.: The generalized Simpson's entropy is a measure of biodiversity. PLoS One **12** (2017). https://doi.org/10.1371/journal.pone.0173305
18. Jost, L.: The relation between evenness and diversity. Diversity **2**, 207–232 (2010). https://doi.org/10.3390/d2020207
19. Kemeny, J.G., Snell, J.L.: Mathematical Models in the Social Sciences. Blaisdell, New York (1962). Reprinted by M.I.T Press, Cambridge, MA, 1972
20. Lamas, G., Robbins, R.K., Harvey, D.J.: A preliminary survey of the butterfly fauna of Pakitza, Parque Nacional Del Manu, Peru, with an estimate of its species richness. Technical report, Publicaciones Del Museo de Historia Natural, Universidad Nacional Mayor de San Marcos, A40, 1–19 (1991)
21. Lamb, E.G., Bayne, E., Holloway, G., et al.: Indices for monitoring biodiversity change: are some more effective than others? Ecol. Indic. **9**, 432–444 (2009)
22. Leyesdorff, L., Rafols, I.: Indicators of the interdisciplinarity of journals: diversity, centrality, and citations. J. Informet. **5**, 87–100 (2011)
23. Lorenz, M.O.: Methods of measuring the concentration of wealth. Publ. Am. Stat. Assoc. **9**, 209–219 (1905). https://doi.org/10.2307/2276207
24. Ma, M.: Species richness vs evenness: independent relationship and different responses to edaphic factors. Oikos **111**, 192–198 (2005)
25. Magurran, A.E.: Ecological Diversity and Its Measurement. Chapman and Hall, London (1991)
26. Magurran, A.E.: Measuring Biological Diversity. Blackwell, Oxford (2004)
27. Nijssen, D., Rousseau, R., Van Hecke, P.: The Lorenz curve: a graphical representation of evenness. Coenoses **13**, 33–38 (1998)
28. Patil, G.P., Taillie, C.: Diversity as a concept and its measurement. J. Am. Stat. Assoc. **77**, 548–561 (1982)
29. Pielou, E.C.: The measurement of diversity in different types of biological collections. J. Theor. Biol. **13**, 131–144 (1966)
30. Pielou, E.C.: Ecological Diversity. Wiley, New York (1975)
31. Prathap, P.: Second order indicators for evaluating international scientific collaboration. Scientometrics **95**, 563–570 (2012)
32. Preston, F.W.: The commonness, and rarity, of species. Ecology **29**, 254–283 (1948)
33. Preston, F.W.: The canonical distribution of commonness and rarity: Part I. Ecology **43**, 185–215 (1962)
34. Ricotta, C.: On parametric evenness measures. J. Theoret. Biol. **222**, 189–197 (2003)
35. Roberts, F.S.: Discrete Mathematical Models, with Applications to Social, Biological, and Environmental Problems. Prentice-Hall, Englewood Cliffs (1976)
36. Rousseau, R.: Concentration and diversity of availability and use in information systems: a positive reinforcement model. J. Am. Soc. Inf. Sci. **43**, 391–395 (1992)
37. Rousseau, R., Van Hecke, P.: Measuring biodiversity. Acta Biotheor. **47**, 1–5 (1999)
38. Rousseau, R., Van Hecke, P., Nijssen, D., et al.: The relationship between diversity profiles, evenness and species richness based on partial ordering. Environ. Ecol. Stat. **6**, 211–223 (1999)
39. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**, 379–423, 623–656 (1948)
40. Shapley, L.S.: A value for n-person games. In: Kuhn, H.W., Tucker, A.W. (eds.) Contributions to the Theory of Games. Annals of Mathematics Studies, vol. 28, pp. 307–317. Princeton University Press, Princeton (1953)
41. Simpson, E.H.: Measurement of diversity. Nature **163**, 688 (1949)
42. Smith, B., Wilson, J.: A consumer's guide to evenness indices. Oikos **76**, 70–82 (1996)
43. Soberon, J., Llorente, B.: The use of species accumulation functions for the prediction of species richness. Conserv. Biol. **7**, 480–488 (1993)

44. Sugihara, G.: Minimal community structure: an explanation of species abundance patterns. Am. Nat. **11**, 770–787 (1980)
45. Takacs, D.: The Idea of Biodiversity: Philosophies of Paradise. Johns Hopkins Press, Baltimore (1996)
46. Tuomisto, H.: A diversity of beta diversities: straightening up a concept gone awry. Part 1, Defining beta diversity as a function of alpha and gamma diversity. Ecography **33**, 2–22 (2010)
47. Tuomisto, H.: A diversity of beta diversities: straightening up a concept gone awry. Part 2, Quantifying beta diversity and related phenomena. Ecography **33**, 23–45 (2010)
48. Tuomisto, H.: An updated consumer's guide to evenness and related indices. Oikos **121**, 1203–1218 (2012)
49. UNEP: Convention on biological diversity (1992). https://www.cbd.int/doc/legal/cbd-en.pdf. Accessed 14 Jul 2018
50. UNEP: Report of the Sixth Meeting of the conference of the Parties to the Convention on Biological Diversity (UNEP/CBD/COP/6/20 (2002). https://www.cbd.int/doc/meetings/cop/cop-06/official/cop-06-20-en.pdf. Accessed 14 Jul 2018
51. Whittaker, R.H.: Evolution and measurement of species diversity. Taxon **21**, 213–251 (1972)
52. Wilsey, B., Chalcraft, D.R., Bowles, C.M., et al.: Relationships among indices suggest that richness is an incomplete surrogate for grassland biodiversity. Ecology **86**, 1178–1184 (2005)
53. Wilson, E.O., Peters, F.M. (eds.): Biodiversity. National Academy Press, Washington (1988)
54. Zhang, H., John, R., Peng, Z., et al.: The relationship between species richness and evenness in plant communities along a successional gradient: a study from sub-alpine meadows of the Eastern Qinghai-Tibetan Plateau, China. PLoS One (2012). https://doi.org/10.1371/journal.pone.0049024

Chapter 9
The Mathematics of Extinction Across Scales: From Populations to the Biosphere

Colin J. Carlson, Kevin R. Burgio, Tad A. Dallas, and Wayne M. Getz

Abstract The sixth mass extinction poses an unparalleled quantitative challenge to conservation biologists. Mathematicians and ecologists alike face the problem of developing models that can scale predictions of extinction rates from populations to the level of a species, or even to an entire ecosystem. We review some of the most basic stochastic and analytical methods of calculating extinction risk at different scales, including population viability analysis, stochastic metapopulation occupancy models, and the species–area relationship. We also consider two extensions of theory: the possibility of evolutionary rescue from extinction in a changing environment and the posthumous assignment of an extinction date from sighting records. In the case of the latter, we provide a new example using data on Spix's macaw, the "rarest bird in the world," to demonstrate the challenges associated with extinction date research.

Keywords Mean time to extinction · Sighting records · Sixth mass extinction · Species–area relationship · Population viability analysis

> It's easy to think that as a result of the extinction of the dodo, we are now sadder and wiser, but there's a lot of evidence to suggest that we are merely sadder and better informed.
> – Douglas Adams, *Last Chance to See*

C. J. Carlson · W. M. Getz (✉)
Department of Environmental Science, Policy and Management, University of California Berkeley, Berkeley, CA, USA
e-mail: cjcarlson@berkeley.edu; wgetz@berkeley.edu

K. R. Burgio
Department of Ecology & Evolutionary Biology, University of Connecticut, Storrs, CT, USA
e-mail: kevin.burgio@uconn.edu

T. A. Dallas
Department of Environmental Science and Policy, University of California Davis, Davis, CA, USA
e-mail: tdallas@ucdavis.edu

H. G. Kaper, F. S. Roberts (eds.), *Mathematics of Planet Earth*, Mathematics of Planet Earth 5, https://doi.org/10.1007/978-3-030-22044-0_9

9.1 Introduction

Every species, like every living organism, has a finite lifespan. From the origin of a species onward, every species changes and adapts to its environment. Some species exist longer than others, but all eventually face extinction (or are replaced by their descendants through evolution). Currently, there are approximately 8.7 million eukaryote species alone. But in the history of the Earth, it is estimated that there have been a daunting 4 billion species altogether, and at least 99% of them are now gone [78].

How long can a species exist? Of the species currently on Earth, some are deeply embedded in the geological record and have changed very little over the span of hundreds of millions of years, such as coelacanths and ginkgo trees. Most species persist for a few millions of years or more, and in periods of environmental stability, extinctions typically occur at a low and steady baseline rate. But at various points in the history of the Earth, extinction rates have suddenly accelerated for brief and eventful periods that biologists term *mass extinction events*. In 1982, based on the marine fossil record, Raup and Sepkoski [84] suggested that five of these mass extinctions happened over the past half billion years. In all five, more than half of all contemporary species disappeared [75], and each extinction was sufficiently drastic to be identified with the end of a geological era: the Ordovician 444 million years ago (*mya*), Devonian 375 mya, Permian 251 mya, Triassic 200 mya, and Cretaceous 66 mya.

In recent years, ecologists have reached the consensus that the biosphere is currently experiencing, or at the very least entering, the sixth mass extinction [61]. Unlike the previous five, which were caused by planetary catastrophes and other changes in the abiotic environments, the sixth mass extinction is the undeniable product of human activities. While anthropogenic climate change is one of the most significant contributors, a number of other factors have exacerbated extinction rates, including habitat loss and fragmentation, biological invasions, urbanization, over-harvesting, pollution, pests, and emerging diseases.

How does the sixth mass extinction scale up against the last five? The number of extinctions alone is an unhelpful metric, as species richness changes over time. A more convenient unit of measurement commonly used by scientists is the number of *extinctions per million species-years* (E/MSY). From a landmark study by Gerardo Ceballos and colleagues, we know that in the geological record, vertebrates normally go extinct at a rate of 2 E/MSY in the periods in-between mass extinctions. But since 1900, that rate is an astounding 53 times higher [20]. One study has suggested that the sixth mass extinction is comparable to other mass extinctions in E/MSY rates, meaning that with enough time, the geological definition of a mass extinction (three quarters extinction) could be achieved in hundreds to thousands of years [7]. Or, to consider another metric: a 1970 study estimated that at a baseline, one species goes extinct per year [67], while a decade later that estimate was revised to just up to one species per hour [79]. Plants, insects, and even micro-organisms all face similarly catastrophic threats; and these across-the-board losses of biodiversity

pose a threat to human survival that some argue could even threaten our own species with extinction.

The crisis of extinction is, for scientists, a crisis of prediction. While extinction is a natural part of ecosystem processes and of the history of the planet, the job of conservation biologists is to protect species that would otherwise be brought to an untimely and avoidable end. To do that, conservationists must sort and prioritize the 8.7 million eukaryotes (and even some prokaryotes) to assess which species face the greatest threat—and which can, and cannot, be saved by human intervention. Assessment is easiest at the finest scales: by marking and tracking all the individuals in a region, a population ecologist can make a statistically informed estimate of the probability of imminent extinction. Above the population level, assessment is much more challenging, requiring sophisticated (and complicated) metapopulation models that are typically data-intensive. If a species is rare enough and the data are "noisy," its extinction may seem uncertain even after the fact; but mathematical models can help assign a probability to the rediscovery of a species once thought extinct, and resolve when (and even why) a species has disappeared long after it is gone. Above the level of a single species, measuring extinction is an altogether different problem, requiring a different type of model to explain how biodiversity arises and is maintained over time. Each of these modeling approaches represents a different aspect of a connected problem, and we deal with each in turn in this chapter. The models we present are seminal and well-known, but extinction risk modeling is a dynamic and rapidly growing field. Consequently, these models only present a handful of many different approaches that link different temporal and spatial scales of extinction together.

Outline of the Chapter We begin by discussing the basic mechanics of extinction as a demographic process at the population scale, including population viability analysis, with a case study on evolutionary rescue processes (Sect. 9.2). In Sect. 9.3, we progress up to the metapopulation scale, including patch occupancy models and island biogeography. At the species scale, we dive deeper into the issue of evolutionary rescue, including the potential for plasticity to buffer species from extinction in a changing environment (Sect. 9.4). Expanding at the species level, we discuss the recently growing literature on using sighting records to determine the odds that species are extinct, with a handful of case studies including Spix's macaw and the ivory-billed woodpecker. In the final Sect. 9.5, we discuss how extinction scales up to the community level, and how extinction rates are inferred from habitat loss using macroecological theory.

9.2 The Population Scale

Even though many make a terminological distinction between *extinction* (the loss of a species) and *extirpation* (the eradication of a population), extinction is still fundamentally a process that begins at the population scale. With the exception of sudden, unexpected catastrophes, extinction at the population scale is almost

always the product either of a declining population or of stochastic variations in an already small population, both of which follow mathematical rules that can be used to quantify extinction risk. Perhaps the most significant body of theory about population extinction deals with the estimation of a population's *mean time to extinction* (MTE, typically T_E in mathematical notation), an important quantity to both theoretical ecologists and to conservation efforts. For both theoretical and applied approaches to extinction, understanding the uncertainty around T_E requires an understanding of the shape of the extinction time distribution, including developing and testing demographic theory that accurately captures both the central tendencies [29] and the long tail [30] of empirical extinction times. We begin by reviewing some of the basic population-scale approaches that scale up to ecosystem-level theory of extinction.

9.2.1 Stochasticity and the Timing of Extinction

The simplest deterministic equation governing the size N of a population as it changes over time t (generally measured in units of either years or generations) is given by

$$\frac{dN}{dt} = rN. \tag{9.2.1}$$

The population is growing if $r > 0$, while the population heads towards extinction if $r < 0$. A slightly more complicated model that captures the phenomenological capping of the growth of a population at a *carrying capacity* K is

$$\frac{dN}{dt} = \begin{cases} rN & \text{if } 1 < N < K, \\ 0 & \text{if } N = K. \end{cases} \tag{9.2.2}$$

Equations (9.2.1) and (9.2.2) both imply that, if $r < 0$, $\ln(N)$ declines linearly with slope r. The mean time to extinction, T_E, for a shrinking population can be derived analytically as the amount of time before the population reaches one individual, $N(T_E)$=1,

$$T_E(N_0) = -\ln(N_0)/r. \tag{9.2.3}$$

Consequently, the maximum achievable extinction time for a given population with a fixed r, given a starting stable population size, would be

$$\max(T_E) = -\ln(K)/r. \tag{9.2.4}$$

But deterministic models only tell a part of the story. In the history of conservation biology, two paradigms emerged that separately explain the process of population extinctions. The *declining population paradigm* explains that populations shrink and vanish due to a combination of internal and external failures, and suggests that the key to conserving populations is to identify and prevent those failures. In contrast, the *small population paradigm* is rooted in ideas of stochasticity, suggesting that even without factors like environmental degradation or disease, smaller, more fragmented populations simply face higher extinction risk due to stochastic population processes [19]. For one thing, stochasticity produces populations with a log-normal distributed size (i.e., most populations are comparatively small relative to a few larger ones) due to Jensen's inequality, which can be applied to stochastic processes to show that if r is stochastic, the expectation $E[r]$ of r will always be greater than the expected real growth rate of the population [13],

$$E[r] > E[(N_t/N_0)^{1/t}]. \tag{9.2.5}$$

As a result, stochastic sub-exponential populations all tend eventually to extinction.

In reality, populations show a combination of deterministic and stochastic processes over time, and their extinction is a product of both. In the late 1980s, the field of *population viability analysis* (PVA) emerged from the need to find appropriate analytical and simulation methods for predicting population persistence over time. According to one history of PVA, Mark Shaffer's work on grizzly bears in Yellowstone [9] helped birth the field through two important developments, which we break down in turn below.

Demographic and Environmental Stochasticity Shaffer's first major contribution was the use of extinction risk simulations that account for—and differentiate between—two major kinds of stochasticity, namely *demographic stochasticity*, which is defined at the scale of the individual and occurs through random variation in demography and reproduction, and *environmental stochasticity*, which occurs at a synchronized scale for an entire population (e.g., a bad year may change vital rates uniformly for all individuals in a population). While the impact of environmental stochasticity is ultimately scale-independent, larger populations become less sensitive to demographic stochasticity as they grow. This is due to the integer-based nature of birth and death processes, where populations made up of fewer individuals will suffer a disproportionate effect from a birth or death event.

Demographic and environmental stochasticity have measurably different effects on T_E in basic population models. A simple modeling framework distinguishing between them was laid out in a 1993 paper by Lande [63]. That framework begins again with Eq. (9.2.2), except that we now regard r as an explicit function of time. In the case of demographic stochasticity, individual variations have no temporal autocorrelation, and at the population scale,

$$r(t) \sim \mathcal{N}(\bar{r}, \sigma_d^2/N), \tag{9.2.6}$$

where σ_d^2 is the variance of a single individual's fitness per unit time. Once again, for populations starting at their carrying capacity,

$$T_E = \left(\frac{1}{\bar{r}} \int_1^K \frac{e^{2r(N-1)/\sigma_d^2}}{N} \, dN \right) - \frac{\ln K}{\bar{r}}. \tag{9.2.7}$$

When $\bar{r} > 0$, MTE scales exponentially with carrying capacity, $T_E \propto e^{2r(N-1)/\sigma_d^2}/K$, while when $\bar{r} < 0$, it scales logarithmically, $T_E \propto \ln(K)$, much like in the deterministic decline given by Eqs. (9.2.3) and (9.2.4). In contrast, in the case of environmental stochasticity, the variance acts on the entire population at once,

$$E[\ln N(t)] = \ln N_0 + (\bar{r} - \sigma_e^2/2)\, t, \tag{9.2.8}$$

and the mean time to extinction is now given by

$$T_E = \frac{2}{V_e c} \left(\frac{K^c - 1}{c} - \ln K \right), \quad c = \frac{2\bar{r}}{\sigma_e^2} - 1. \tag{9.2.9}$$

In the case of environmental stochasticity, if the "long-run growth rate" ($\tilde{r} = \bar{r} - \sigma_e^2/2$) is zero or negative, MTE again scales logarithmically with K. When long-run growth is positive, the dynamic is a bit more complicated,

$$T_E \approx 2K^c/(\sigma_e^2 c^2) \quad \text{if} \quad c \ln K \gg 1. \tag{9.2.10}$$

In this case, the scaling of MTE with K curves up if and only if $\bar{r}/\sigma_e^2 > 1$ (i.e., if and only if the intrinsic growth rate exceeds environmental variation).

Minimum Viable Populations and Effective Population Size The second major contribution of Shaffer's work was the introduction of the concept of a *minimum viable population* (MVP). In Shaffer's original work, MVP is defined as the smallest possible population for which there is a 95% chance of persistence (a 5% or lower chance of extinction) after 100 years. In their foundational treatment of the minimum viable population concept, Gilpin and Soulé [42] identify four special cases—*extinction vortices*—in which a population is likely to tend towards its MVP and ultimate extinction. The first, the *R Vortex*, is perhaps the most obvious: demographic stochasticity (variation in r) reduces populations and increases variation in r, a positive feedback loop of demographic stochasticity directly driving populations to extinction. The *D Vortex* occurs when the same processes—potentially in concert with external forces—produce increased landscape fragmentation (see Sect. 9.3.1 for an explanation of D), which not only reduces local population sizes (increasing local extinction rate) but also has subtle effects on population genetic diversity. The final two vortices—the *F Vortex* and *A Vortex*—both concern the genetic and evolutionary trajectories of small stochastic populations. In the former, inbreeding and demographic stochasticity form a feedback cycle, while in the latter, maladaptation

is the underlying mechanism of extinction. Both are especially relevant in research surrounding phenomena like climate change, but fully understanding them requires a mathematical language for the genetic behavior of near-extinction populations.

In heavily subdivided populations with low dispersal, increased inbreeding can lead to decreased genetic diversity and the accumulation of deleterious or maladapted alleles that make the total population less viable than its size might indicate. As a consequence, intermediate-size populations with low genetic diversity can behave, mathematically, like small populations. *Effective population size*, N_e, quantifies that phenomenon, expressing the genetically or reproductively "effective" number of individuals in a population. In some cases, measuring population size with N_e may more readily allow the computation of a meaningful and predictive MVP, by removing some of the variability between different populations of the same size and by more accurately capturing the long-term reproductive potential of the available genetic material. (Relatedly, it is worth noting that in one unusual study, it was found that there is no statistical link between species MVP and global conservation status [14].)

A number of different approaches exist for the estimation of N_e. Sewall Wright, who created the concept of effective population size, offered one interpretation based on neighborhoods. In his model, offspring move a distance away from their parent based on a two-dimensional spatial normal distribution with standard deviation σ [107]. If individuals have a density D, then

$$N_e = 4\pi\sigma^2 D. \tag{9.2.11}$$

Wright [108] also provides a more commonly invoked method of calculating N_e based on sex structure, using N_m and N_f to, respectively, denote the number of breeding females and males in the population,

$$N_e = \frac{4N_m N_f}{N_m + N_f} \tag{9.2.12}$$

In such an approach, a population of all males or all females would have $N_e = 0$ because no new offspring could be produced in the next generation, rendering the population functionally extinct. That method of deriving N_e is still frequently cited in population conservation work, as small populations tend to stochastically deviate from a 50:50 sex ratio, sometimes severely impacting long-term survival.

A more genetics-based method of calculating N_e comes from the Wright–Fisher model of a two-allele one-locus system, referred to as the *variance effective population size* [21]. In that model, variance between generations $\sigma^2(a)$ for allele A with frequency a is given as $a(1-a)/(2N)$, resulting in an effective population size

$$N_e = \frac{a(1-a)}{2\sigma^2}. \tag{9.2.13}$$

Alternatively, for a locus with a greater degree of polymorphism, or multi-locus microsatellite data, genetic diversity θ and mutation rate μ are related by

$$N_e = \frac{\theta}{4\mu}. \tag{9.2.14}$$

A more commonly used metric in current literature is *inbreeding effective population size*. To construct that metric, we start by defining population-level measures of heterozygosity. In the simplest Hardy–Weinberg formulation for a two-allele system with allele frequencies a and $1 - a$, the expected fraction of heterozygote offspring is $E(H) = 2a(1-a)$. By counting the real fraction of heterozygotes and comparing, we can measure the assortiveness of mating,

$$f = \frac{E(H) - H}{H}. \tag{9.2.15}$$

That value f is called the inbreeding coefficient, ranging from 0 to 1; again according to Wright [3], N_e should be calculated such that it satisfies

$$N_e = \frac{1}{2\Delta f}, \tag{9.2.16}$$

where Δf is the change per generation (in a declining or small population, genetic diversity decreases at a rate determined by the population size and inbreeding).

Returning to the extinction vortex concept with N_e in mind clarifies the genetic component of those extinction processes. While the *D Vortex* reduces N_e as a byproduct of fragmentation (in fact, decreasing neighborhood size), the last two extinction vortices bring N_e below the MVP through specifically genetic modes of extinction. In the *F Vortex*, a positive feedback loop between increased inbreeding (hence f, the inbreeding coefficient) and decreases in effective population size drive a population to extinction over a few generations. A notorious real-world example of such a process might be the near-extinction (or extinction, depending on one's species concept) of the Florida panther, a subspecies of *Puma concolor* ultimately rescued through outbreeding with Texas panthers. All things considered, their rescue was both fortuitous and improbable, as the species was assigned a 5% or less chance of avoiding imminent extinction in 1995 [55]. Finally, in the *A Vortex* (for adaptation), decreased N_e acts as a buffer to the strength of selection acting on phenotypes that are closely paired with environmental variation or change, leading to mismatch between them that reduces both r and N (and N_e) until extinction (a process we cover in much greater detail in Sect. 9.4.1) . Obviously, the four vortices are not independent processes and probably often exist in combination in real-world cases of population extinction.

Population Viability Analysis Through Simulation Usually, MVP is often calculated through simulation methods, which benefit from a greater ease of incorporating age, sex structure, and other population-scale heterogeneities. Even though these

methods are still the foundation of most population-level extinction analyses, they date as far back as P. H. Leslie's population analyses in the late 1940s in the framework of discrete matrix models and linear systems theory. Formulations of the Leslie model and the theory behind such models can be found in several expository texts [18, 39], with a brief outline provided here.

In the Leslie model, the population is divided into n age classes, where $N_i(t)$ is used to denote the number of individuals in age class i at time t. In each age class, the parameter s_i $(0 < s_i \leq 1)$ is used to represent the proportion of individuals aged i that survive to age $i+1$, in which case the variables $N_i(t)$ and $N_{i+1}(t+1)$ are linked by the equation

$$N_{i+1}(t+1) = s_i N_i(t). \tag{9.2.17}$$

We either terminate this sequence of equations at age n by assuming that $s_n = 0$ (i.e., no individuals survive beyond age n), or we interpret N_n as the group of individuals in the population aged n and older and use the equation

$$N_n(t+1) = s_{n-1} N_{n-1}(t) + s_n N_n(t) \tag{9.2.18}$$

to imply that all individuals aged n and older are subject to the survival parameter s_n (i.e., individuals older than age n are indistinguishable from individuals aged n). If we now interpret $N_0(t)$ as all newborn individuals born just after individuals have progressed one age class, then $N_0(t)$ can calculated using the formula

$$N_0(t) = \sum_{i=1}^{n} b_i N_i(t), \tag{9.2.19}$$

where b_i is the average (expected) number of progeny produced by each individual aged i. In this model we have not differentiated between the sexes; so, for example, if each female aged i is expected to produce three young and the population has a 1:1 sex ratio (same number of males to females), then $b_i = 1.5$ for this age class. If we now apply Eq. (9.2.17) for the case $i = 0$, we obtain the equation

$$N_1(t+1) = s_0 N_0(t) = s_0 \sum_{i=1}^{n} b_i N_i(t). \tag{9.2.20}$$

Equations (9.2.17)–(9.2.20) can be written compactly in matrix notation,

$$\mathbf{N}(t+1) = L\,\mathbf{N}(t), \tag{9.2.21}$$

where

$$\mathbf{N} = \begin{pmatrix} N_1 \\ \vdots \\ N_n \end{pmatrix}, \quad L = \begin{pmatrix} s_0 b_1 & \cdots & s_0 b_{n-1} & s_0 b_n \\ s_1 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & s_{n-1} & s_n \end{pmatrix}.$$

The *Leslie matrix* L is nonnegative, since all its elements are nonnegative, with at least one positive element. Further, if there exists some integer $p > 0$ such that L^p is positive (i.e., all its elements are positive), then it is known by the *Perron–Frobenius Theorem* that the matrix L has a dominant positive eigenvalue λ_p (known as the *Perron root*) and a corresponding eigenvector $\mathbf{v}_p$ whose elements are all positive; λ_p and $\mathbf{v}_p$ characterize the long-term behavior of $\mathbf{N}$,

$$\mathbf{N}(t) \sim (\lambda_p)^t \mathbf{v}_p. \tag{9.2.22}$$

That is, $\mathbf{N}(t)$ grows like $(\lambda_p)^t$ as t gets very large, and the ratio of different age classes matches the ratio of elements of $\mathbf{v}_p$. This implies that, if $\lambda_p > 1$ ($\lambda_p < 1$), $\mathbf{N}(t)$ will grow (decline) geometrically as λ_p^t and approach the *stable age-distribution* characterized by the ratio of consecutive elements of $\mathbf{v}_p$. Thus, this model predicts that the population will go extinct whenever the largest eigenvalue of L is less than one ($0 < \lambda_p < 1$). On the other hand, if $\lambda_p > 1$, then we expect density-dependent effects at some point to rein in the unfettered growth by causing survival rates to decline. In particular, if the survival rate s_0 of the youngest age class is the most dependent of all the survival rates on the total biomass density $B = \sum_1^n w_i N_i$, where $w_i > 0$ is the average weight of an individual in age class i, then we should replace s_0 in Eq. (9.2.20) with an expression such as

$$s_0 = \frac{\hat{s}_0}{1 + (B/K_0)^\gamma}, \tag{9.2.23}$$

where $\hat{s}_0$ is the density-independent survival rate, K_0 is the density at which $\hat{s}_0$ is halved, and $\gamma > 1$ is a so-called abruptness parameter (which controls the abruptness in the onset of density, approaching a step down function as γ gets large [38]). Similar modifications can be made to the other survival parameters s_i, depending on their sensitivity to changes in population density.

Stochastic equivalents of these deterministic models typically treat the survival rates s_i as probabilities that each individual survives each time period, rather than as the proportion of individuals surviving each time period; and b_i itself is a random variable drawn from an appropriately defined distribution (usually the binomial distribution). Stochastic models of this sort can be made even more complex by adding more population structure (e.g., genetic variability) or increased levels of complexity (e.g., modeling at the metapopulation scale, discussed in Sect. 9.3, or adding underlying environmental variation or other landscape structures). Though MVP or extinction rates might be difficult to calculate analytically for models of this level of complexity, repeated simulation can easily allow empirical derivation of these properties of a system [76] and is perhaps the most widespread practice for

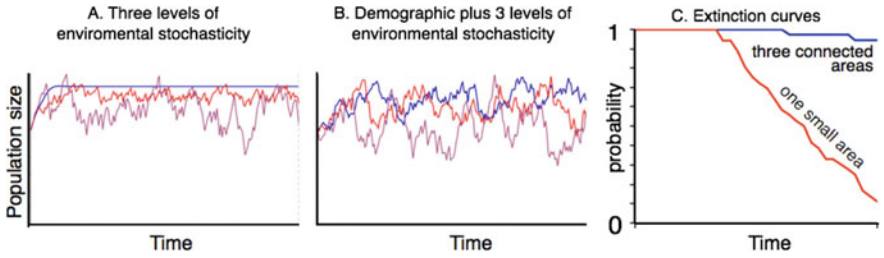


Fig. 9.1 An example of PVA without (**a**) and with (**b**) demographic stochasticity, with no (blue), medium (red), and high (purple) environmental stochasticity. With repeated simulation, an "extinction curve" can be plotted from the probability of population survival over time (**c**). The analysis can be used to make decisions about management and conservation: here, illustrating that three populations with migration between them survive much longer in a poached population of rhinos than a single population. An interactive tutorial of PVA, which can be adjusted to produce anything from the simplest population dynamics to a stochastic, structured metapopulation experiencing harvesting can be found at http://www.numerusinc.com/webapps/pva

estimating population extinction risk in conservation research. An example using an interactive web app [40] is shown in Fig. 9.1.

9.2.2 Case Study: PVA, Disease, and Evolutionary Rescue

In 2015, an epidemic of unknown identity eliminated more than half of the population of the critically endangered saiga antelope (*Saiga tatarica*) in the short span of 3 weeks. While the causative agent was ultimately identified as a species of *Pasteurella*, the mechanism by which a normally asymptomatic non-pathogenic bacterium killed at least 130,000 antelopes is still in question [77]. Literature explaining the die-off, or predicting the consequences for the species, remains comparatively limited; the fate of the species remains uncertain, and it may yet face extinction in the coming years.

Disease is rarely responsible for the extinction of a cosmopolitan species. But for already-threatened species like the saiga, it can be one of the most rapid, unpredictable and unpreventable mechanisms of extinction. Disease has been implicated in a handful of notable wildlife extinctions, like that of the thylacine (*Thylacinus cynocephalus*) or Carolina parakeet (*Conuropsis carolinensis*), and has been the definitive mechanism of extinction for species like the eelgrass limpet (*Lottia alveus*) [25]. While most diseases co-evolve with their hosts to an optimal virulence that prevents the species from reaching extinction, diseases that can persist in the environment may be released from such constraints and be more likely to evolve "obligate killer" strategies (like that of anthrax [37]). Fungal pathogens in particularly tend to have rapid intra-host growth rates and high transmission potential, which can result in population collapses before optimal virulence levels can be attained [35].

Two notable fungal diseases have recently demonstrated the destructive potential of environmentally transmitted pathogens. Perhaps the most significant example of disease-driven extinctions is the trail of destruction caused by the chytrid fungus *Batrachochytrium dendrobatidis* (Bd). Bd has been found in at least 516 species of amphibians [80] and has driven decline or extinction in at least 200 [96], including at least two-thirds of the genus *Atelopus* alone [82]. According to some estimates, current extinction rates that amphibians face (largely but not entirely due to chytrid) are roughly 200 times the background rate; including declining species, that estimate is closer to an even more staggering 25–45,000 [73]. White-nose syndrome (*Geomyces destructans*), a similar fungal epizootic, has similarly spread through bat populations in the eastern United States, causing widespread population-level die-offs since the mid-2000s. While white-nose syndrome has yet to drive any entire species to extinction, significant concern remains regarding its ongoing spread; one study in 2010 using population viability analysis suggested a 99% extinction risk for the little brown bat *Myotis lucifugus* in under two decades. Even in a best-case scenario where white-nose mortality was reduced to one-twentieth of its rate, substantially reducing extinction risk, bats would still be reduced to one percent of their original population size.

White-nose syndrome (WNS) has also become a potential case study for evolutionary rescue, one of the most controversial phenomena in extinction research. The premise that rare genes for resistance or tolerance can bring a disease-ridden population back from the brink of extinction has theoretical support, and potentially indicated from the rapid evolutionary response of certain hosts documented throughout the literature [4]. But WNS constitutes one of the most interesting and controversial examples because, while populations show some sign of recovery from the disease at the time of this writing, no definitive genetic mechanism for resistance has been isolated—a necessary component of demonstrating evolutionary rescue from disease-induced extinction [4]. Consequently, speculation about evolutionary rescue is controversial and so far has been conducted in primarily theoretical settings. In an age-structured matrix population model proposed by Maslo and Fefferman, two scenarios for recovery from WNS are considered [71]. In one model, bats' adaptive immunity leads to re-stabilization at much lower levels overall, but a much faster recovery to a stable balance of juveniles (J) and adults (A), with subscript t denoting the number of individuals in these two age classes at time t. In that model, in the absence of WNS,

$$\begin{pmatrix} J_{t+1} \\ A_{t+1} \end{pmatrix} = \begin{pmatrix} 0.95 & 0.35 \\ 0.95 & 0.87 \end{pmatrix} \begin{pmatrix} J_t \\ A_t \end{pmatrix}. \tag{9.2.24}$$

In a second model, recovery comes not from adaptive immunity but from innate immunity through a genetic mechanism for resistance. In that scenario, a robust type (R) is present in the gene pool with frequency p, and the remainder of individuals are wild type (WT), resulting in the matrix model

$$\begin{pmatrix} J_{t+1} \\ A_{t+1} \end{pmatrix} = p_t \begin{pmatrix} 0.86 & 0.32 \\ 0.86 & 0.78 \end{pmatrix} \begin{pmatrix} J_t^{\mathrm{R}} \\ A_t^{\mathrm{R}} \end{pmatrix} + (1 - p_t) \begin{pmatrix} 0.52 & 0.27 \\ 0.52 & 0.46 \end{pmatrix} \begin{pmatrix} J_t^{\mathrm{WT}} \\ A_t^{\mathrm{WT}} \end{pmatrix}. \quad (9.2.25)$$

In this model, an 11-year stabilization period ultimately leads to population recovery with a positive net growth rate (calculated as the dominant eigenvalue $\lambda = 1.05$), potentially saving populations from extinction. Despite the lack of genetic evidence for evolutionary rescue, Maslo and Fefferman propose that observed similarities between the dynamics they observe and real data on white-nose outbreaks suggest that evolutionary rescue may be happening in real time.

9.3 The Metapopulation Scale

Populations rarely exist in isolation, but are often connected to other populations through dispersal processes, creating a metapopulation. Metapopulations are considered to be in a relatively constant state of flux, as local extinctions of species in habitat patches are buffered by recolonization by dispersal. In this way, dispersal can be beneficial or detrimental to metapopulation persistence. Under high dispersal, patches become homogeneous and population dynamics tend to become synchronous. This synchrony is destabilizing, in that periods of low population sizes will be experienced by all patches, increasing the likelihood of stochastic extinction of the entire metapopulation. On the other hand, too little dispersal will result in spatial clustering of a species, as the species will be confined to the set of patches that can be successfully reached and colonized and similarly potentially increasing extinction risk [1, 2].

The importance of dispersal to patch-level colonization and metapopulation persistence highlights that extinction processes occur at two scales in metapopulations. Specifically, extinction can occur both at the local patch-level (i.e., a single population in the network of habitat patches) or at the entire metapopulation level (i.e., either through catastrophic events or cascading local extinctions). Extinctions of single patches can occur as a result of demographic, environmental, or genetic stochasticity (addressed in more detail in Sect. 9.2.1), or through extrinsic events related to habitat loss or natural enemies [48]. Metapopulation level extinction can also result from environmental stochasticity at the regional scale [17], provided this stochasticity is spatially autocorrelated, such that it is expected to promote synchronous dynamics among habitat patches [45].

9.3.1 Basic Metapopulation Models and Extinction

In the classic metapopulation model described by Richard Levins, the balance between patch colonization (c) and local extinction (e) determines patch occupancy dynamics. In this case, local habitat patches are either occupied or unoccupied,

and both patch number and the spatial orientation of patches are undescribed. It is important to note that in a metapopulation, there are two levels of extinction; individual habitat patches may go extinct, or the entire metapopulation may go extinct. Dispersal among habitat patches can rescue patches from extinction or allow for the recolonization of extinct patches. This becomes more important when we consider dispersal dynamics, subpopulation synchrony, and environmental stochasticity.

The basic formulation of the Levins model is

$$\frac{dP}{dt} = cP(1-P) - eP, \tag{9.3.1}$$

where the balance between e and c determines long-term persistence of the metapopulation [66]. A necessary condition for metapopulation persistence in this model is

$$\frac{e}{c} < 1, \tag{9.3.2}$$

where, at equilibrium, the patch occupancy is given as

$$\hat{P} = 1 - \frac{e}{c} \tag{9.3.3}$$

In this model, the mean time to extinction of any given population is the inverse of the rate (i.e., $T_E = 1/e$), providing a link to the models at the population scale discussed above.

We can take the Levins model a step further to explicate the relationship between patch occupancy and overall mean time to extinction T_M at the metapopulation scale. Starting with the assumption that each of the H patches has its own average extinction time T_L (which should be the inverse of e), we have

$$T_M = T_L \exp\left(\left(\hat{P}H\right)^2 / \left(2H(1-\hat{P})\right)\right). \tag{9.3.4}$$

Consequently, using Eq. (9.3.3), we can also express T_M as

$$T_M = T_L \exp\left(\frac{H}{2}\left(cT_L + \frac{1}{cT_L} - 2\right)\right), \tag{9.3.5}$$

showing that metapopulation extinction time increases exponentially, not linearly, with the MTE of individual habitat patches [47].

The Levins model is mathematically equivalent to a logistic model, a well-developed model often used to examine single species population dynamics. The simplicity of the Levins model has resulted in a sizable body of literature surrounding and extending the model. For instance, in the original Levins model all patches are equidistant from one another, identical in quality, and can only be in one of two

potential states (occupied or unoccupied), but each of these conditions is frequently adjusted in derivative stochastic patch occupancy models (SPOMs). Researchers have shown that despite the simplicity, Levins-type dynamics can emerge from more complicated stochastic metapopulation models [32], and extensions of the Levins model continue to provide insight into the influence of habitat patch size and topography (i.e., spatial orientation of habitat patches) on metapopulation persistence [41].

Island Biogeography and Metapopulation Capacity A simple extension of the Levins model considers a set of spatially explicit patches of variable size, where a distance matrix D describes the distance between all patches in the metapopulation. The model borrows elements of Island Biogeography Theory [70], such that distance between patches (D_{ij}) and patch area (A_i) influence extinction and colonization processes, where the extinction constant (e) is modified for each patch based on area ($e_i = e/A_i$) and colonization becomes a property of distance (D_{ij}), patch area (A_i), and dispersal rate (α),

$$c_i = e^{-\alpha D_{ij}} A_j p_j(t). \tag{9.3.6}$$

This suggests that the mean time to extinction of a habitat patch ($1/e_i$) is determined by the area of the patch. This makes the occupancy probability of each patch in the metapopulation, described in terms of matrix M,

$$M_{ij} = e^{-\alpha D_{ij}} A_i A_j, \tag{9.3.7}$$

and the leading eigenvalue of this matrix M describes the persistence of the metapopulation (*metapopulation capacity* λ_m [49]). The condition for metapopulation persistence is that the dominant eigenvalue of M must be greater than the ratio between extinction and colonization rates,

$$\lambda_M > e/c. \tag{9.3.8}$$

Since habitat patches vary in their size and connectedness to other patches, it is possible to determine the relative importance of each habitat patch to metapopulation persistence in this framework [46, 49], potentially informing conservation and management decisions [102]. While spatially explicit, this approach does assume that dispersal among habitat patches is determined by patch area and distance to other patches, ignoring population dynamics in each patch.

Incorporating Patch Dynamics The above extension of the Levins model allows for patches to vary in size and connectedness. Another extension is to consider the abundances of habitat patches within the metapopulation, thus considering the dynamics of each patch and the effects of dispersal among local populations [89],

$$N_i(t+1) = R_i(t) N_i(t) e^{-N_i/K}. \tag{9.3.9}$$

This expression assumes that the growth rate of each habitat patch is R_i and that the carrying capacity is a constant K. If we assume that the population growth rates (r_i) are *iid* Gaussian random variables, this causes R_i values to be log-normally distributed and allows us to define persistence thresholds for the metapopulation based on the variance in the population growth rates r_i. The threshold for metapopulation persistence relies on exceeding a threshold value (σ_{th}) in terms of the variance among local patch population growth rates (r_i). This threshold is

$$\sigma_{th} > \sqrt{2|\mu_i|}, \tag{9.3.10}$$

where μ_r is the mean local population growth rate over time. This model can be extended to yield many interesting conclusions. For instance, if populations have influence on where their offspring go, population growth rates may be maximized by seeding offspring in less than suitable "sink" habitat if habitat quality fluctuates with time, and when the "source" habitat occasionally experiences catastrophes [54]. The complexity of metapopulation dynamics in the face of environmental stochasticity, variable patch quality, dispersal, and competition has fueled some great theoretical work [12, 72]. An obvious next step is to scale from single species metapopulations to multi-species communities (i.e., metacommunities), which allows for the modeling of how species interactions, predator-prey dynamics, and community assembly relate to persistence [65].

9.4 The Species Scale

Extinction is defined at the scale of the species, but it is also at this level of taxonomic resolution that it is perhaps hardest to quantify—and, to summarize—due to considerable diversity of approaches and applications. We explore in this section two applied extensions of that body of theory, corresponding to two common quantitative frameworks for species-level extinctions. In the first, the complete loss of suitable habitat leads to an inevitable—if not immediate—extinction. Species can escape extinction through three primary channels: acclimation, adaptation, and migration. Species distribution models are often used to calculate extinction risk at the community scale in that framework (described in greater detail below), but they can only at best include the last of those three rescue processes. Evolutionary models, on the other hand, can link demography and genetics to the overall risk of extinction in a changing environment. We explore that application here in the context of both adaptation and phenotypic plasticity.

The second framework is based on the notion that population extinctions become species extinctions; and so the framework for population (and metapopulation) viability analysis described above acts as a sufficient method for estimating species extinction risk. In many cases, that may be a safe assumption, as near-extinction species are reduced down to a single persistent population or a handful in isolated

refugia. But in real applications, persistence in small isolated refugia may be difficult to study, or even observe with any regularity; consequently, an entire body of literature has been developed to relate extinction risk to the sightings of rare species. That body of theory allows two applications: the posthumous assignment of extinction dates to extinct species, and sighting-based hypothesis testing for a species of unknown extinction status. We explore both applications briefly below.

9.4.1 Adaptation and Plasticity in a Changing Environment

Bounding uncertainty is the seminal challenge to extinction research, and in the real world, species' potential to acclimate and adapt to changing environments confers an unknown degree of robustness that has the potential to give species a chance at evading extinction. As discussed above, evolutionary rescue has been a particularly tantalizing—and controversial—idea in the context of disease research. But more broadly, evidence suggests that extinction risk is heavily complicated by species' variable ability to track changing climates (and, more broadly, changing environments).

Most models that estimate the potential for evolutionary rescue approach the problem by explicitly modeling fitness curves and the speed of natural selection. In a foundational paper by Gomulkiewicz and Holt [43], an environmental change beginning at time 0 is followed by changes determined by fitness W such that

$$N_t = \bar{W}_{t-1} N_{t-1} = \prod_{i=1}^{t-1} W_i N_0. \tag{9.4.1}$$

If the population has a critical density N_c below which extinction is certain—essentially, a pseudo-extinction threshold in a PVA framework—extinction time is evolutionarily fixed without adaptation (i.e., $W_t = W_0$),

$$T_E = \frac{\ln N_c - \ln N_0}{\ln \bar{W}_0}. \tag{9.4.2}$$

To address evolutionary potential, Gomulkiewicz and Holt adapt Lande's equations, which describe the rate of natural selection on a single phenotypic trait [62]. In their notation, the trait z has an optimum phenotype normalized to zero, making d_t the distance of observed phenotypes from optimal phenotype at each time step, and d_0 the initial distance (i.e., the initial mean phenotype of the population). Any individual phenotype z is normally distributed around d_t in a distribution p that determines fitness,

$$p_t[z] \sim \mathcal{N}(d_t, {\sigma_z}^2) \tag{9.4.3}$$

The corresponding fitness function with width ω_z is expressed as

$$W(z) = W_{\max} e^{-z^2/(2\omega_z)}, \tag{9.4.4}$$

where $W_{\max}$ is the fitness at $z = 0$. The same expression can also be used to describe the overall tendency of the system,

$$\bar{W}_t = W_{\max} \sqrt{\omega_z/(\sigma_z^2 + \omega_z)}\, e^{-d_t^2/(2{\sigma_z}^2 + 2\omega_z)}. \tag{9.4.5}$$

The expression can be mildly simplified by defining $\hat{W}$ such that it is the growth rate of the optimum mean phenotype population,

$$\hat{W} = W_{\max} \sqrt{\omega_z/(\sigma_z^2 + \omega_z)}. \tag{9.4.6}$$

How does the actual distribution of phenotypes change over time? In real systems, evolution is seldom a direct progression towards the optimum, even under hard selection with ample genetic variation. If the trait z has a heritability h^2, they define an "evolutionary inertia,"

$$k = \frac{\omega_z + (1 - h^2){\sigma_z}^2}{\omega_z + {\sigma_z}^2}; 0 \le k \le 1 \tag{9.4.7}$$

$$d_t = k^t d_0 \tag{9.4.8}$$

which together produce a governing expression for the system,

$$t \ln \hat{W} - \frac{{d_0}^2}{2(\omega_z + {\sigma_z}^2)} \frac{1 - k^{2t}}{1 - k^2} = \ln \frac{N_c}{N_0}, \tag{9.4.9}$$

If this equation has no roots when solving for t, then this indicates the population will fall and rise without any real extinction risk. But when it does, the roots are estimates of the time until the population falls below the critical threshold (T_E) and the time until recovery could be evolutionarily possible (T_P in their notation, where N_t passes back above N_c). The interval between these two values is characterized by a small population that, due to demographic stochasticity, would require much more intensive conservation efforts (e.g., managed ex situ breeding) than normal to possibly survive that interval. The time to recovery (growth switches from negative to positive even though $N_t < N_c$) is

$$T_R = \frac{1}{\ln k^2} \left(\ln \ln \hat{W} - \ln \frac{{d_0}^2}{2(\omega_z + {\sigma_z}^2)} \right). \tag{9.4.10}$$

From this expression, Gomulkiewicz and Holt derive a useful finding: "t_R increases logarithmically with the degree of initial maladaptation ... but is independent of the initial population density."

The model developed by Gomulkiewicz and Holt sets useful theoretical bounds on the genetically coded evolution of a trait. But in the real world, phenotypic plasticity represents some of the most difficult to quantify potential for species to escape extinction. In an extension of similar models developed by Chevin et al. [22], the trait z has a developmental trajectory with both a genetic component and the potential for phenotypic plasticity in response to an environmental gradient ε. Their model uses a "reaction norm" approach to plasticity (popularized by Schlichting et al. [94]), breaking down that phenotypic trait into an adaptive genetic component a and a plastic component b that responds to the environmental gradient. They express the distribution of the phenotype $p(z)$ at generation n in an environment changing at rate $\varepsilon = \eta t$ as

$$p(z) \sim \mathcal{N}(\bar{z}, {\sigma_z}^2), \tag{9.4.11}$$

$$\bar{z} = \bar{a} + b\eta\big(T(n-\tau)\big), \tag{9.4.12}$$

$$\sigma_z^2 = {\sigma_a}^2 + {\sigma_e}^2, \tag{9.4.13}$$

where T is the generation time, developmental plasticity takes effect at time τ during ontogeny, and the strength of plasticity b (the slope of a phenotypic reaction norm) does not evolve over time. Assuming there is an optimum phenotype $\theta = B\varepsilon$, they define a changing population size with a maximum growth rate $W_{\max}$, such that

$$W(z) = W_{\max} \exp\left(-\frac{(z-\theta)^2}{2\omega_z} - \frac{b^2}{2\omega_b}\right), \tag{9.4.14}$$

where both ω's represent the strength of stabilizing selection (the width of fitness curves, comparable to above). From there, they make the link to overall population dynamics, where the intrinsic growth rate r of the population can be scaled with generation time and related to selection on z,

$$r = \frac{\ln(\bar{W})}{T} = \frac{\ln(W_{\max})}{T} - \frac{\ln(1+\sigma_z^2/\omega_z) + b^2/\omega_b}{2T} - \frac{(\bar{z}-\theta)^2}{2T(\omega_z+\sigma_z^2)}, \tag{9.4.15}$$

where the first two terms become the maximum possible growth rate $r_{\max}$ if z reaches the optimum θ.

From the expression for population dynamics, Chevin et al. derive a formula for the critical rate of environmental change, above which plasticity and adaptation cannot prevent extinction,

$$\eta_c = \sqrt{\frac{2r_{\max}\gamma}{T}}\,\frac{h^2\sigma_z^2}{|B-b|}. \tag{9.4.16}$$

From this expression, it is easy to determine the long-term tendency of the population to extinction or survival as a function only of the degree of plasticity and the associated strength of costs (ω_b). The greater the extent of plasticity, the more the costs of plasticity separate out population trajectories; but when plasticity has a weak slope, the extinction isoclines converge towards the same threshold. While this conceptualization of adaptation to environmental change as a single-trait system with readily measured costs of adaptive plasticity is obviously an idealization, it also clearly illustrates a number of important points. While adaptive genetic variation has a clear direct relationship to evolutionary rescue, plasticity also plays an important role; and quantifying plasticity without quantifying its costs can provide a misleading perspective on the feasibility of adaptation and acclimation.

Is Evolutionary Rescue Real? Evolutionary rescue is not a "silver bullet," and the application of evolutionary theory to real populations and metapopulations is far from straightforward. For one thing, evolutionary rescue requires a sufficiently large population that a species is buffered against extinction long enough for higher fitness phenotypes to become predominant [50]. Additional complications include, but are not limited to

- **Initial environmental conditions.** Bell and Gonzalez showed that populations that begin at intermediate stress levels may react the slowest to environmental "deterioration," producing a U-shaped curve in adaptive rescue [10]. They explain this as a product of two competing processes driving evolutionary rescue: as baseline stress increases, overall mutation rates decline, but the proportion of beneficial mutations (or, perhaps more accurately, the associated fitness differential) increases. Populations beginning in "mildly stressful conditions" may simply be at the low point of both processes. Bell and Gonzalez similarly show that populations with a history of minor environmental deterioration have a much greater probability of evolutionary rescue in a fast-changing environment.
- **The velocity of environmental change.** As Chevin et al.'s model highlights, environmental changes that are too rapid almost invariably drive species to extinction, when selection simply cannot operate fast enough to keep pace; this finding is readily confirmed in environmental settings. Rapid environmental changes can also functionally reduce mutation rates at a population scale. A study of *E. coli* by Lindsey et al. showed that "The evolutionary trajectory of a population evolving under conditions of strong selection and weak mutation can be envisioned as a series of steps between genotypes differing by a single mutation," and some "priming mutations" may be necessary to arrive at further genotypic combinations with substantially higher fitness [68]. Consequently, if environmental changes are too rapid, higher fitness genotypes may be "evolutionary inaccessible."
- **Dispersal rates and metapopulation connectivity.** Simulated metapopulation models by Schiffers et al. showed that higher dispersal rates can severely limit the propensity of populations to experience local adaptation, especially in a heterogeneous environment (a phenomenon they refer to as "genetic swamping"), and thereby potentially limit evolutionary rescue [93]. However, for an entire

species to persist, intermediate (local) dispersal may be necessary to allow adaptive mutations to spread, a finding shown experimentally by Bell and Gonzalez.

- **Linkage disequilibrium.** Schiffers et al.'s study, which simulated genomes in an "allelic simulation model," produced an unusual result suggesting that linkage between adaptive loci may not actually increase the rate of adaptation. The interaction this could have with the "priming mutation" process is complex and poorly explored in a theoretical context.

A final important consideration should be made with regard to what Schiffers et al. distinguish as *complete* vs. *partial evolutionary rescue.* In their models, they find that when adaptive traits originated but spread poorly (as a combination of linkage disequilibrium, habitat heterogeneity, and dispersal limitations), it substantially reduced population sizes and ultimately produced an "effective reduction in the suitable habitat niche." This type of partial evolutionary rescue could be most common in real-world scenarios, where adaptation in larger populations experiencing the slowest rates of environmental change may allow persistence but not maintain a species throughout its entire range, and may still be followed by a substantial reduction in overall habitat occupancy.

If current research on global climate change is any indication, this type of partial evolutionary rescue may ultimately be a poor buffer against extinction. Climate change may set the events of an extinction in motion, but research suggests that habitat loss from climate change is rarely the direct and solitary causal mechanism of an extinction [15]. Instead, climate change may reduce a population to small enough levels at which other mechanisms drive extinction. Small populations are especially susceptible to stochastic crashes in population size and may also be especially susceptible to stochastic collapse due to other factors within-species (Allee effects in breeding, inbreeding) or from interactions with other species (competition, invasion, disease). Ultimately, the synergy between these drivers may produce a greater overall extinction risk that many modeling approaches might not directly quantify, but that could be most likely to drive species to extinction and ecosystems into novel assemblages [8].

9.4.2 After Extinction: Lazarus Species, Romeo Errors, and the Rarest Birds in the World

The job of conservation biologists and extinction researchers is far from over after the extinction of a species. The *auto-ecology* of an extinct species (its basic biology, ecology, natural history, distribution, and other species-level characteristics) often becomes a permanent unknown, assumed to be lost to the annals of history. But as statistical tools for ecological reconstruction become more sophisticated, researchers have the opportunity to explore basic questions about extinction in retrospect. In particular, the same body of theory that governs the timing of

extinction in a declining population can be applied in a retrospective sense as well, to estimate the likely extinction date of a species. (Or, more formally, the estimation of the MTE from a given point can be used to pinpoint T_E , even with the same data, after extinction has already occurred.) These methods have been used both for ancient species like the megalodon [81] and for more recent extinctions like that of the dodo [86]. But perhaps most interestingly, the theory can be applied when the uncertainty bounds on T_E contain the present date, meaning that the extinction of a species is not taken as a certain part of history. Even ancient "Lazarus species" can be rediscovered, like the coelacanth, believed to have gone extinct 66 million years ago but rediscovered in the last century. How can we confidently say the coelacanth continues to exist, but the megalodon is likely to never be rediscovered?

Basic Statistical Methods for the Sighting Record Once a species is suspected to be extinct, at what point do we stop looking for them? With limited resources for conservation, trying to find and conserve a species that is no longer around waste resources better used elsewhere/ but making a type I error and assuming a species is falsely extinct (and abandoning conservation efforts) can lead to a "Romeo error," whereby giving up on the species can lead to actual extinction [24]. Since 1889, 351 species thought to be extinct have been "rediscovered" [92], highlighting just how big of a problem this may be. In order to answer these questions, determining the probability that a species is still extant, despite a lack of recent sightings, is an important tool in making evidence-based decisions conservation managers must make about allocating resources.

Consider the plight of the ivory-billed woodpecker *(*Campephilus principalis), a charismatic and iconic part of the North American fauna. The ivory-billed woodpecker's decline was gradual, and unlike its gregarious and easily-spotted compatriots (such as the passenger pigeon, *Ectopistes migratorius*, or the Carolina parakeet, *Conuropsis carolinensis*, both extinct in a similar time period), sightings of the woodpecker were already rare previous to its decline. So while the bird's last "credible" sighting was in 1944, the precise date of its extinction remains controversial, and some believe the bird still exists based on unverified observations as recent as 2004 (with audiovisual evidence reviewed in a highly controversial 2005 paper in *Science* [36]). These controversial observations led to one of the most costly surveys in history, yet yielded no new evidence. In some circles, the search continues; in 2016, two ornithologists—Martjan Lammertink and Tim Gallagher—traveled through Cuba searching for remaining populations of the elusive woodpecker. Was Lammertink and Gallagher's search justified from a statistical standpoint?

But how do we determine the likelihood that a species is extinct? How long does it have to be since the last time an individual was seen before we can say, with some certainty, that the species is, in fact, gone? The most obvious step is to assemble all available evidence of when the species was around. The first place to look is in the specimen record, since this is the "gold standard" of evidence. However, other data can be brought to bear, including observations, photos, and audio recordings. All these forms of evidence are collectively referred to as *sightings*. In 1993, Andrew

Solow developed an approach to resolve the extinction date of a species based on sighting records [97]. In Solow's notation, sightings in a period of surveillance between time 0 and time T occur at the dates $(t_1, t_2, ..., t_n)$ as a random process governed by a fixed sighting rate m that becomes 0 at T_E, the true date of extinction. The probability of the data conditional on a current time T and an extinction date T_E, is

$$P(T_n \leq t_n | T_E \geq T) = (t_n/T)^n. \tag{9.4.17}$$

In that light, Solow says, hypothesis testing is easy: against the null hypothesis that extinction has yet to happen (i.e., $T_E > T$), we can test the alternate hypothesis that the species is extinct ($T_E < T$). For a given last sighting at T_N, we can provide a p-value for the test with desired significance level α equivalent to

$$P(T_N \leq \alpha^{1/n} T | T_E < T) = \alpha (T/T_E)^n \tag{9.4.18}$$

for values of $\alpha^{1/n} T < T_E < T$; for values of T_E lower than or equal to that critical value $\alpha^{1/n} T$, the value of P is equal to 1 and the null hypothesis is rejected with full certainty. Solow explains, by way of example, that with 10 sightings and 95% confidence, the critical value of T_E/T is 0.74, and so the null hypothesis is sure to be rejected (extinction is confidently confirmed) if the true extinction date occurs within the first 74% of the $(0, T)$ window.

Solow similarly constructs a Bayesian approach, where the likelihood of the sighting data given H_0 is

$$\int_0^\infty m^n e^{-mT} dP(m), \tag{9.4.19}$$

and given H_A is

$$\int_0^\infty m^n e^{-mT_E} dP(m). \tag{9.4.20}$$

From these and other assumptions, he derives the Bayes factor for the hypothesis test (a metric that does not depend on prior assumptions, which expresses the posterior: prior odds of H_0),

$$B(t) = (n-1) / \left((T/t_n)^{n-1} - 1 \right). \tag{9.4.21}$$

Finally, from the original formulation of Solow's approach, we can also derive a maximum likelihood estimate of the extinction date [98],

$$\hat{T_E} = \frac{n+1}{n} t_n, \tag{9.4.22}$$

and, in addition, a $1 - \alpha$ upper confidence interval bound,

$$T_E^u = t_n / \alpha^{1/n}. \tag{9.4.23}$$

Does this approach make sense? If an extinction happens abruptly on the scale of sightings data (say, an epidemic wipes a species out within a year), then sighting rates might remain relatively constant throughout the sighting record. Similarly, applying this method to paleontological records may make sense, as prior information about variation in specimen preservation might be limited (and so a constant rate parameter is the best possible prior). But there are also a number of situations where the constant sighting rate m simply does not suffice. Lessons from population ecology remind us that extinction is, at its most fundamental scale, a process of declining abundance. If sightings are dependent on abundance (which they generally are), replacing m with a non-constant function has the potential to sharply refine the process of extinction date estimation.

Similarly, not all sightings are created equally. If you are holding a dead body of an individual of the species in question, that is good evidence the species was present the year the specimen was collected. Conversely, if some person claims they saw an extremely rare species with no corroborating evidence, that person may have misidentified the individual, or in some cases even lied, meaning that this sighting could be invalid. Roberts et al. found that these approaches are sensitive to the data used and can, unsurprisingly, lead to very different estimates of extinction dates [87]. They partitioned sighting data into three categories: (1) physical evidence, (2) independent expert opinion, and (3) controversial sightings in order of certainty. They found that adding independently verified observations to the analysis can sometimes lead to earlier predicted extinction times, since the "gaps" within the sighting record are closed up, whereas, by nature, later controversial sightings, if treated as legitimate (i.e., on par with physical evidence), can greatly push the estimates of extinction to later years. To account for this uncertainty, a few approaches have been proposed recently. These approaches largely expand on Solow's 1993 Bayesian equation above, modified to consider multiple levels of uncertainty in the data [64, 99, 103]. For an overview of the assumptions and relative strengths of these approaches, see Boakes et al. [11].

Finally, some nonparametric approaches to extinction date estimation focus on the last few sightings of a species, rather than the entire record of their observations. Solow [98] notes two such methods in a review that covers these methods of estimations in much greater depth. The first, originally suggested by Robson and Whitlock [88], just uses the last two sightings,

$$T_E = t_n + (t_n - t_{n-1}), \tag{9.4.24}$$

with a fairly clear reasoning: if a large gap exists between the last two sightings, conservation biologists should wait at least that long before pronouncing a species certain to be extinct.

In contrast, the second and far more complex method designed by Solow (and implemented by Roberts and Solow in their 2003 study of the dodo [86]) accounts for the fact that the last few sightings of the species should, in most circumstances, follow a Weibull distribution. The method, *optimal linear estimation* (OLE), estimates T_E through linear algebra,

$$T_E = \sum_{i=1}^{k} w_i t_{n-i+1}, \text{ where } w = (e' \Lambda^{-1} e)^{-1} \Lambda^{-1} e. \tag{9.4.25}$$

Here, e is a column vector consisting of k 1's and Λ is a $k \times k$ matrix with elements

$$\Lambda_{ij} = \frac{\Gamma(2\hat{v} + i)\Gamma(\hat{v} + j)}{\Gamma(\hat{v} + i)\Gamma(j)}, \text{ where } \hat{v} = \frac{1}{k-1} \sum_{i=1}^{k-2} \ln \frac{t_n - t_{n-k+1}}{t_n - t_{i+1}}. \tag{9.4.26}$$

While the OLE method is obviously much less transparent, it has been recorded as one of the most successful methods available for predicting extinction [23], and has the added bonus of being adjustable through sensitivity analysis to examine how different extent of sighting data changes the overall estimate.

Case Study: Spix's Macaw Perhaps the most fruitful body of research concerning extinction date estimation has been within ornithology, where data on the last sightings of rare species are often more available than for other groups, due to tremendous global interest in bird sightings and observation by non-scientists. The most popular methods for sighting date research have often been developed in association with data on notable extinct birds, including the dodo, the passenger pigeon, and the ivory-billed woodpecker. In fact, one of the most expansive reviews of sighting date estimators, conducted by Elphick, estimated the extinction date of 38 extinct or near-extinct birds from North America (including Hawaii, a hotspot of bird extinction) [34]. But for rarer birds around the world, basic data on their extinction may be somewhat more lacking.

One such bird, the Spix's macaw (*Cyanopsitta spixii*) has been called "the world's rarest bird" [56] and has been the subject of two popular animated movies (Rio and Rio 2). Currently, Spix's macaw is considered critically endangered (possibly extinct in the wild) by the IUCN (2016), with a small number of captive individuals ($\sim$130) found around the world. Not seen in the wild since 2000, a video of a Spix's macaw in Brazil made headlines in 2016. The video was subsequently examined by ornithologists, and the consensus that the bird was, in fact, a Spix's macaw, though many still believe the bird was likely an escaped captive bird.

Sightings of the Spix's macaw are sporadic, and after the first known specimen being shot in 1819 by Johann Baptist Ritter von Spix (though he believed the bird to be a Hyacinth Macaw), it was not recorded again until a wild-caught individual was procured by the Zoological Society of London in 1878. Collecting sighting records of the Spix's macaw relies mostly on data from trappers/poachers and inferring

data from captive individuals. Given the illicit nature of wildlife poaching, better data may exist in the husbandry records of the wild-caught individuals currently in captivity, but those data are not freely available. Verifiable observations are few and far between, as this species was not subject to any intensive study or searches until the mid-1980s, when only a handful of individuals were found and, of those remaining, most were caught by poachers.

For this case study, we collected sighting and specimen data from GBIF (Global Biodiversity Information Facility; www.gbif.org) and Juniper's authoritative book on Spix's macaw. We found physical evidence (specimens and wild-caught captive birds) for sightings in the years 1819, 1878, 1884, 1901, 1928, 1954, 1977, 1984, 1985, 1986, and 1987. Due to their rarity and the demand for them, we assumed individuals were caught in the wild the same year they were procured by the receiving institution or zoo. We considered all observations of the Spix's macaw reported in Juniper's book as verified, as there aren't many and these few have been rigorously scrutinized: 1903, 1927, 1974, 1989, 1990, and 2000. Our only controversial sighting is the recent video taken in 2016. Taking the approach by Roberts et al. we partitioned the data into three datasets: (1) physical data only, (2) physical plus verifiable observation data, and (3) all data (including the controversial sighting). By eliminating the controversial sighting (in analyses 1 and 2), we inherently test a methodological question: would extinction date estimators have pronounced the apparently extant species dead?

Our analysis was conducted using the beta version of the `R` package `sExtinct`, which allows a handful of different extinction analyses to be implemented. (We encourage prospective users to test the demos available with the package.) Our analysis uses two of the most common methods. First, we used the original Solow maximum likelihood approach, plotting the probability of persistence in Fig. 9.2. The maximum likelihood estimates are given in that method as

- Specimens only: $T_E = 2040$,
- Uncontroversial sightings: $T_E = 2035$,
- All sightings: $T_E = 2052$.

The method suggests, even with the most limited dataset, that the species still appears to exist. In contrast, the OLE method tells a different story:

- Specimens only: $T_E = 1988$ (95% CI: 1987–2006),
- Uncontroversial sightings: $T_E = 2002$ (95% CI: 2000–2018),
- All sightings: $T_E = 2021$ (95% CI: 2016–2045).

All things considered, both analyses suggest a chance the 2016 sighting may have been legitimate, and there is a possibility that a wild population of Spix's macaws may be out there, yet undiscovered in the Amazon rainforest. But, the OLE method—for all its documented strength as an approach—would likely have been far hastier to dismiss the species as extinct before its 2016 "rediscovery." Furthermore, it currently only predicts another 5 years of persistence for the species, and with some researchers hoping to use extinction date estimators as a method of Red Listing, the Spix's macaw clearly remains a severely threatened species.

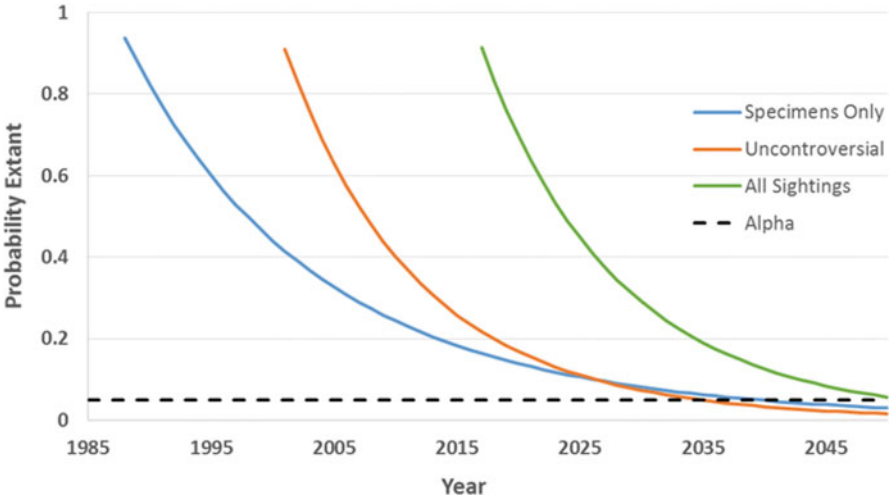


Fig. 9.2 Estimates of likely extinction date of the Spix's macaw based on extinction estimating equations in [97]. The lines represent the estimated probability the species is extant each year; the blue line is the results using physical evidence only (specimens/wild-caught individuals), the orange line for uncontroversial sightings *and* physical evidence, and the green line is the results for all sightings, including controversial. The dotted line is a significance level of 0.05. Once the probability drops below this level, the species is considered likely extinct

Hope Springs Eternal: The Ivory-Billed Woodpecker and the Hunt for More Lazarus Species To briefly reconsider Lammertink and Gallagher's continuing search for the ivory-billed woodpecker: regardless of how the sighting record for the ivory-billed woodpecker is analyzed, all indications point to an extremely low likelihood that the species is extant [34, 44, 99]. In the work of Elphick et al., estimates based on physical evidence suggested a T_E of 1941 (upper 95% CI: 1945) and including expert opinion sightings only moves T_E towards 1945 (upper 95% CI: 1948). With other models hardly disagreeing on the scale of a full century, the hard evidence available to modelers casts serious doubts on the validity of the species' "rediscovery" in 2004 [95], or further, justifies the subsequent, costly search to find more conclusive evidence of the ivory-billed woodpecker's existence. Some argue the search continues as long as hope does, but statistics has a somewhat different answer in this case. And with other species like the Spix's macaw still potentially within the bounds of rescue, the resources of conservation organizations might be better devoted to saving those species than to chasing the ghosts of woodpeckers past.

Once it is determined that there is an acceptable level of probability that a species is extant, one possible way to further leverage the data collected would be use the data to build species distribution models (SDMs) to aid in the search and rescue effort. In basic terms, SDMs use information about the conditions where a species has occurred (and where it has not occurred) to determine the realized

niche of the species. This niche can be projected onto geographic space to help identify areas that appear highly suitable for the species but perhaps have not been searched yet. This approach has been successful in identifying new populations of threatened species (e.g., see [74]), with the author identifying new populations of four of the eight rare plant species in the study. While SDMs are commonly used in a variety of different ecological and conservation applications, there is a deep literature on comparisons of SDM methods (see Qiao et al. [83] for an overview), so much caution must be used in selecting which methods are best for the available occurrence and environmental data. This approach—of determining the probability a species is still extant and using SDMs to identify the areas they are most likely to be—may provide a way forward for conservation agencies for making cost-effective decisions of which species to pursue and where to look for them.

9.5 The Community Scale and Beyond

Suppose that, in a twisted experiment motivated by an ecology-related childhood trauma, a mad scientist was developing a scheme to reduce global biodiversity to one half of the Earth's total species. Hunting, fishing, and poaching could achieve that goal slowly but would be particularly inefficient for eradicating insects; and while a generalist disease might help eradicate a handful of mammals or a sizable fraction of amphibians, the majority of species would still remain. But perhaps realizing that habitat loss might be the most efficient tool to destruction, the mad scientist might cut the Gordian knot by simply bisecting the Earth and destroying one half. Would his plan come to fruition?

Our mad scientist's plan is riddled with flaws. If one half of the species were endemic to each half of the Earth with no overlap, his plan would succeed. But a handful of species in any clade of life are globally cosmopolitan; and no matter how his plan was executed, the handful of species occurring on both halves of the Earth would leave him with far, far more than half the species he started with.

With renewed vigor, the mad scientist sets out on a newly ambitious project: what percentage scorched earth would be required to achieve his goal? He begins by counting every species on his sidewalk block, then in his neighborhood, and up to bigger scales. With enough grant funding and undergraduate assistants, he has eventually covered a measly 6.25% of the Earth when he realizes he has counted half of Earth's species. To enact his master plan, he's tasked with destroying the remaining 93.75%. Going by land area alone (his grudges, we suppose, do not extend to the ocean), he only needs preserve 3.6 million square miles of land—roughly (conveniently?) the land area of the United States.

The process our nationalist, isolationist villain has enacted is the empirical construction of the species–area relationship (SAR), one of the oldest and most powerful scaling laws in macroecology. Because the synthesis of different factors at global scales is challenging, and habitat loss is one of the easiest extinction drivers to measure, the SAR gives us a powerful tool for approximating extinction rates—at the price of not knowing specifically which species will go extinct.

9.5.1 The Species–Area Relationship

The biogeographer Olof Arrhenius began the process of formalizing the SAR in a classic 1921 paper entitled "Species and Area" in the *Journal of Ecology* [6]. Arrhenius observed that, by expanding the area of focus, the number of species continues to increase at a diminishing rate (but, never reaching an asymptote [106]). The canonical formula for the SAR has been called the Arrhenius SAR and is formulated as

$$S = cA^z,$$

where c is a constant fit to the data and z is a slope, conventionally taken as 0.25. The application of this formula to extinction rate estimation is relatively obvious; by changing the amount of area, we can change the number of species,

$$S' = c(A')^z,$$

and calculate the number of extinctions

$$E(A') = S - S'.$$

In our mad scientist's failed scheme, reducing the area of the Earth by half would leave us with far more than half the species,

$$\frac{S'}{S} = \left(\frac{0.5A}{A}\right)^{0.25} = (0.5)^{0.25} = 0.84.$$

In a 2004 *Nature* paper that has become the most cited study on extinction since the millennium, a group of researchers led by Chris Thomas refined the global extinction rate estimate by analyzing species' habitat losses from climate change and applying the SAR. Their extinction–area relationship took three forms applied to n species, with a given A_i area per species before change, and A_i' subsequent to habitat loss,

$$E_1 = 1 - \left(\frac{\sum\limits_{i\in(1,n)} A_i'}{\sum\limits_{i\in(1,n)} A_i}\right)^{0.25},$$

$$E_2 = 1 - \left(\frac{1}{n}\sum_{i\in(1,n)} \frac{A_i'}{A_i}\right)^{0.25},$$

$$E_3 = \frac{1}{n}\sum_{i\in(1,n)} \left(1 - \left(\frac{A_i'}{A_i}\right)^{0.25}\right).$$

Using those three methods in combination with species distribution models, the authors estimated that 15–37% of species on Earth might face climate-driven extinction by 2050. This result is by far one of the most important ones produced in any study of extinction and has supported a number of the most expansive conservation programs worldwide.

9.5.2 Everything You Know About the Species–Area Relationship Is Wrong

Like many "laws" of ecology, the conventional SAR has problems and pitfalls, and with the tremendous array of approaches developed to study it, it has even been called ecology's "most protean pattern" [69]. Subsequent to the publication of Thomas *et al.*'s study, one of the most seminal debates in extinction research has centered around its conclusion that climate change is likely to act as the most consequential driver of the sixth mass extinction. Different approaches to the species–area relationship and comparable or derivative macroecological methods have sprung up in the wake of Thomas's work. Here, we review a few of the different approaches that can be used to predict extinction rates at the community level.

z: A Dynamic Scaling Property The most immediate problem with applying the species–area relationship is that the slope z, normally set to 0.25, is neither universal nor scale-independent. In part, this is because of two different constructions of the SAR. The slope of 0.25 derives from the experimental work of Macarthur and Wilson on island ecosystems, which applied the SAR to the richness of species on islands of different sizes. For islands (and for application of the island SAR to extinction), a slope of 0.25 is justified under a set of three (relatively common) circumstances delineated by Harte and Kitzes: "(1) total abundance in the new area A is proportional to area, (2) individuals found in A are chosen by a random draw from all individuals in A_0, and (3) the number of individuals of each species in A_0 follows a canonical lognormal abundance distribution" [51].

However, the continental "nested" SAR (constructed from nested areas on a continental scale) does not always follow the same property. This is in part because the conventionally used SAR assumes self-similarity (or, in more tangible terms, picking two patches of different area always yields a roughly-the-same-slope difference in species). As it turns out, self-similarity works within some sites but not others, and within the Western Ghats mountains of India alone, scaling up from vegetation sampling plots to broader scales brings z down from values closer to 0.5 to values approaching 0 [52]. Selecting an appropriate slope based on scale is an important part of appropriate use of the SAR to predict extinction rates, and as analyses approach the continental scale, the appropriateness of the SAR method decreases as z approaches zero.

Alternate Approach Based on the Endemics–Area Relationship In the Thomas et al. study, the application of the species–area relationship followed three methods, and while some explicitly predicted extinction risk at the scale of a single species, all rely on the prediction of reduced species richness based on habitat loss. In place of this indirect calculation of decreased richness, a more direct approach uses what is called the *endemics–area relationship* (EAR), which calculates the number of endemic species restricted to a given area (all of which should be committed to extinction when the area is destroyed). As pointed out by He and Hubbell, the SAR and the EAR are not mirror curves except in a single special case when species are completely randomly distributed in space; else, the "forwards" and "backwards" methods of extinction calculation are not, they argue, comparable [53].

Prediction of extinction based on the EAR may be more appropriate for measuring the immediate effects of habitat loss and is likely to better account for the "geometry of habitat clearing" [58]. Storch et al. [100] developed an approach to the SAR and the EAR that scales the area by the mean geographic range size in the focal clade/area and scales richness by the average number of species in that mean geographic range. When plotted, the SAR curves upwards while the EAR is roughly linear with a slope of 1 across most scales. Starting from basic knowledge about the average geographic range size of a given species, this result indicates that extinction from habitat loss can be predicted based on the EAR across scales fairly accurately.

Alternate Approach Based on Maximum Entropy Two "unifying" theories have dominated discussions about macroecology. The first is the unified neutral theory (UNT) of biogeography and ecology (proposed by Stephen Hubbell), which is beyond the scope of this chapter; the second is the maximum entropy theory of ecology (METE) proposed by John Harte. The METE deserves special mention here, due to a particular focus in the METE literature on improving the applicability of the SAR to extinction rate prediction. What differentiates both the UNT and the METE from more general conceptions of the SAR is the explicit treatment of species abundance as a component of community assembly. The theory of the METE is far too complex to encapsulate in this chapter (and an entire book by Harte exists for that purpose), but a few useful derivations are worth mentioning. One is the derivation by Kitzes and Harte of an extinction probability that is applicable at the species scale [51] based on proportional area loss (A_0/A, shortened to β) and corresponding reduction in abundance (n from n_0) with a general probability distribution

$$P(n|n_0, A_0, A) = ce^{-\lambda n}, \tag{9.5.1}$$

for which they provide rough approximations,

$$c \approx \frac{1}{(An_0/A_0) + 1}, \quad \lambda \approx \ln\left(1 + \frac{A_0}{An_0}\right). \tag{9.5.2}$$

Drawing on similar concepts from the pseudo-extinction thresholds we discuss above in Sect. 9.4.1, they suggest that the probability that a remainder $r_c = n/n_0$ is left after habitat loss is

$$\text{Prob}\left[\frac{n}{n_0} > r_c\right] = \int_{r_c n_0}^{n_0} c e^{-\lambda n}\, dn = \frac{[n_0\beta/(1+n_0\beta)]^{r_c n_0} - [n_0\beta/(1+n_0\beta)]^{n_0}}{(1+n_0\beta)\ln(1+1/n_0\beta)}. \tag{9.5.3}$$

Given a starting population and a critical population size, analogous results can be derived for the Thomas et al. calculations; higher level predictions can be made based on the distribution of abundances and critical abundances within the community.

In a subsequent publication [59], this *extinction–area relationship* is extended even further to extrapolate a MaxEnt-based probability that a given number of species will remain after habitat loss. It assumes a log-series distribution ϕ of abundance for species with a mean μ_ϕ, with a single shape parameter p,

$$\phi(n_0) = \frac{-p^{n_0}}{\ln(1-p)n_0}, \quad \mu_\phi = \frac{-p}{(1-p)\ln(1-p)}. \tag{9.5.4}$$

They similarly propose an upper-truncated geometric species-specific abundance distribution, which provides the probability that n individuals remain in a fractionally reduced area a (β in their other notation) based on a shape parameter q,

$$\Pi(n|a, n_0) = \frac{(1-q)q^n}{1-q^{n_0+1}}, \tag{9.5.5}$$

where q is solved implicitly based on a and n_0 from the equation

$$an_0 = \frac{q}{1-q} - \frac{(n_0+1)q^{n_0+1}}{1-q^{n_0+1}}. \tag{9.5.6}$$

The probability that a species is found in area A after habitat loss follows a distribution g which takes the form

$$g(a, n_c) = \sum_{n_0=1}^{\infty} (1 - \Pi(n \leq n_c|a, n_0))\phi(n_0), \tag{9.5.7}$$

which scales up to a community-level richness after area loss,

$$p(S|S_0, g) = \binom{S_0}{S} g^S (1-g)^{S_0-S}, \tag{9.5.8}$$

where

$$g(a, n_c, \mu_\phi) = \sum_{n_0=1}^{\infty} \left(1 - \frac{q^{n_c+1} - 1}{q^{n_0+1} - 1} \frac{-p^{n_0}}{n_0 \ln(1-p)}\right) \tag{9.5.9}$$

or, if the pseudo-extinction threshold is set to zero (i.e., no species has 0% survival odds until all individuals are dead) and area loss is severe, this expression can be reduced to eliminate the q term,

$$g(a, n_c, \mu_\phi) = -\frac{a}{\ln(1-p)} \sum_{n_0=1}^{\infty} \frac{p^{n_0}}{an_0 + 1}. \tag{9.5.10}$$

This METE approach thus provides a *probabilistic species–area relationship* (PSAR) that can be used to provide not only an expected extinction rate under habitat loss but also a range of confidence. This becomes an especially important tool in a small community of only a few dozen species or fewer (or in communities with pervasive low abundance across species), where deviations from SAR-based predictions may be greater due to stochastic processes.

How does the PSAR scale up against the Thomas-SAR? It has a clear advantage in the prediction of individual species extinction risk (but correspondingly requires more data on abundance/demography that may be absent for many poorly known taxa). Kitzes and Harte provide two illustrations. First, assuming the normal slope of 0.25, the PSAR predicts a 44% chance of extinction for a species that loses 90% of its habitat. Second, if we assume a pseudo-extinction threshold of 50 individuals, the Thomas-SAR under-predicts the extinction risk if n_0 is less than 1000 but over-predicts otherwise.

Tying Up Loose Threads, Thinking Across Scales The various different approaches to predicting extinction at the broadest scales have driven substantial controversy among different interpretations of macroecological theory. But one of the most important problems is that estimates of extinction from these methods are still poorly connected, by and large, to the rest of the extinction literature—and to the other types of models we discuss above. One of the most innovative and unusual approaches in the literature was presented by Rybicki and Hanski [90], who simulated a stochastic patch occupancy model (similar to those presented in Sect. 9.3.1) with spatially heterogeneous environmental conditions across patches. While their model incorporates the standard mainstays of an SPOM (colonization, extinction, a dispersal kernel), it also incorporates a phenotype and niche breadth that produce a Gaussian fitness function (like many of the models discussed in Sect. 9.4.1).

Tying together a number of the important ideas discussed above, the work of Rybicki and Hanksii made several advances into new territory. For one, they make a semantic distinction between the EAR (which they define as the $S = cA^z$ relationship applied to the area lost a) and the "remaining species–area relationship" (RAR),

$$S - S_{\text{loss}} = c(A_{\text{new}}/A)^z. \tag{9.5.11}$$

The EAR and RAR, as two methods of calculating extinction risk, are not interchangeable or symmetric counterparts. Rybicki and Hanskii highlight a discrepancy between Storch et al.'s suggested EAR slope of roughly 1, and He and Hubbell's values which were a tenth smaller [53], which they suggest can be resolved by the fact that Storch fit the EAR while He and Hubbell were calculating the RAR. Their simulations agree with the results of He and Hubbell that the slope of the RAR may be half or less that of the SAR.

Their empirical approach to simulation leads to a valuable conclusion that stands in opposition to previous work. While Kinzig and Harte [58] and He and Hubbell [53] both strongly suggest that the SAR over-estimates extinction risk, the results of Rybicki and Hanskii's simulations suggest that in the short term, the RAR under-estimates extinction while the continental SAR ($z \approx 0.1$) is adequate. Their result ties the population scale to the community scale, as they attribute it to species' populations *outside* destroyed or fragmented habitat falling below critical thresholds and facing extinction despite the lack of total endemic extirpation. In the long term, they suggest, the island SAR ($z = 0.25$) may be the best predictor of total losses. Finally, they explore the difference between leaving a single patch of habitat and fragmenting habitat and conclude all models underestimate extinction risk in scenarios of extreme fragmentation. To address that problem, they propose a modified species–area relationship

$$S = cA^z e^{-b/\lambda_M}, \tag{9.5.12}$$

where λ_M is the metapopulation capacity (see Sect. 9.3.1) and b is another scaling parameter like c and z. If n is the number of habitat fragments, they suggest, the metapopulation capacity scales linearly with A^3/n^2, meaning that the *fragmented landscape species–area relationship* (FL-SAR) can be expressed as

$$S_{\text{new}}/S = (A_{\text{new}}/A)^2 e^{-bn^2/A^3}. \tag{9.5.13}$$

While the data to fit such an expression might be challenging to collect (and so the FL-SAR may not be an immediately useful conservation planning tool), the FL-SAR provides an important and much needed link between the population processes we discuss above and our broader understanding of the rate of extinction at landscape and community scales.

9.6 Last Chance to See

What don't we know about extinction yet?

As predictive tools gain precision, our estimates of the extinction rates of well-known groups like mammals and birds also become more precise. But the majority of the world's species are not yet known; most animal diversity is harbored by insects or parasites (especially nematodes), and the vast majority of species in those groups are undiscovered or undescribed. Their extinction rates are just as poorly quantified as their diversity, average range size or abundance distribution, or the hotspots of their biodiversity. But some basic estimates suggest that 7% of the planet's invertebrates may have already gone extinct—at which rate evidence would suggest that 98% of extinctions on Earth are currently going undetected [85]. It is also especially difficult to compare these extinction rates to historical baselines, because the fossil record for most invertebrates and other taxa are incomplete or nearly absent.

An especially poignant problem is the detection and estimation of coextinction rates—the secondary extinction of species dependent on others for their ecological niche—which Jared Diamond suggested in 1989 was one of the four horsemen of mass extinction (in his words, "overhunting, effects of introduced species, habitat destruction, and secondary ripple effects") [26]. Among the most obvious candidates for coextinction are two main groups: pollinators (which can have a strict dependency on host plants) and endosymbionts (parasites and mutualists, which may exhibit strict specificity in their association with plant or animal hosts). While both groups are believed to be severely at risk of secondary extinction, quantifying their extinction rate can be challenging, as there is rarely a 1:1 correspondence between hosts and dependent species. An approach popularized by Koh simulates host extinctions in a random order and predicts the number of corresponding coextinctions from the *affiliation matrix*; by fitting a function to real affiliation matrices, Koh et al. found that if host specificity is 1:1 then the slope is linear, but when affiliates use a greater number of hosts, the coextinction function is concave upward,

$$\bar{A} = (0.35\bar{E} - 0.43)\bar{E}\ln\bar{s} + \bar{E}, \tag{9.6.1}$$

where E gives primary extinction risk, A secondary extinction risk, and s is host specificity [60]. Subsequent work has shown that even though parasites and mutualists may experience a reduced rate of extinction from host switching, the majority of threatened species on Earth might still be mutualists and parasites (due to the tremendous diversity of such species, e.g., the estimated 300,000 species of helminth alone [27]), and most of those extinctions are poorly cataloged [31]. More data are needed on host-symbiont association networks to better inform the role that nonrandom structure in those networks might play in increasing or decreasing extinction rates; some work has suggested that species preferentially favor more stable host species, the underlying cause of a "paradox of missing co-extinctions" [101]. Similarly, the potential for species to switch hosts and thereby avoid extinction is unknown, but likely mitigates global extinction risk. In parasitology, the Stockholm Paradigm suggests that host-parasite associations diversify in changing climates and environments as a function of (1) phenotypic plasticity,

(2) trait integration, and (3) phylogenetic conservatism of "latent potential" which together produce a pattern of *ecological fitting* that might benefit parasites (and thereby other symbionts) in the face of the sixth mass extinction [16]. A more in-depth treatment of the theoretical ecology of ecological fitting can be found in the recent work of Araujo et al. [5].

Is saving microbes and parasites from extinction a reasonable goal? Some argue that it is [28], but others have recently suggested it's "time to get real about conservation" and focus on the fact we're not adequately preventing catastrophic population crashes in megafauna like elephants [33] or giraffes. Regardless of animal type or conservation status, the development of demographic theory and predictive modeling are our best options to understand and mitigate extinction risk in natural populations. One such advance is the development of *early warning signals* of population collapse. This is a developing body of literature that is built around the fact that populations on the verge of collapse often produce detectable statistical signals [91]. If researchers are able to detect these signals in time series data before it is too late, mitigation efforts and prioritization of at-risk populations may prevent population collapse. Current work is attempting to scale the detection of early warning signals to the metapopulation level by developing spatial early warning signals [57], which could be used to optimize reserve design and address the influence of dispersal, stochasticity, and local population dynamics on metapopulation persistence.

The pressure for more accurate, predictive tools will only grow in the next few decades of research. A recent review by Mark Urban surveyed studies of climate change-driven extinction risk and found that, despite the variation between different modeling methods and scopes, projected extinction rates are not only rising but one in six species might be imminently threatened with extinction [104]. Similarly, in a study of roughly 1000 species of plants and animals, about half had experienced population extinctions driven by climate change [105]. As extinction rates accelerate due to global change and we fully enter the sixth mass extinction, the need for better analytical and simulation tools—that produce precise estimates from limited data—will only grow. In light of the constant need to test, revise, and re-test models of extinction, to a mathematically trained ecologist or an ecologically minded mathematician, this field of research is a critical opportunity to apply the principles of ecosystem science towards a high-impact and worthy goal.

References

1. Abbott, K.C.: Does the pattern of population synchrony through space reveal if the Moran effect is acting? Oikos **116**(6), 903–912 (2007)
2. Abbott, K.C.: A dispersal-induced paradox: synchrony and stability in stochastic metapopulations. Ecol. Lett. **14**(11), 1158–1169 (2011)
3. Allendorf, F.W., Ryman, N.: The role of genetics in population viability analysis. In: Population Viability Analysis, pp. 50–85. University of Chicago Press, Chicago (2002)

4. Altizer, S., Harvell, D., Friedle, E.: Rapid evolutionary dynamics and disease threats to biodiversity. Trends Ecol. Evol. **18**(11), 589–596 (2003)
5. Araujo, S.B., Braga, M.P., Brooks, D.R., et al.: Understanding host-switching by ecological fitting. PLoS One **10**(10), e0139, 225 (2015)
6. Arrhenius, O.: Species and area. J. Ecol. **9**(1), 95–99 (1921)
7. Barnosky, A.D., Matzke, N., Tomiya, S., et al.: Has the earth's sixth mass extinction already arrived? Nature **471**(7336), 51–57 (2011)
8. Bartlett, L.J., Newbold, T., Purves, D.W., et al.: Synergistic impacts of habitat loss and fragmentation on model ecosystems. Proc. R. Soc. B **283**(1839), 20161, 027 (2016)
9. Beissinger, S.R.: Population viability analysis: past, present, future. In: Population Viability Analysis, pp. 5–17. University of Chicago Press, Chicago (2002)
10. Bell, G., Gonzalez, A.: Adaptation and evolutionary rescue in metapopulations experiencing environmental deterioration. Science **332**(6035), 1327–1330 (2011)
11. Boakes, E.H., Rout, T.M., Collen, B.: Inferring species extinction: the use of sighting records. Methods Ecol. Evol. **6**(6), 678–687 (2015)
12. Bonsall, M.B., Hastings, A.: Demographic and environmental stochasticity in predator–prey metapopulation dynamics. J. Anim. Ecol. **73**(6), 1043–1055 (2004)
13. Boyce, M.S.: Population growth with stochastic fluctuations in the life table. Theor. Popul. Biol. **12**(3), 366–373 (1977)
14. Brook, B.W., Traill, L.W., Bradshaw, C.J.: Minimum viable population sizes and global extinction risk are unrelated. Ecol. Lett. **9**(4), 375–382 (2006)
15. Brook, B.W., Sodhi, N.S., Bradshaw, C.J.: Synergies among extinction drivers under global change. Trends Ecol. Evol. **23**(8), 453–460 (2008)
16. Brooks, D.R., Hoberg, E.P.: How will global climate change affect parasite–host assemblages? Trends Parasitol. **23**(12), 571–574 (2007)
17. Bull, J.C., Pickup, N.J., Pickett, B., et al.: Metapopulation extinction risk is increased by environmental stochasticity and assemblage complexity. Proc. R. Soc. Lond. B Biol. Sci. **274**(1606), 87–96 (2007)
18. Caswell, H.: Matrix Population Models. Wiley, Hoboken (2001)
19. Caughley, G.: Directions in conservation biology. J. Anim. Ecol. **63**, 215–244 (1994)
20. Ceballos, G., Ehrlich, P.R., Barnosky, A.D., et al.: Accelerated modern human–induced species losses: entering the sixth mass extinction. Sci. Adv. **1**(5), e1400, 253 (2015)
21. Charlesworth, B., Charlesworth, D.: The Evolutionary Effects of Finite Population Size: Basic Theory, vol. 42, chap. 5, pp. 195–244. Roberts and Company Publishers, Englewood (2012)
22. Chevin, L.M., Lande, R., Mace, G.M.: Adaptation, plasticity, and extinction in a changing environment: towards a predictive theory. PLoS Biol. **8**(4), e1000, 357 (2010)
23. Clements, C.F., Worsfold, N.T., Warren, P.H., et al.: Experimentally testing the accuracy of an extinction estimator: Solow's optimal linear estimation model. J. Anim. Ecol. **82**(2), 345–354 (2013)
24. Collar, N.: Extinction by assumption; or, the Romeo error on Cebu. Oryx **32**(4), 239–244 (1998)
25. De Castro, F., Bolker, B.: Mechanisms of disease-induced extinction. Ecol. Lett. **8**(1), 117–126 (2005)
26. Diamond, J.M., Ashmole, N., Purves, P.: The present, past and future of human-caused extinctions [and discussion]. Philos. Trans. R. Soc., B: Biol. Sci. **325**(1228), 469–477 (1989)
27. Dobson, A., Lafferty, K.D., Kuris, A.M., et al.: Homage to Linnaeus: how many parasites? How many hosts? Proc. Natl. Acad. Sci. **105**(Supplement 1), 11482–11489 (2008)
28. Dougherty, E.R., Carlson, C.J., Bueno, V.M., et al.: Paradigms for parasite conservation. Conserv. Biol. **30**, 724–733 (2015)
29. Drake, J.M.: Extinction times in experimental populations. Ecology **87**(9), 2215–2220 (2006)
30. Drake, J.M.: Tail probabilities of extinction time in a large number of experimental populations. Ecology **95**(5), 1119–1126 (2014)
31. Dunn, R.R., Harris, N.C., Colwell, R.K., et al.: The sixth mass coextinction: are most endangered species parasites and mutualists? Proc. R. Soc. Lond. B Biol. Sci. **276**(1670), 3037–3045 (2009)

32. Elías-Wolff, F., Eriksson, A., Manica, A., et al.: How Levins' dynamics emerges from a Ricker metapopulation model. Theor. Ecol. **2**(9), 173–183 (2016)
33. Ellison, A.M.: It's time to get real about conservation. Nature **538**(7624), 141 (2016)
34. Elphick, C.S., Roberts, D.L., Reed, J.M.: Estimated dates of recent extinctions for North American and Hawaiian birds. Biol. Conserv. **143**(3), 617–624 (2010)
35. Fisher, M.C., Henk, D.A., Briggs, C.J., et al.: Emerging fungal threats to animal, plant and ecosystem health. Nature **484**(7393), 186–194 (2012)
36. Fitzpatrick, J.W., Lammertink, M., Luneau, M.D., et al.: Ivory-billed woodpecker (Campephilus principalis) persists in continental North America. Science **308**(5727), 1460–1462 (2005)
37. Frank, S., Schmid-Hempel, P.: Mechanisms of pathogenesis and the evolution of parasite virulence. J. Evol. Biol. **21**(2), 396–404 (2008)
38. Getz, W.M.: A hypothesis regarding the abruptness of density dependence and the growth rate of populations. Ecology **77**(7), 2014–2026 (1996)
39. Getz, W.M., Haight, R.G.: Population Harvesting: Demographic Models of Fish, Forest, and Animal Resources, vol. 27. Princeton University Press, Princeton (1989)
40. Getz, W.M., Muellerklein, O.C., Salter, R.M., et al.: A web app for population viability and harvesting analyses. Nat. Resour. Model. **30**, e12120 (2016)
41. Gilarranz, L.J., Bascompte, J.: Spatial network structure and metapopulation persistence. J. Theor. Biol. **297**, 11–16 (2012)
42. Gilpin, M.E., Soulé M.E.: Minimum viable populations: processes of extinction. Conservation Biology: The Science of Scarcity and Diversity, pp. 19–34. Sinauer Associates Inc., Sunderland (1986)
43. Gomulkiewicz, R., Holt, R.D.: When does evolution by natural selection prevent extinction? Evolution **49**(1), 201–207 (1995)
44. Gotelli, N.J., Chao, A., Colwell, R.K., et al.: Specimen-based modeling, stopping rules, and the extinction of the ivory-billed woodpecker. Conserv. Biol. **26**(1), 47–56 (2012)
45. Gouveia, A.R., Bjørnstad, O.N., Tkadlec, E.: Dissecting geographic variation in population synchrony using the common vole in central Europe as a test bed. Ecol. Evol. **6**(1), 212–218 (2016)
46. Grilli, J., Barabás, G., Allesina, S.: Metapopulation persistence in random fragmented landscapes. PLoS Comput. Biol. **11**(5), e1004, 251 (2015)
47. Hanski, I.: Single-species metapopulation dynamics: concepts, models and observations. Biol. J. Linn. Soc. **42**(1-2), 17–38 (1991)
48. Hanski, I.: Metapopulation dynamics. Nature **396**(6706), 41–49 (1998)
49. Hanski, I., Ovaskainen, O.: The metapopulation capacity of a fragmented landscape. Nature **404**(6779), 755–758 (2000)
50. Hao, Y.Q., Brockhurst, M.A., Petchey, O.L., et al.: Evolutionary rescue can be impeded by temporary environmental amelioration. Ecol. Lett. **18**(9), 892–898 (2015)
51. Harte, J., Kitzes, J.: The use and misuse of species-area relationships in predicting climate-driven extinction. In: Saving a Million Species, pp. 73–86. Springer, Basel (2012)
52. Harte, J., Smith, A.B., Storch, D.: Biodiversity scales from plots to biomes with a universal species–area curve. Ecol. Lett. **12**(8), 789–797 (2009)
53. He, F., Hubbell, S.P.: Species-area relationships always overestimate extinction rates from habitat loss. Nature **473**(7347), 368–371 (2011)
54. Jansen, V.A., Yoshimura, J.: Populations can persist in an environment consisting of sink habitats only. Proc. Natl. Acad. Sci. U. S. A. **95**, 3696–3698 (1998)
55. Johnson, W.E., Onorato, D.P., Roelke, M.E., et al.: Genetic restoration of the Florida panther. Science **329**(5999), 1641–1645 (2010)
56. Juniper, T.: Spix's Macaw: the race to save the world's rarest bird. Simon and Schuster, New York (2004)
57. Kéfi, S., Guttal, V., Brock, W.A., et al.: Early warning signals of ecological transitions: methods for spatial patterns. PLoS One **9**(3), e92, 097 (2014)

58. Kinzig, A.P., Harte, J.: Implications of endemics–area relationships for estimates of species extinctions. Ecology **81**(12), 3305–3311 (2000)
59. Kitzes, J., Harte, J.: Beyond the species–area relationship: improving macroecological extinction estimates. Methods Ecol. Evol. **5**(1), 1–8 (2014)
60. Koh, L.P., Dunn, R.R., Sodhi, N.S., et al.: Species coextinctions and the biodiversity crisis. Science **305**(5690), 1632–1634 (2004)
61. Kolbert, E.: The Sixth Extinction: an Unnatural History. A&C Black, London (2014)
62. Lande, R.: Natural selection and random genetic drift in phenotypic evolution. Evolution **30**, 314–334 (1976)
63. Lande, R.: Risks of population extinction from demographic and environmental stochasticity and random catastrophes. Am. Nat. **142**, 911–927 (1993)
64. Lee, T.E., McCarthy, M.A., Wintle, B.A., et al.: Inferring extinctions from sighting records of variable reliability. J. Appl. Ecol. **51**(1), 251–258 (2014)
65. Leibold, M.A., Holyoak, M., Mouquet, N., et al.: The metacommunity concept: a framework for multi-scale community ecology. Ecol. Lett. **7**(7), 601–613 (2004)
66. Levins, R.: Some demographic and genetic consequences of environmental heterogeneity for biological control. Bull. Entomol. Soc. Am. **15**(3), 237–240 (1969)
67. Levins, R.: Extinction. Lectures on Mathematics in the Life Sciences, vol. 2, pp. 77–107 (1970)
68. Lindsey, H.A., Gallie, J., Taylor, S., et al.: Evolutionary rescue from extinction is contingent on a lower rate of environmental change. Nature **494**(7438), 463–467 (2013)
69. Lomolino, M.V.: Ecology's most general, yet protean pattern: The species-area relationship. J. Biogeography **27**(1), 17–26 (2000)
70. MacArthur, R.H., Wilson, E.O.: Theory of Island Biogeography (MPB-1), vol. 1. Princeton University Press, Princeton (2015)
71. Maslo, B., Fefferman, N.H.: A case study of bats and white-nose syndrome demonstrating how to model population viability with evolutionary effects. Conserv. Biol. **29**(4), 1176–1185 (2015)
72. Matthews, D.P., Gonzalez, A.: The inflationary effects of environmental fluctuations ensure the persistence of sink metapopulations. Ecology **88**(11), 2848–2856 (2007)
73. McCallum, M.L.: Amphibian decline or extinction? Current declines dwarf background extinction rate. J. Herpetol. **41**(3), 483–491 (2007)
74. McCune, J.: Species distribution models predict rare species occurrences despite significant effects of landscape context. J. Appl. Ecol. **53**(6), 1871–1879 (2016)
75. McKinney, M.L., Lockwood, J.L.: Biotic homogenization: a few winners replacing many losers in the next mass extinction. Trends Ecol. Evol. **14**(11), 450–453 (1999)
76. Melbourne, B.A., Hastings, A.: Extinction risk depends strongly on factors contributing to stochasticity. Nature **454**(7200), 100–103 (2008)
77. Milner-Gulland, E.: Catastrophe and hope for the saiga. Oryx **49**(04), 577–577 (2015)
78. Mora, C., Tittensor, D.P., Adl, S., et al.: How many species are there on earth and in the ocean? PLoS Biol. **9**(8), e1001127 (2011)
79. Myers, N.: The Sinking Ark. Pergamon Press, Oxford (1979)
80. Olson, D.H., Aanensen, D.M., Ronnenberg, K.L., et al.: Mapping the global emergence of Batrachochytrium dendrobatidis, the amphibian chytrid fungus. PLoS One **8**(2), e56802 (2013)
81. Pimiento, C., Clements, C.F.: When did Carcharocles megalodon become extinct? A new analysis of the fossil record. PLoS One **9**(10), e111086 (2014)
82. Pounds, J.A., Bustamante, M.R., Coloma, L.A., et al.: Widespread amphibian extinctions from epidemic disease driven by global warming. Nature **439**(7073), 161–167 (2006)
83. Qiao, H., Soberón, J., Peterson, A.T.: No silver bullets in correlative ecological niche modelling: insights from testing among many potential algorithms for niche estimation. Methods Ecol. Evol. **6**(10), 1126–1136 (2015)
84. Raup, D.M., Sepkoski Jr., J.J.: Mass extinctions in the marine fossil record. Science **215**(4539), 1501–1503 (1982)

85. Régnier, C., Achaz, G., Lambert, A., et al.: Mass extinction in poorly known taxa. Proc. Natl. Acad. Sci. **112**(25), 7761–7766 (2015)
86. Roberts, D.L., Solow, A.R.: Flightless birds: when did the dodo become extinct? Nature **426**(6964), 245–245 (2003)
87. Roberts, D.L., Elphick, C.S., Reed, J.M.: Identifying anomalous reports of putatively extinct species and why it matters. Conserv. Biol. **24**(1), 189–196 (2010)
88. Robson, D., Whitlock, J.: Estimation of a truncation point. Biometrika **51**(1/2), 33–39 (1964)
89. Roy, M., Holt, R.D., Barfield, M.: Temporal autocorrelation can enhance the persistence and abundance of metapopulations comprised of coupled sinks. Am. Nat. **166**(2), 246–261 (2005)
90. Rybicki, J., Hanski, I.: Species–area relationships and extinctions caused by habitat loss and fragmentation. Ecol. Lett. **16**(s1), 27–38 (2013)
91. Scheffer, M., Bascompte, J., Brock, W.A., et al.: Early-warning signals for critical transitions. Nature **461**(7260), 53–59 (2009)
92. Scheffers, B.R., Yong, D.L., Harris, J.B.C., et al.: The world's rediscovered species: back from the brink? PLoS One **6**(7), e22531 (2011)
93. Schiffers, K., Bourne, E.C., Lavergne, S., et al.: Limited evolutionary rescue of locally adapted populations facing climate change. Philos. Trans. R. Soc. Lond. B: Biol. Sci. **368**(1610), 20120,083 (2013)
94. Schlichting, C.D., Pigliucci, M.: Phenotypic evolution: a reaction norm perspective. Sinauer Associates Incorporated, Sunderland (1998)
95. Sibley, D.A., Bevier, L.R., Patten, M.A., et al.: Ivory-billed or pileated woodpecker? Science **315**(5818), 1495–1496 (2007)
96. Skerratt, L.F., Berger, L., Speare, R., et al.: Spread of chytridiomycosis has caused the rapid global decline and extinction of frogs. EcoHealth **4**(2), 125–134 (2007)
97. Solow, A.R.: Inferring extinction from sighting data. Ecology **74**(3), 962–964 (1993)
98. Solow, A.R.: Inferring extinction from a sighting record. Math. Biosci. **195**(1), 47–55 (2005)
99. Solow, A.R., Beet, A.R.: On uncertain sightings and inference about extinction. Conserv. Biol. **28**(4), 1119–1123 (2014)
100. Storch, D., Keil, P., Jetz, W.: Universal species-area and endemics-area relationships at continental scales. Nature **488**(7409), 78–81 (2012)
101. Strona, G., Galli, P., Fattorini, S.: Fish parasites resolve the paradox of missing coextinctions. Nat. Commun. **4**, 1718 (2013)
102. Taylor, S., Drielsma, M., Taylor, R., et al.: Applications of rapid evaluation of metapopulation persistence (REMP) in conservation planning for vulnerable fauna species. Environ. Manag. **57**(6), 1281–1291 (2016)
103. Thompson, C., Lee, T., Stone, L., et al.: Inferring extinction risks from sighting records. J. Theor. Biol. **338**, 16–22 (2013)
104. Urban, M.C.: Accelerating extinction risk from climate change. Science **348**(6234), 571–573 (2015)
105. Wiens, J.J.: Climate-related local extinctions are already widespread among plant and animal species. PLOS Biol. **14**(12), e2001104 (2016)
106. Williamson, M., Gaston, K.J., Lonsdale, W.: The species–area relationship does not have an asymptote! J. Biogeogr. **28**(7), 827–830 (2001)
107. Wright, S.: Isolation by distance under diverse systems of mating. Genetics **31**(1), 39 (1946)
108. Wright, S.: The interpretation of population structure by f-statistics with special regard to systems of mating. Evolution pp. 395–420 (1965)

Part IV
Socio-Economics and Infrastructure

Chapter 10
Modeling Food Systems

Hans G. Kaper and Hans Engler

Abstract When enough food is produced but sizable fractions of the population suffer from malnutrition or are overweight, it is time to get a better understanding of the global food system. This chapter introduces food systems and food security as timely research topics for Mathematics of Planet Earth (MPE).

Keywords Food systems · Food security · Planetary boundaries · Social justice · Models · Data

10.1 Introduction

The most important thing to know about the global food system is also one of the least appreciated: there is enough food for everyone on the planet to live a healthy and nutritious life. In fact, the UN Food and Agriculture Organization (FAO) tells us that there are about 2800 kcal per person per day available [16]. But the global food system is deeply inequitable. With about 1 billion people going hungry on the planet [18] and about 2 billion people with insufficient nutrients [41], undernutrition and malnutrition are affecting more than half the world's population. At the same time, there are about 1.5 billion people who are overweight or obese [43]. Clearly, when enough food is produced but sizable fractions of the population suffer from malnutrition or are overweight, we need to get a better understanding of the global food system. The purpose of this chapter is to introduce some of the challenges to food system modeling and describe some of the techniques that have been used to analyze food systems.

H. G. Kaper (✉) · H. Engler
Mathematics and Statistics, Georgetown University, Washington, DC, USA
e-mail: hans.kaper@georgetown.edu; engler@georgetown.edu

H. G. Kaper, F. S. Roberts (eds.), *Mathematics of Planet Earth*, Mathematics of Planet Earth 5, https://doi.org/10.1007/978-3-030-22044-0_10

10.1.1 Food Security and Food Systems

A population enjoys *food security* when all its members, at all times, have access to enough food for an active and healthy life. Limited or uncertain availability of nutritionally adequate and safe foods, or limited or uncertain ability to acquire acceptable foods in socially acceptable ways causes *food insecurity*. A *food system* is the medium that enables a population to achieve food security.

A food system is essentially a supply chain that operates within the broader context of economics, subject to biophysical and socioeconomic constraints. Figure 10.1 gives a schematic representation of a food system. It shows three sectors: *producers*, who engage in farming, livestock raising, horticulture, aquaculture, and fishing; *food chain actors*, who are engaged in "post-farm gate" activities like food processing, packaging, trading, shipping, storing, and retailing; and *consumers*, the ultimate stakeholders of the entire enterprise.

To provide some perspective on the relative economic significance of the three sectors, we note that in developed countries more than 80% of the market value from annual food sales is created by post-farm gate establishments [5, 8]. These food chain actors also consume a considerable share of the natural resources used in the food system, such as fossil fuels for manufacturing and use in home kitchens, forest products for packaging, and fresh water ("blue water") for food-related energy consumption and in home kitchens [7, 48].

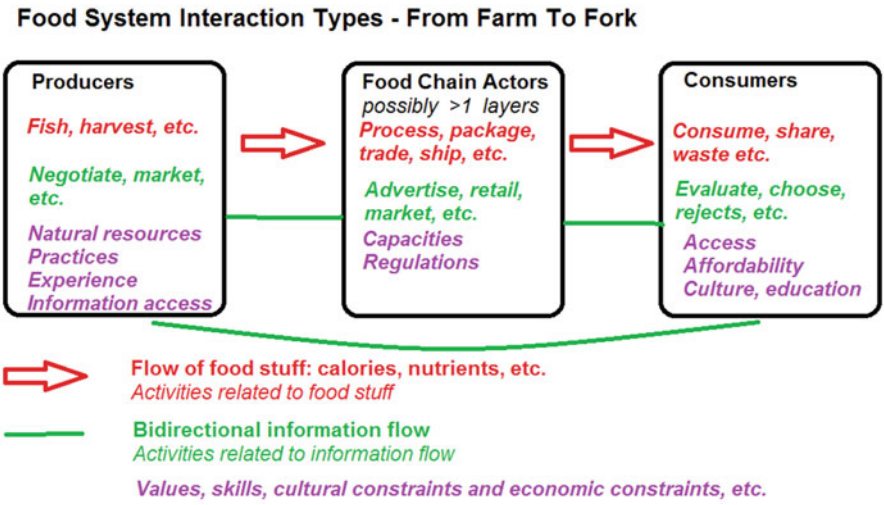


Fig. 10.1 Schematic representation of a food system, showing producers, food chain actors, and consumers, and their respective activities and contributions to the system

Foodstuff generally flows from producers to food system actors to consumers, as indicated by the red arrows in the diagram. Foodstuff can be quantified in many ways—for example, by its energy content (measured in calories), nutrient content (grams of protein), mass, volume, or monetary value. However, *how* foodstuff flows through the system depends in a fundamental way on decisions by a myriad of actors—decisions which are based on information that flows in both directions among the three sectors, as indicated by the green lines in the diagram. In economics, the process leading from information to decision is encapsulated in the law of supply and demand, but this "law" is unlike any law of nature—it may be the result of cultural constraints, social norms, and individual preferences. This poses a fundamental difficulty for food system modeling, since this kind of information is difficult, if not impossible to quantify.

Although food systems function within the broader context of economics, they are more than purely economic systems. To be acceptable, a food system must be both *sustainable* and *equitable*. At every stage, food systems require input of resources of one kind or another (not shown in the diagram of Fig. 10.1). In economic studies, these resources are designated collectively as "capital." They can be "natural" (water, minerals, energy, etc.); "hard" (manufactured goods like tractors, machines, roads, and bridges); "soft" (rules and regulations, organizations); "human" (knowledge and skills); or "social" (cultural institutions, networks, community). All are part of the input to the supply chain. But equally important are the indirect costs and impacts that are associated with food systems. For example, food production affects the ecology and diversity of the environment; food availability and consumption influence personal well-being and public health; the way food is distributed and shared affects the social fabric, sometimes in unanticipated ways; sustainable and safe practices require consensus and public debate. In addition, there are external global phenomena like climate change, natural disasters, and extreme events that threaten food security and introduce elements of risk and uncertainty into the analysis. In short, food systems and food security offer a plethora of challenges for interdisciplinary research. In this chapter, we will make a case that mathematicians working on Mathematics of Planet Earth (MPE) can find exciting opportunities to contribute to this research agenda.

10.1.2 Outline of the Chapter

In Sect. 10.2, we present a framework for the study of food systems that takes into account both the limits of our physical environment and a human value system. In Sect. 10.3, we make the case that a food system is a complex adaptive system requiring appropriate methodologies to capture its key features. In subsequent sections we discuss various modeling techniques: network models in Sect. 10.4, agent-based models in Sect. 10.5, equation-based models in Sect. 10.6, system dynamics models in Sect. 10.7, statistical models in Sect. 10.8, and economic models in Sect. 10.9. Section 10.10 is devoted to data, data assimilation, data

visualization, data sources, and issues associated with missing data. In the final Sect. 10.11, we briefly summarize the discussion.

10.2 A Framework for Discussing Food Systems

Economics has dominated the discussion of food systems since the early nineteenth century. The English economists Thomas Malthus (1766–1834), best known for his theory that population growth will always tend to outrun the food supply, and David Ricardo (1772–1823), one the most important figures in the development of economic theory, published their seminal papers on land use and agriculture in 1798 [33] and 1815 [53], respectively, Today, agricultural economics is a subdiscipline of economics focused entirely on the agricultural sector.

Similarly, aquaculture economics addresses the economics of fisheries, especially the management of fisheries. Since marine diseases are common in the world's oceans, the study of infectious diseases is an integral element of aquaculture economics.

As explained in Sect. 10.1.1, food systems are supply chains that require inputs of one form or another at every stage. Consequently, resource economics—the allocation of resources, with the goal of optimizing output—plays a significant role in the study of food systems, especially in the presence of the biophysical and socioeconomic constraints required for a sustainable and equitable operation [10, 52].

Economic geography—the study of the relation between location and the economics of production—provides yet another perspective on food systems. The beginning of economic geography is generally identified with the publication of a pamphlet by Johann von Thünen in 1826 [63], where the author developed a conceptual mathematical model of a food system to determine the optimal location of different agricultural production sectors relative to an urban center. Economic geography was advanced substantially through the work of Paul Krugman (for example, see [32]), for which he was awarded the Nobel Prize in Economics in 2008.

Concerns about the impact of food systems on the environment and the loss of biological diversity have led to the specialties of environmental economics and ecological economics [11]. A useful reference in this context is the *System of Environmental–Economic Accounting (SEEA)—Central Framework*, an integrated environmental and economic accounting framework for organizing data [60]. The SEEA—Central Framework was adopted in 2014 by a group of international organizations, including the United Nations, the European Union, and the International Monetary Fund.

Not every aspect of food systems falls under the purview of economics. The relation between food and nutrition is a recurrent theme, for example, in the discussions of obesity and public health. Globalization and urbanization have increased the risks of infectious diseases and raised concerns about food safety and public health.

In this section, we present a framework for the discussion of food systems that goes beyond economics, accounting for critical issues of sustainability and social justice. The framework provides a convenient organizing principle for the development of mathematical models.

10.2.1 Planetary Boundaries

In 2009, a group of Earth system and environmental scientists led by Johan Rockström (Stockholm Resilience Centre) and Will Steffen (Australian National University) argued that humanity must stay within defined boundaries for a range of essential Earth-system processes to avoid catastrophic environmental change [49]. They proposed a framework to measure stress to the Earth system in terms of nine stress factors: climate change, ocean acidification, biochemical flows, freshwater use, land-use change, biosphere integrity, ozone layer depletion, air pollution, and chemical pollution. They also presented control variables (indicators) to introduce a metric for each stress factor, suggesting upper or lower bounds as appropriate, for seven of the nine indicators. They referred to these bounds as *planetary boundaries*. Crossing even one of the planetary boundaries would risk triggering abrupt or irreversible environmental changes, and the authors suggested that, if one boundary were transgressed, there would be a more serious risk of breaching the other boundaries. In other words, the planetary boundaries jointly constitute an *ecological ceiling*, beyond which humanity's pressure threatens Earth's life-supporting capacity.

Table 10.1 lists the stress factors and their indicators, the ecologically acceptable (upper or lower) bounds on the indicators ("planetary boundaries"), the estimated current values of the indicators and their trends (improving or worsening). Figure 10.2 gives a graphical representation of the current status. The indicators are measured from the center; a necessary condition for sustainable development is that all indicators are within the environmental ceiling. The red sectors indicate that the boundaries for genetic diversity, nitrogen and phosphorus use have already been surpassed.

It can be argued that, with our limited understanding of the fundamental processes controlling the stress factors, it is impossible to present reasonable numbers, or that the bounds are much more malleable than the boundaries suggest, or that, with better or worse management, boundaries could be moved. The concept of planetary boundaries, however, is now generally accepted and has since been adopted, for example, by the United Nations for ecosystem management and environmental governance. Expert commentaries can be found in [1, 2, 6, 38, 39, 50, 51], a scholarly discussion of boundaries and indicators in [21], and an update of the original framework in [55]. We also refer to the recent book by Raworth [46] for a discussion of planetary boundaries in the context of economics.

Table 10.1 Ecological stress factors of Earth's system, together with their indicators, planetary boundaries, estimated current values, and trends

Stress factor	Indicator	Planetary boundary	Current value, trend
Climate change	Atmospheric CO_2 concentration	$\leq$350 ppm	400 ppm, worsening
Ocean acidification	Average saturation of aragonite ($CaCO_3$) at ocean surface	$\geq$80% of pre-industrial level	84%, intensifying
Biochemical flows	(1) Phosphorus applied to land as fertilizer	$\leq$6.2 Mt/yr	14 Mt/yr, worsening
	(2) Nitrogen applied to land as fertilizer	$\leq$6.2 Mt/yr	150 Mt/yr, worsening
Freshwater use	Blue water consumption	$\leq$4000 km^3/yr	2600 km^3/yr, intensifying
Land-use change	Forested land area	$\geq$75% of forest-covered land prior to human alteration	62%, worsening
Biosphere integrity	Species extinction rate	$\leq$10 pms/yr	100–1000 pms/yr, worsening
Ozone layer depletion	Stratospheric O_3 concentration	$\geq$275 DU	283 DU, improving
Air pollution	TBD		
Chemical pollution	TBD		

Abbreviations used: *ppm* parts per million, *Mt* million tons, *pms* per million species, *DU* Dobson units. Adapted from [46, Appendix, Table 2], data from [54]

10.2.2 Social Justice

The concept of planetary boundaries is a recognition of the fact that there are biophysical and ecological constraints to the Earth system. The planetary boundaries define a sustainable operating space for humanity; beyond the environmental ceiling lie unacceptable degradation and potential tipping points. What is missing in this framework is a recognition that global environmental stresses also pose threats to human well-being. Eradicating poverty and achieving social justice are inextricably linked to ensuring ecological stability and renewal. To make the connection, we must complement the environmental ceiling with a *social foundation*—a set of generally accepted social priorities which, if not met, imply unacceptable human deprivation.

The guidelines for the UN Conference on Sustainable Development (known as Rio+20, which took place in June 2012) suggest a set of twelve social priorities, which form the basis of a social foundation. Table 10.2 lists the social priorities and their indicators, and estimated values of the indicators for the year(s) listed. Figure 10.3 gives a graphical representation of the current status. The indicators are measured from the center; a socially equitable system is realized when the social

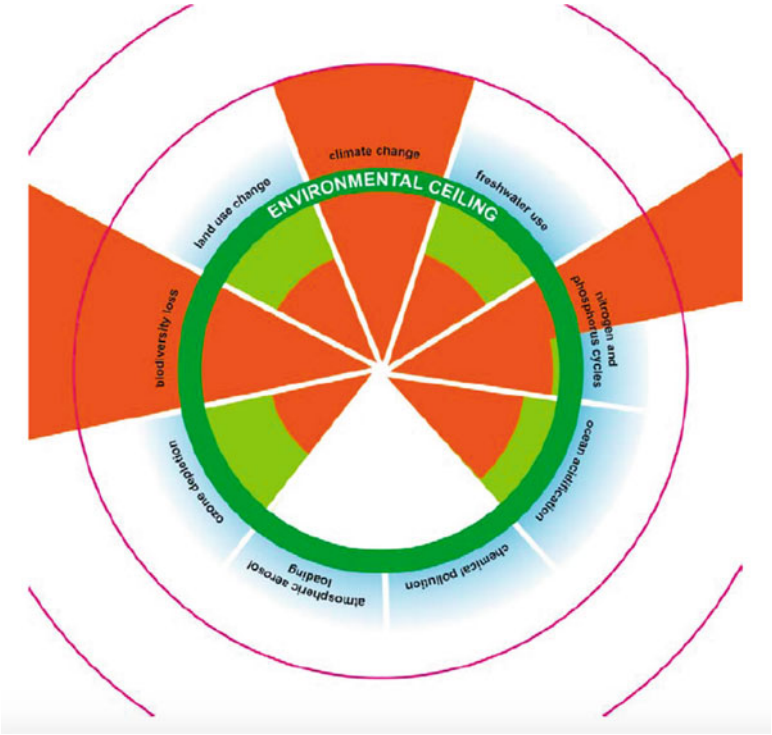


Fig. 10.2 Estimated status of the indicators for seven of the ecological stress factors [55]

foundation is achieved in all sectors. The red sectors indicate not only that we are falling short on social justice at the global level, but also that there are significant discrepancies among the various indicators.

The *Sustainable Development Goals* (SDGs), which were adopted by the United Nations in 2015 to improve human lives and protect the environment [59], incorporate the concepts of planetary boundaries and social priorities.

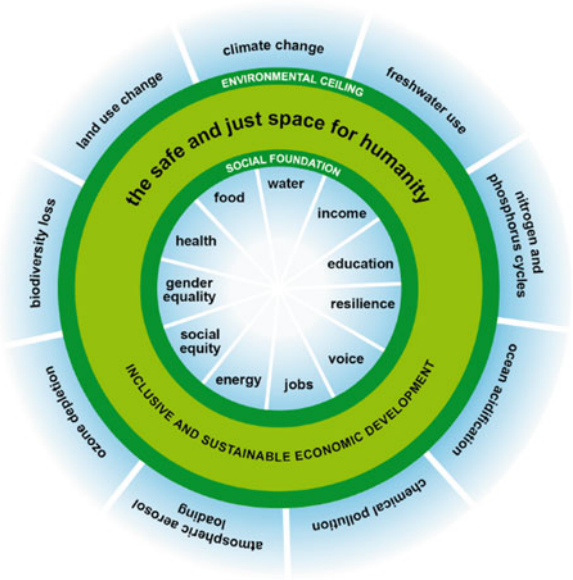
10.2.3 Doughnut Economics

By placing the indicators for social justice inside the planetary boundaries, Raworth [44–46] achieved a visual representation of a *safe and just operating space* for humanity. An economic system that stays inside the ring bounded externally by the environmental ceiling and internally by the social foundation allows for sustainable development while maintaining social justice for all. Since the ring is reminiscent of a (two-dimensional) doughnut, this framework is sometimes described as *doughnut economics*.

Table 10.2 Social priorities and their indicators, estimated current values, and year

Social priority	Indicator (fraction of global population, unless otherwise stated)	%	Year
Food	Population undernourished	11	2014–2016
Health	(1) Population living in countries with under-5 mortality rate	46	2015
	(2) Population living in countries with life expectancy at birth < 70 years	39	2013
Education	(1) Adult population aged 15+ who are illiterate	15	2013
	(2) Children aged 12–15 out of school	17	2013
Income and work	(1) Population living on less than $3.10 per day	29	2012
	(2) Proportion of people aged 15–24 seeking but unable to find work	13	2014
Water and	(1) Population without access to improved drinking water	9	2015
sanitation	(2) Population without access to improved sanitation	32	2015
Energy	(1) Population lacking access to electricity	17	2013
	(2) Population lacking access to clean cooking facilities	38	2013
Networks	(1) Population stating that they are without access to someone to count on for help in times of trouble	24	2015
	(2) Population lacking access to the Internet	57	2015
Housing	Global urban population living in slum housing in developing countries	24	2012
Gender equality	(1) Representation gap between women and men in national parliaments	56	2014
	(2) Worldwide earnings gap between women and men	23	2009
Social equity	Population living in countries with a Palma ratio ≥ 2	39	1995–2012
Political voice	Population living in countries scoring $\leq 0.5/1.0$ in the VAI	52	2013
Peace and justice	(1) Population living in countries scoring $\leq 50/100$ in the CPI	85	2014
	(2) Population living in countries with a homicide rate ≥ 10 per 10,000	13	2008–2013

Column 3: % of global population unless otherwise stated; Palma ratio: ratio of the income share of the top 10% of people to that of the bottom 40%.
Abbreviations: *VAI* Voice and Accountability Index, perceptions of the extent to which a country's citizens are able to participate in selecting their government, as well as freedom of expression, freedom of association, and a free media; *CPI* Corruption Perception Index, perceived levels of misuse of public power for private benefit, as determined by expert assessments and opinion surveys. Adapted from [46, Appendix, Table 1], data from FAO, World Bank, WHO, UNDP, UNESCO, UNICEF, OECD, IEA, Gallup, ITU, UN, Cobham and Sumner, ILO, UNODC, and Transparency International



10.2.4 A Simplified Framework

A simplified framework for food system modeling was presented in a recent report prepared by the U.S. National Academy of Sciences for the Institute of Medicine (IOM) and the National Research Council (NRC) [62]. It is based on the conceptual model shown in Fig. 10.4. (Although the diagram refers to the agricultural sector, it is equally applicable to other sectors of the food system.)

The food system is comprised of four components: a component labeled "input" is added to the three standard components (producers, food chain actors, and consumers) indicated in Fig. 10.1. The number of planetary boundaries is reduced to four (air, biota, land, water), collectively labeled *Natural Resources*; similarly, the number of social priorities is reduced to four (health, markets, policy, well-being), collectively labeled *Human Systems*. The figure highlights the role of feedback mechanisms indicated by the arrows in the diagram, which can influence the evolution of the food system either positively (reinforcing) or negatively (inhibiting). For example, natural resources like air, soil, water, and biota (pollinators, natural enemies of food pests) are essential for agricultural production, as well as the manufacture of many foods like bread, cheese, and wine. Yet, depletion and effluents from the food system influence the future status of natural resources. Likewise, the food system depends on a host of human systems that govern our health, markets, policy, and general well-being. These human systems provide the labor, entrepreneurship, capital, and technology needed to produce and distribute food. Once again, the food system generates feedbacks that influence human systems at a future period.

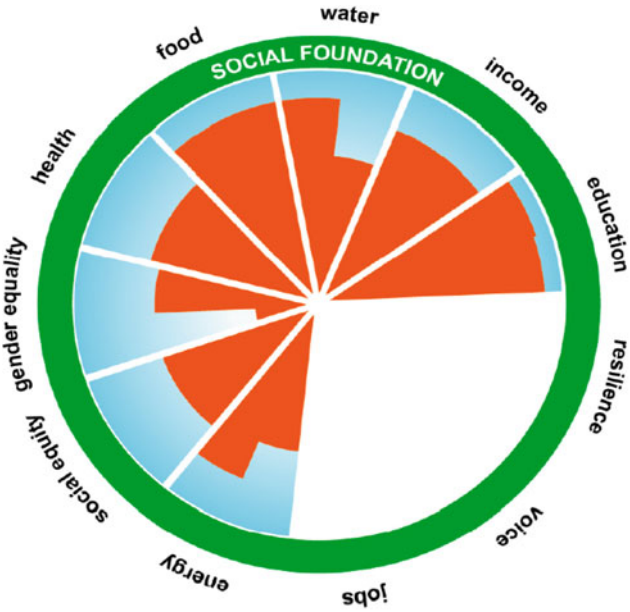


Fig. 10.3 Estimated status of the indicators for eight of the social boundaries [44]

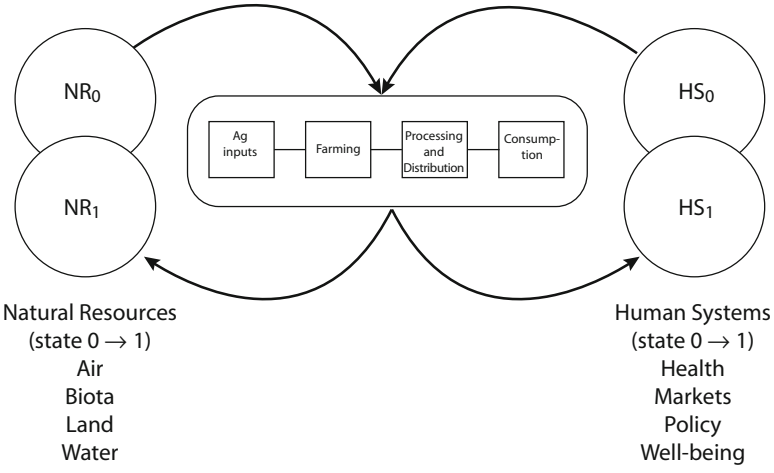


Fig. 10.4 Food system as a dynamic process transforming the state of natural resources and human systems from one period to the next [62]

10.3 A Modeler's Perspective

Food systems have many of the characteristics of a *complex system*. While there is no generally accepted definition of a complex system, complex systems can be

characterized by what they do [25]. Four properties stand out, each of which adds complexity to a system:

1. A complex system has *internal structure*. The structure may consist of many interacting components, a network that describes which components of the system interact, feedback mechanisms across multiple scales of space and/or time, and symmetries. The components of many complex systems are heterogeneous and form a hierarchy of subsystems.
2. A complex system shows *emergent behaviors*. Such behaviors arise from the interaction of subsystems and are not evident from an analysis of each and every subsystem. Chaotic dynamics, tipping points, and phase transitions are examples of emergent behaviors.
3. A complex system can adapt to its environment and evolve without external guidance. *Adaptation* and *evolution* are characteristic of critical infrastructure systems and fundamental to biological and social systems.
4. *Uncertainty* is pervasive in complex systems. Quantifying uncertainties and determining how uncertainties propagate throughout the system is a key aspect of reliable prediction and control.

Like all complex systems, food systems pose challenges for mathematical modeling. Clearly, no single model can capture an entire food system; specific questions require specific models tailored to the scale of the phenomenon of interest. In the following sections we present various modeling techniques that have been used in the study of food systems, with case studies where relevant.

10.4 Network Models

Network models are fundamental to understanding interactions within complex systems. They are also critical for the analysis of a system's resilience to shocks and other disturbances.

A network consists of a set of *nodes* (or *vertices*) and a set of *edges* connecting nodes. A network is more than a graph. Nodes can be complex systems themselves, so a network exposes the hierarchical structure of the system under consideration. Nodes can have locations, demographics, and content. Edges are not just connections; they can have weights, capacities, lengths, and other attributes characterizing the interactions among the nodes. Networks can change over time, they often have multiple types of links, and may include higher order interactions (motifs) in addition to pairwise ones. Many emergent phenomena in complex systems cannot be understood without understanding the structure of the underlying network.

Often, networks have certain attributes that can be quantified. Examples are its size (measured by the number of nodes or, less frequently, the number of edges), density (the ratio of the number of edges to the number of possible edges), average degree (the average number of edges connected to a node), characteristic path length (the average number of steps it takes to get from one member of the network

to another), diameter (the shortest distance between the two most distant nodes), average clustering coefficient (the average of the ratio of existing links connecting a node's neighbors to each other to the maximum possible number of such links), node centrality (an index used to rank the most important nodes), and node accessibility (an index measuring how accessible the rest of the network is from a given start node). All these measures can be meaningfully computed from the structure of the network alone, and several of them can be used to define types of networks (small-world, preferential attachment, etc.). We refer the reader to the article by Newman [42] for a survey of the structure and function of complex networks.

Networks play an important role, for example in transportation modeling. Two frameworks dominate both the applied and theoretical research on this topic. The *gravity model* framework [29] is used to explain interactions between different locations from basic geographic and demographic properties. Currently the most complete framework for economic geography is the *New Trade Theory*, which is based on the work that won Paul Krugman the Nobel Prize in Economics in 2008 [32]; see [35] for an application to a trade/transportation problem.

Case Study An example of network modeling can be found in [14]. The authors were interested in the future development of the global wheat trade system and its resilience to future shocks of differing length and severity. Note that the concept of resilience encompasses not only the robustness of a system to damage resulting from shocks but also the speed at which it recovers from shocks [47].

As a first step, the authors used available data for the global wheat trade for the period 1986–2011 to develop a "backbone" network (referred to as the *empirical network model*) by including only those edges corresponding to the largest trades by volume. The empirical network, shown in Fig. 10.5a, is a simplification which retains much of the structure of the historic wheat trade system.

The authors then developed a preferential attachment (PA) model to emulate the empirical network. In this PA model, the probability of a trade (import or export) between countries i and j occurring at time t depends on the fitness $x_i(m)$ of country i and the fitness $x_j(m)$ of country j at time t,

$$P_t[x_i(m), x_j(m)] = \frac{\alpha x_i(m) x_j(m) + \varepsilon}{1 + \beta x_i(m) x_j(m)}. \tag{10.4.1}$$

Here, the fitness $x_i(m)$ of country i is the ratio $x_i(m) = I_i(m)/m$, where m is the number of directed edges in the network and $l_i(m)$ is number of edges connected to country i; similarly for the fitness $x_j(m)$ of country j. The expression (10.4.1) shows that a PA model favors fitness; to paraphrase, a PA model corresponds to a system where the rich get richer. The small probability ε is introduced to enable the formation of an edge between countries i and j when one or both have zero fitness (that is, one or both are not yet engaged in any trade partnerships). The free parameters α and β are chosen to match several characteristics of the empirical network; the details of the matching procedure and a comparison of the two network models can be found in [14]. The PA network model is shown in Fig. 10.5b.

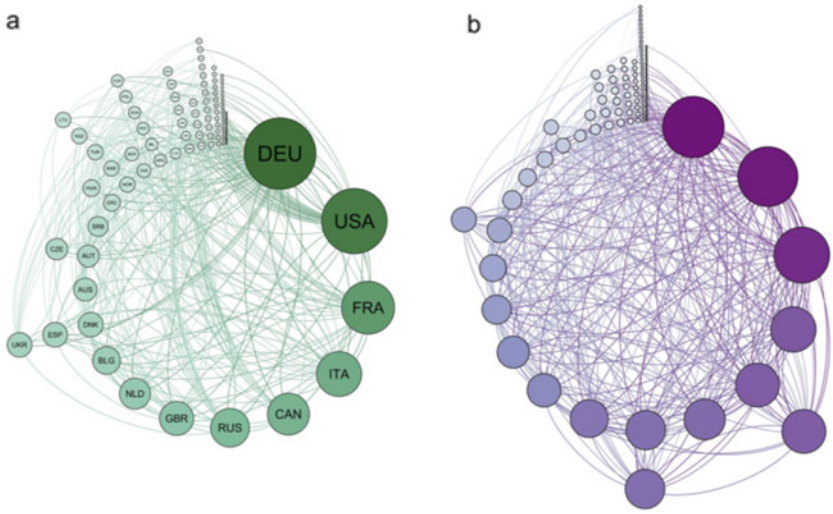


Fig. 10.5 (**a**) Empirical network model and (**b**) preferential attachment (PA) network model of the global wheat trade system at the end of the year 2013. Nodes are arranged clockwise, and in size and color, by total degree. Node labels in the empirical network model correspond to the ISO 3166-1 alpha-3 codes for the countries they represent [14]

An analysis of the time evolution of historical metrics in the empirical network and predictions of future trends in the PA network model enabled the authors to present various conclusions concerning the evolution of the global wheat trade network, the effect of factors such as extreme weather events or agro-terrorism on its dynamics, and its resilience to short-term and repeated shocks.

10.5 Agent-Based Models

Agent-based models (ABMs) are a class of computational models that are used to simulate the actions and interactions of autonomous *agents*. An agent can be an individual, or a group of individuals, or any other independent entity like a community, an institution, or a country. Agents act within a specific environment and interact with and influence other agents according to a given set of rules. The effect of their interactions often emerges in the behavior of the system as a whole, as *collective phenomena* such as patterns, structures, and self-organization, which were not explicitly programmed into the model but emerge as a result of the agents' interactions.

Because of their versatility and relatively straightforward implementation, ABMs are used in several scientific domains including biology, ecology (where they are referred to as *individual-based models*, IBMs), and social sciences to gain explanatory insight into the collective behavior of agents obeying simple rules.

A typical agent-based model has three elements: (1) a set of agents, their attributes and behaviors; (2) a set of agent relationships and methods of interaction: an underlying topology of connectedness, which defines how and with whom agents interact; and (3) the agents' environment: agents interact with their environment in addition to other agents. ABMs are typically implemented as computer simulations, either using custom software or via ABM toolkits [56]. This software can be then used to test how changes in individual behaviors will affect the system's emerging overall behavior.

Case Study An example of ABM modeling is described in [37]. The ABM concerns key land-use/land-cover dynamics and livelihood decisions on Isabela Island in the Galápagos Archipelago of Ecuador.

The common guava (*Psidium guajava*), which was introduced to the Galápagos for cultivation in 1858, now covers more than 40,000 ha of land on Isabela Island, mostly within the agricultural zone and the adjacent protected area of Galápagos National Park (GNP). Its spread in the agricultural zone is an obstacle to cultivation and can substantially reduce farm productivity. As a consequence, some households have decided to alter their land-use patterns by allowing infested fields to lie fallow, effectively abandoning portions of the farm that are dominated by guava. These abandoned fields and farms act as source populations that promote the spread of guava into neighboring farms and the GNP. The spread of guava depresses household wealth and assets related to agriculture and necessitates the hiring of contract labor from the Ecuadorian mainland to eradicate problem populations in farm fields, as well as the imposition of control measures in the GNP.

Figure 10.6 gives a schematic representation of the agent-based model designed for this study. The assumption is that distal factors (economic markets, policy, environmental variation) influence more proximate, local (political-economic, socio-cultural, biophysical) landscapes, which in turn affect agent characteristics, livelihood decisions, and ultimately land-use patterns. Feedbacks among many of the factors are present in this framework and specified in the model. Livelihood decision making involves the selection of one of three options: agriculture, fisheries, or tourism, which are the largest economic sectors on Isabela Island. Total guava coverage is the primary output of the model and is used to measure land-cover change. The model is implemented in the NetLogo software platform [56].

The authors' primary goal was to evaluate broad population and environment trends in the study area. By separately testing several levels of direct income subsidy (increasing annual farmer incomes) and control cost subsidy (reducing the guava control cost per hectare), the authors were able to estimate the approximate level of income subsidy required to bring farmers' incomes in line with the other two economic sectors (fisheries and tourism), and the approximate level of control cost subsidy to eliminate the cost of clearing guava; for details, see Ref. [37].

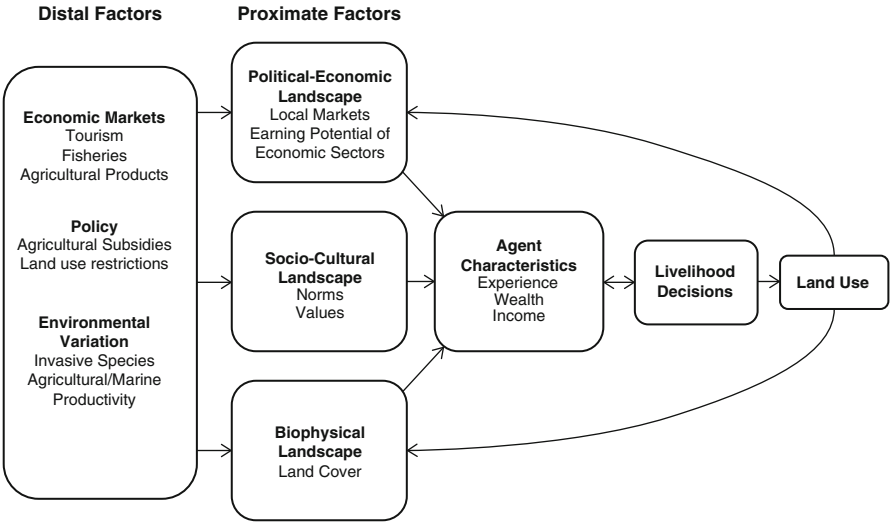


Fig. 10.6 Conceptual framework for modeling livelihood decisions and land use/land cover processes on Isabela Island in the Galápagos Archipelago of Ecuador [37]

10.6 Equation-Based Models

If the state of a system on the scale of interest is completely characterized at any time by one or more *state variables* and the mutual dependencies among the state variables are known—for example, following certain laws of nature—then it is possible to develop an *equation-based model* (EBM) of the system. Generally speaking, while individual agents and their mutual interactions (that is, attributes on the microscale) are the primary concern in ABMs, the focus of EBMs is usually on aggregate (macroscopic) quantities. EBMs are common in the physical sciences, where state variables like mass, momentum, and energy are governed by universal laws like conservation laws, Newton's laws of motion, and the laws of thermodynamics. Since there are no such laws in food systems or, more generally, in social systems, EBMs are less common in this context.

Nevertheless, it is sometimes possible to study a particular observable phenomenon using a conceptual model that embodies general principles derived from data or even common sense. Such models can be *static*, describing an equilibrium state of the system, or *dynamic*, relating the rate of change of the state variables to the current state of the system or, if there are memory effects, to the state of the system over an immediate past time interval. This class of EBMs generally comprises differential or difference equations, depending on whether time is taken as a continuous or discrete variable. Since there are usually uncertainties associated with the data, for example because the measurements were subject to error or the data were incomplete, statistical techniques may have to be applied to "clean" the

data and quantify the uncertainties. Processes that occur on scales that are not captured by the model introduce additional uncertainties, giving rise to stochastic equations.

Case Study An example of EBM modeling is described in [4]. The case originated from a project on the Pajaro Valley of California. This region is known for water-intensive berry farming and has agencies actively seeking sustainable agricultural practices. The crops under consideration—lettuce, strawberries, raspberries, and blackberries—have different growing seasons, with raspberries in particular requiring model parameters that vary with time.

Each farmer has a certain amount of arable land and a choice of crops to plant on the land. The goal is to develop a flexible modeling and optimization framework to aid the farmers in selecting crop portfolios that offer the best financial outcome, under sustainable water usage limitations, over specified time frames. The crop portfolio available to each farmer should not deviate significantly from the existing crop state, and the optimization problem must take into account constraints, possibly dynamic, to enforce planting and harvesting schedules.

A flowchart depicting the strategy developed for the optimization problem is given in Fig. 10.7.

In the preprocessing step, the growing and harvesting schedules for the crops under consideration are outlined in a calendar that tracks when land becomes available and what is being harvested. The calendar fixes the constraints which enforce the details of the growing season and leads to the determination of the state variables. At each decision point, an inequality constraint must be enforced to ensure that the total percentage of farm allocation does not exceed 100%. Land may be left fallow, which is beneficial in scenarios involving water restrictions or soil treatments used to increase nutrients (and thereby yields) in subsequent planting periods.

The objective function incorporates the information a farmer uses to make crop planting decisions. Examples of such information include profitability, limited changes in crop portfolios from year to year, meeting consumer demand on an

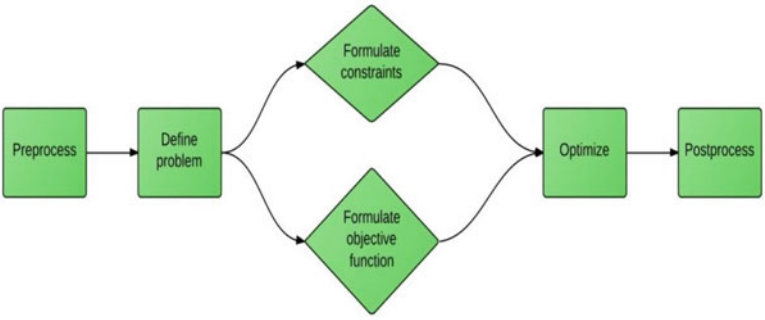


Fig. 10.7 Conceptual framework for modeling crop selection decisions [4]

annual basis, and minimizing the use of costly resources. The objective function can be defined as a single target, or it can include multiple, often competing, targets. For example, minimizing the use of costly resources does not necessarily increase profitability. Many crops have minimum resource requirements for a successful harvest, and limiting their use often competes with profitability objectives.

The state variables chosen by the authors for each crop i ($i = 1, \ldots, N$), together with their units, are

Y_i	Yield from one harvest	[boxes/acre]
W_i	Water usage	[acre-ft/acre]
P_i	Sales price	[$/box]
D_i	Demand	[% crop/year]
C_i	Operational planting cost	[$/acre]

Profit models can be as complex or as simple as needed once the farming model is in place. For instance, a simple representation of profit could be

$$\text{Profit} = \sum_{i=1}^{N} A_i (Y_i P_i - P_w W_i - C_i), \tag{10.6.1}$$

where A_i is the number of acres planted with crop i, and P_w is the current price of water [$/acre-ft]. The maximization of profit must be accomplished under the constraints of minimal water usage over the entire growing season,

$$\text{Water} = \sum_{i=1}^{N} A_i W_i, \tag{10.6.2}$$

and minimal deviation from a given demand vector $d = (d^1, \ldots, d^N)^{\mathrm{T}}$.

The optimization problem can be solved with existing software, for example in the Matlab suite of programs; details can be found in [4].

10.7 System Dynamics Models

A special class of EBMs is known as *system dynamics models* (SDMs). These are basically systems of equations written in a graphical language, showing the system's state variables and their causal relationships. The following description is based on the system dynamics entry in Wikipedia [65], which also gives several examples.

In a SDM, the system of interest is represented as a *causal-loop diagram*. We have already encountered such a diagram in Fig. 10.4. Another simple example is given in Fig. 10.8. There are two feedback loops in this diagram. The positive

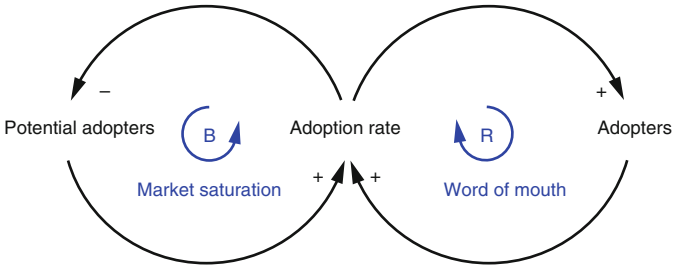
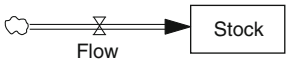


Fig. 10.8 Causal-loop diagram showing the adoption process of a new product

reinforcement loop (labeled R) on the right indicates that the more people have already adopted the new product, the stronger the word-of-mouth impact. There will be more references to the product, more demonstrations, and more reviews. This positive feedback should generate sales that continue to grow. The second feedback loop on the left is negative reinforcement (or "balancing" and hence labeled B). Clearly, growth cannot continue forever, because as more people adopt, fewer potential adopters remain. Both feedback loops act simultaneously, but at different times they may have different strengths: one might expect growing sales in the initial years, and then declining sales in the later years.

In general, a causal-loop diagram does not specify the structure of a system sufficiently to permit determination of its behavior from the visual representation alone. To perform a more detailed quantitative analysis, a causal-loop diagram is



transformed to a *stock-and-flow diagram*, a stock being any entity that accumulates or depletes over time (a state variable) and a flow being the rate of accumulation of the stock.

Figure 10.9 shows the stock-and-flow diagram for the adoption process described in Fig. 10.8. In this example, there are two stocks, "Potential adopters" (P) and "Adopters" (A), and one flow, "New adopters" (N). The diagram gives a visual representation of the ways in which a Potential adopter can become part of the flow: as an Innovator who adopts the product without prodding, or as an Imitator who adopts the product after learning about it from an Adopter. For every New adopter, the number of Potential adopters decreases by one, while the number of Adopters increases by one. Hence, the sum $P + A$ is constant at all times and equal to P_0, the number of Potential adopters at the start when no one has yet had an opportunity to adopt the new product. The model results in an S-shaped curve for the number of Adopters, which rises slowly in the beginning, then increases rapidly and gradually slows down as the market for the new product saturates.

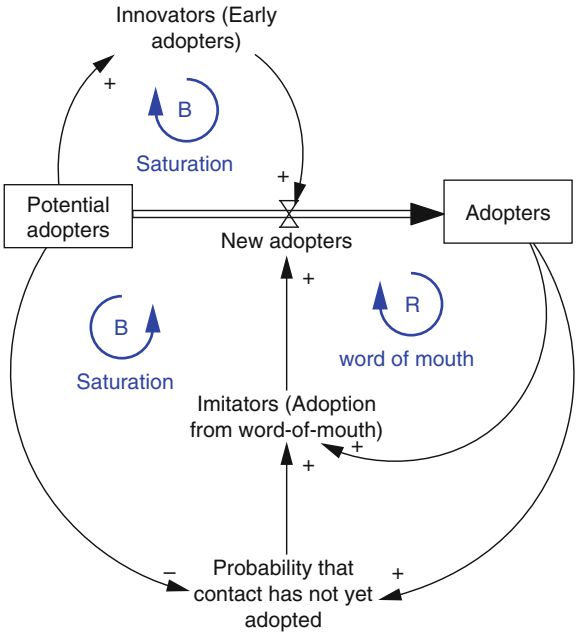


Fig. 10.9 Stock-and-flow diagram for the adoption process described in Fig. 10.8

The language of SDMs was developed in the engineering community in the 1950s to better understand the behavior of complex systems. The dynamics of the system are simulated by updating all variables in small time increments, with positive and negative feedbacks and time delays structuring the interactions and control.

10.8 Statistical Models

Statistical models (SMs) are similar to EBMs in the sense that they describe aggregate properties of populations. But what distinguishes a statistical model from other mathematical models is that a statistical model is non-deterministic. Some of the variables do not have specific values; instead, they have probability distributions—that is, some of the variables are *stochastic*. Typically, a process model is assumed to be given, and the emphasis of stochastic modeling is on the incorporation of the data into the model, the prediction of bulk behavior, and the assessment of uncertainties.

Since most models of food systems are data-driven, statistics play an important role. The data may have errors because observations are error-prone, and they may be incomplete because of limitations of the observational procedure. Moreover, any model fails to capture sub-scale phenomena, which are then represented by

parameters whose values must be guessed. All this introduces uncertainties into the model, which need to be assessed using statistical methods.

For example, to investigate how crop yield varies with precipitation, we assume a linear relationship of the form "yield $= b_0 + b_1 \cdot$ precip," where b_0 and b_1 are (unknown) constants. After n observations we have a set of data points $(\text{precip}_i, \text{yield}_i)$, $i = 1, 2, \ldots, n$, in the sample space S of all possible pairs (precip, yield). The data points do not necessarily lie on a straight line, so we modify the linear relationship to a linear regression model, "$\text{yield}_i = b_0 + b_1 \cdot \text{precip}_i + \varepsilon_i$." Here, ε_i is a stochastic error term; without it, the model would be deterministic. Assuming that the errors are independent and identically distributed (iid) with a Gaussian distribution $N(0, \sigma^2)$ with mean 0 and variance σ, we now have a statistical model with three degrees of freedom, namely b_0, b_1, and σ, which we denote collectively by a single symbol, say $\theta = (b_0, b_1, \sigma)$. Each possible value of θ determines a distribution P_θ on S. If Θ is the set of all possible values of θ, then $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ is the set of all probability distributions on S. *Statistical inference* is the process of deducing properties of the probability distribution $\mathscr{P}$ from the data set.

Among the most common techniques of statistical inference are the classical (or frequentist) paradigm and the Bayesian paradigm. In the frequentist paradigm, the properties of a statistical proposition are quantified by repeated sampling of the data set. The Bayesian paradigm requires the specification of prior distributions for all unknown parameters, which are subsequently improved upon using available data. Parameters of prior distributions may themselves have prior distributions, leading to Bayesian hierarchical modeling, or they may be interrelated, leading to Bayesian networks.

10.9 Economic Models

Economists have studied food systems mostly for their role within in the overall economy, maximizing GDP being one of the main drivers. In this section we highlight two common modeling techniques, input/output models and computable general equilibrium models.

10.9.1 Input–Output Models

Input/Output (I/O, or simply IO) models are matrix equations that show how output from one industrial sector may become an input to another industrial sector. (IO can also stand for Industrial Organization.) The column entries of a typical inter-industry IO matrix represent inputs to an industrial sector, while row entries represent outputs from a given sector. The size and sparsity of the matrix depend on the granularity of the IO model. Researchers typically use matrix algebra tools to analyze IO models.

The history of IO models in economic analysis goes back to the work of Wassily Leontief (1906–1999), who was the first to use a matrix representation of a national (or regional) economy. This work earned him the Nobel Prize in Economics in 1973. In contemporary environmental accounting, one uses an extended IO framework, called *Environmentally extended input-output analysis* (EEIOA), often in combination with the SEEA—Central Framework discussed at the beginning of Sect. 10.2. An example can be found in [64].

Input/output models offer universal ways to describe economic sectors, also on a regional scale, and thus have a profound influence on the ways in which economic sectors (including the food sector) are described and data are collected. They are capable of predicting price structures and responses to small changes. A comprehensive modern reference is [36].

10.9.2 Computable General Equilibrium Models

Computable General Equilibrium (CGE) models are nonlinear versions of I/O models that take the behavior of economic agents into account and can incorporate the effects of external factors such as changes of market rules or environmental change.

A CGE model consists of a set of equations describing the model variables and their interdependencies, together with a (usually very detailed) database. The equations often assume cost-minimizing behavior by producers, average-cost pricing, and household demands based on optimizing behavior. Since CGE models always contain more variables than equations, some variables must be set outside the model; these variables are termed *exogenous*. The remaining variables, which are determined by the model, are called *endogenous*. In the language of mathematics, the endogenous variables are the *state variables*; once we know their values, we know the state of the system. The exogenous variables affect the state of the system but are not affected by it; they can be *external forces* or *parameters*.

For example, rainfall is exogenous to a system describing the process of farming and crop output. There are causal factors that determine the level of rainfall—so rainfall is endogenous to a weather model—but these factors are not themselves part of the model used to explain the level of crop output. Variables defining technology, consumer preferences, and government instruments (such as tax rates) are usually exogenous.

CGE models are a useful tool to assess the effect of policy changes. One runs the model twice, once to calculate the equilibrium without the change and once to calculate the equilibrium with the change. Then one quantifies the impact of the change on the state of the system by comparing the equilibria from the two experiments. Time does not enter, and the transient dynamics in the process of adjustment to a new equilibrium are ignored. This procedure is known as the *comparative-static* CGE model.

In mathematical terms, let $x \in \mathbb{R}^n$ denote the vector of state variables and $\lambda \in \mathbb{R}^m$ the vector of parameters, and let the function f represent the system of equations describing the state variables and their interdependencies in the absence of external forces, $f : (\lambda, x) \mapsto f(\lambda, x)$ for all $(\lambda, x) \in \mathrm{dom}(f) \subset \mathbb{R}^{m+n}$. The function f is defined in such a way that $f(\lambda, x) = 0$ if and only if the state of the system represented by x is realizable as an equilibrium state for the vector of parameters λ. (The value 0 is arbitrary and chosen for the sake of convenience.) Then the comparative-static CGE procedure can be described as follows.

Assume that the system is in a known equilibrium state x^* for a given vector λ^* of parameter values, so $f(\lambda^*, x^*) = 0$. If some or all of the parameters are perturbed, λ^* changes to $\lambda^* + \Delta\lambda$. If the system settles into a new equilibrium, say $x^* + \Delta x$, then this state must also be realizable, so $f(\lambda^* + \Delta, x^* + \Delta x) = 0$. The CGE model is an algorithm to solve this equation for Δx, given $\Delta\lambda$. In practice, f is a complicated function of many variables, so an "exact" solution is out of the question, except in very special cases. One must then resort to an approximation procedure, for example by linearizing the equation, $(D_\lambda f)^* \Delta x + (D_x f)^* \Delta\lambda = 0$.

However, there is a caveat; namely, if f is nonlinear, there is the possibility that the equation $f(\lambda, x) = 0$ has multiple solutions, and their number may even depend on λ. Each solution has its own basin of attraction, and a large change in the parameters can drive the system from one basin of attraction to another. The linear approximation (or any other, more sophisticated approximation) is essentially a *local* approximation, so it is not a priori clear that the CGE algorithm captures the correct solution. This problem persists if, instead of a one-step CGE procedure, one uses a multistep implementation, where the change $\Delta\lambda$ is split into several subintervals and the state variables are updated after each step.

A more realistic approach would be to take a global approach and analyze the bifurcations associated with the vector field f. Then one would know the critical parameter values, where the nature of the solution could change significantly, and follow the solution closely as parameters cross critical values. The study of bifurcation theory falls outside the scope of this chapter; we refer the reader to the literature [24, 34].

Today there are many CGE models of different systems; one of the better known is the GTAP model of world trade [27]. Recently, CGE has been a popular way to estimate the economic effects of measures to reduce greenhouse gas emissions and to estimate the effects of extreme weather events on certain components of the food system. The issue of missing and incomplete data comes up regularly in IO as well as in CGE models and will be discussed below, in Sect. 10.10.4.

Software packages are available to solve CGE models; among the better known are GAMS [19] and GEMPACK [26]. Their "black-box" characteristics tend to make them difficult to analyze [12, 22, 26].

Both IO models and CGE models face the problem of balancing economic data systems in the presence of incomplete data. There is an extensive economic literature to address this problem. A recent example using a maximum entropy approach can be found in [23]. A broader discussion of issues related to missing and incomplete data for food systems will be given below, in Sect. 10.10.4.

10.10 Data and Information Systems

Food system modeling is a data-driven activity. Data are essential to inform the mathematical models that help us understand the inner workings of a food system. This section is devoted to methods to assimilate data into a mathematical model and techniques to integrate the information in a timely manner and in a format that is useful for decision makers.

10.10.1 Data Assimilation

Data assimilation is the process of linking data with a mathematical model. It is an essential technique in any scientific discipline that is data-rich and for which well-founded mathematical predictive models exist. The technique originated in engineering and has found widespread application in many other disciplines, most notably in weather prediction, where it has extended the ability to predict weather more or less accurately from hours to days. The technique is used in a variety of modes; for example, to estimate state variables at a certain time using all available observational data, including those made at a later time (*reanalysis* or *smoothing mode*); or to estimate state variables using only past and present observations (*analysis* or *filtering mode*); or even to estimate state variables that are inaccessible to observations such as future states or states between measurements (*forecasting* or *predicting mode*). In any of these modes, problems can be approached with a variety of techniques, including optimization methods, maximum likelihood methods, and Bayesian methods. The following example illustrates the application of data assimilation methodology to reanalysis, filtering, and forecasting.

Consider a process with four real-valued state variables, $\{X_i : i = 1, \ldots, 4\}$. The state variables are related by the simple mathematical model

$$X_{i+1} = \alpha X_i + \xi_i, \quad i = 1, 2, 3, \tag{10.10.1}$$

where α is a known positive constant, and the random process error terms ξ_i are either identically zero or have a standard normal distribution, $\xi \sim N(0, 1)$. The data consist of two observations, $\{Y_i : i = 2, 3\}$, which are related to the state variables through the identities

$$Y_i = X_i + \zeta_i, \quad i = 2, 3, \tag{10.10.2}$$

where the random observation error terms ζ_i are independent and identically distributed with a standard normal distribution, $\zeta_i \sim N(0, \tau^2)$. Schematically, the entire model looks like this:

$$\begin{array}{ccccccc} X_1 & \Longrightarrow & X_2 & \Longrightarrow & X_3 & \Longrightarrow & X_4 \\ & & \Downarrow & & \Downarrow & & \\ & & Y_2 & & Y_3 & & \end{array}$$

The mathematical model is represented in the top row, the data model in the bottom row; the arrows represent the dependencies in the combined model. The problem of estimating X_1 from the observations Y_2 and Y_3 is a reanalysis problem, estimating X_2 from Y_2 or X_3 from Y_2 and Y_3 is a filtering problem, and estimating X_4 from Y_2 and Y_3 is a forecasting problem.

When the processes are time-dependent, the data may arrive as a time series—a sequence of realizations of a discrete-time stochastic process. In that case, filtering methods can be applied. A well-known filtering method is due to Kalman (*Kalman filtering*); extensions of the Kalman filter are the *Extended Kalman filter* and the *Ensemble Kalman Filter*. We refer the interested reader to the literature for details; see, for example, [31].

10.10.2 Data Visualization

Data visualization is a general term that describes any effort to help people understand the significance of data by placing it in a visual context. Patterns, trends, and correlations that might go undetected in text-based data can be exposed and recognized easier with data visualization.

A discussion of best ways to visually represent data falls outside the scope of the chapter, as does a discussion of available visualization software. We refer reader to the existing literature. The classical treatise by Tufte [57] is a good introduction.

While the Food and Agricultural Organization (FAO) of the United Nations maintains integrated (and interactive) near-real-time information systems on food and agriculture, they reflect only currently available data, include many dubious estimates of important quantities, and no estimates at all of other important quantities [17]. Moreover, information concerning the demand for food is fragmented and incomplete. For planning purposes, it would certainly be useful to have a dynamic and comprehensive display of energy flows throughout the global food delivery system.

In 2011, the International Commission on Sustainable Agriculture and Climate Change (ICSACC) issued a series of recommendations for improving global food security [3]. Of particular interest is Recommendation 7, to create a comprehensive, shared, integrated information system that encompasses human and ecological dimensions of agricultural and food systems. Such a system would track changes in land use, food production, climate, the environment, human health, and well-being worldwide by regular monitoring on the ground and by public-domain remote sensing networks.

10.10.3 Data Sources

There are numerous data sources for topics relevant to food systems and food security. But, as one might expect, they are heterogeneous, in various formats, spread across the globe or in cyberspace, and not always easily located. Since it is impossible to list a fair selection, we mention a few data sources to indicate the breadth of coverage at the international, national, and regional level.

At the international and multinational level, the Food and Agriculture Organization (FAO) of the United Nations provides food and agriculture data for over 245 countries and territories and covers all FAO regional groupings from 1961 to the most recent year available [16]. The World Bank maintains numerous country-level data sets and indicators that capture both the economic growth and the human priorities of ongoing development programs around the world [66]. The European Environmental Agency (EEA) gathers data and produces assessments on agriculture and topics related to the environment in the EU member states [13]. The International Food Policy Research Institute (IFPRI) has developed websites, portals, and applications built around open data from its own and external sources [28]. Data and appealing visualizations on international development issues of all kinds are available at the Gapminder website [20].

At the national level, the UK Government Department for Environment, Food and Rural Affairs (DEFRA) publishes a yearly Food Statistics Pocketbook with information about the food and farming industry in the United Kingdom [58]. In the United States, the Department of Agriculture (USDA) has several agencies that perform research to provide analysis and statistics, including the Economic Research Service (ERS), Foreign Agricultural Service (FAS), and National Agricultural Statistics Service (NASS). A catalog of publicly available USDA data can be found at [61]. Information about hunger in America can be found, for example, at the website of Feeding America, a not-for-profit network that is the nation's largest domestic hunger-relief organization [15].

Additional sources of data are the websites of non-government organizations (NGOs); state, provincial, and local agencies; and not-for-profit organizations such as producer associations, associations of food chain actors, and consumer organizations. Many universities have academic units whose mission includes the study of food systems, food policy, and food security. The University of Minnesota's Food Protection and Defense Institute maintains the World Factbook of Food, a reference repository of data related to food which provides the user with a wide range of food and agriculture data at the food and country level. Johns Hopkins University's Center for a Livable Future lists a number of Food Policy Networks [30].

Scraping the web is a useful technique to obtain information that is not directly available in published form. An example is found in [9], where data on interstate live cattle trade were obtained by scraping online records of cattle auction houses.

With current technology it is possible to develop visualization tools to study food systems and food security in the broader socioeconomic context. An example can

be found at the website [40], which displays data on demographic and economic diversity and food security in Montgomery County, Maryland.

Lastly, the report [62] includes four tables—one each on metrics, methodologies, data sources, and models—that provide samples of existing resources for assessing health, environmental, social, and economic effects of food systems (Appendix B).

10.10.4 Missing Data

Problems due to missing, incomplete, or otherwise corrupted data are common in economics and are especially pressing when modeling food systems. In developed economies, post-farm gate data about foodstuff being processed, transported, and distributed are generally detailed and accurate, since there is typically a business interest involved. But when these data are proprietary, they may be inaccessible. At the production level, data quality and availability are improving due to the rapid deployment of sensor technology, for example in precision agriculture. At the consumer level, fine-grained information about food purchases by individual households is sometimes available, for example from store checkout records. However, all these data sources come with their own problems. For example, remote-sensing data for precision agriculture may have observational gaps, and store checkout data may have geographical or social biases. Far less is known about the actual food consumption in households compared to, for example, food that is wasted. Typically, consumers do not record what is going into the trash or what is being left on the plate at a restaurant.

In developing economies, the logistics for gathering data are often prohibitively expensive. As a result, much less information is available, and if it is available, its reliability is often debatable.

10.11 Concluding Remarks

The purpose of this chapter is to introduce the mathematics research community to a range of problems in the areas of food systems and food security that offer opportunities for modeling, analysis, computational and data science.

A food system is a network with many actors and relationships, from the production stage through the distribution stage to the consumption stage. Modeling such a system is a challenge; yet, models are needed to improve our understanding of the system and thus the well-being of future generations. There are few, if any, universal laws; there are significant uncertainties, including climate change, population growth, and increasing urbanization; and there is little appreciation for the impact of food shortages and other catastrophic impacts on the social fabric.

In Sect. 10.2, we presented a framework for the study of food systems that reflects both issues of sustainability and social justice and matches the *Sustainable*

Development Goals (SDGs) adopted by the United Nations in 2015 [59]. After a general discussion of the challenges facing the development of mathematical models of food systems in Sect. 10.3, we devoted several sections to the various modeling techniques that have been used to study aspects of the food system, with case studies where appropriate (Sects. 10.4–10.9). In Sect. 10.10, we outlined data assimilation methods for integrating data into mathematical models and hinted at ways to improve the presentation of results that are scientifically justified and useful for decision makers. We concluded with a sampling of data sources available to the research community.

The problems are difficult, the tools are sparse, but the challenges must be faced if we want to contribute to one of the major existential problems facing humanity.

Acknowledgements Much of the work presented in this chapter was inspired by discussions at the weekly seminars of the SAMSI Working Group on Food Systems during the academic year 2017–2018. The authors thank professor Mary Lou Zeeman (Bowdoin College), who planted the seeds of our interest in food systems and contributed to an earlier draft of the chapter. The authors also thank the anonymous referees of this and an earlier version of the chapter for insightful comments and encouraging remarks.

References

1. Allen, M.: Planetary boundaries: tangible targets are critical. Nat. Rep. Clim. Chang. **3**(10), 114–115 (2009). https://doi.org/10.1038/climate.2009.95
2. Bass, S.: Planetary boundaries: keep off the grass. Nat. Rep. Clim. Chang. 113–114 (2009). https://doi.org/10.1038/climate.2009.94
3. Beddington, J., Asaduzzaman, M., Fernandez, A., et al.: Achieving food security in the face of climate change: summary for policy makers from the commission on sustainable agriculture and climate change. Technical report, CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS), Copenhagen (2011). https://cgspace.cgiar.org/bitstream/handle/10568/10701/Climate_food_commission-SPMNov2011.pdf?sequence=6
4. Bokhiria, J.B., Fowler, K.R., Jenkins, E.W.: Modeling and optimization for crop portfolio management under limited irrigation strategies. J. Agric. Environ. Sci. **3**, 209–237 (2014)
5. Boyer, P., Butault, J.: The food Euro: what food expenses pay for? Letter of the Observatory on formation of prices and margins of food products. FranceAgriMer **1**, 6 (2013)
6. Brewer, P.: Planetary boundaries: consider all consequences. Nat. Rep. Clim. Chang., 117–118 (2009). https://doi.org/10.1038/climate.2009.98
7. Canning, P., Rehkamp, S., Waters, A., Etemadnia, H.: The role of fossil fuels in the U.S. food system and the American diet. Technical Report. ERR-224, Department of Agriculture, Economic Research Service, Washington, DC (2017)
8. Canning, P., Weersink, A., Kelly, J.: Farm share of the food dollar: an IO approach for the United States and Canada. Agric. Econ. **47**, 505–512 (2016)
9. Carroll, I.T., Bansal, S.: Livestock market data for modeling disease spread among US cattle. bioRxiv Preprint (2015). http://dx.doi.org/10.1101/021980
10. Clark, C., Conrad, J.: Natural Resource Economics: Notes and Problems. Cambridge University Press, Cambridge (1997)
11. Costanza, R.: Ecological Economics: The Science and Management of Sustainability. Columbia University Press, New York (1992)

12. Dixon, P.B., Parmenter, B.R.: Computable general equilibrium modelling for policy analysis and forecasting. Handb. Comput. Econ. **1**, 3–85 (1996)
13. European Environmental Agency (EEA): Agriculture. https://www.eea.europa.eu/themes/agriculture
14. Fair, K.R., Bauch, C.T., Anand, M.: Dynamics of the global wheat trade network and resilience to shocks. Nat. Sci. Rep. **7**, 7177 (2017). https://doi.org/10.1038/s41598-017-07202-y
15. Feeding America: Hunger in America. http://www.feedingamerica.org/research/
16. Food and Agriculture Organization (FAO): Food and Agriculture Data. http://www.fao.org/faostat/en/
17. Food and Agriculture Organization (FAO): Information Systems for Food Security and Nutrition. http://www.fao.org/3/a-au836e.pdf
18. Food and Agriculture Organization (FAO): The state of food insecurity in the world 2014: strengthening the enabling environment for food security and nutrition. United Nations (2015)
19. GAMS Software GmbH.: General Algebraic Modeling System (GAMS), Frechen (2017). URL https://www.gams.com/docs/intro.htm
20. Gapminder: Gapminder. https://www.gapminder.org/
21. Garver, J., Goldberg, M.S.: Boundaries and Indicators: Conceptualizing and Measuring Progress Toward an Economy of Right Relationship Constrained by Global Economic Limits, chap. 5, pp. 149–190. Columbia University Press, New York (2015). https://doi.org/10.1038/climate.2009.93
22. Gillig, D., McCarl, B.A.: Introduction to Computable General Equilibrium Model (CGE): Course notes. Technical report, Department of Agricultural Economics, Texas A&M University (2002)
23. Golan, A., Judge, G., Robinson, S.: Recovering information from incomplete or partial multisectoral economic data. Rev. Econ. Stat. **LXXVI**(3), 541–549 (1994)
24. Guckenheimer, J., Holmes, P.: Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields, third printing, revised and corrected edn. Springer, New York (1990)
25. Guckenheimer, J., Ottino, J.M.: Foundations for complex systems research in the physical sciences and engineering. Technical report, U.S. National Science Foundation, Alexandria (2008)
26. Harrison, W.J., Pearson, K.R.: Computing solutions for large general equilibrium models using GEMPACK. Comput. Econ. **9**(2), 83–127 (1996)
27. Hertel, T.W., Hertel, T.W.: Global Trade Analysis: Modeling and Applications. Cambridge University Press, Cambridge (1997)
28. International Food Policy Research Institute (IFPRI): Agricultural S&T Indicators. http://library.ifpri.info/open-data/
29. Isard, W.: Gravity and Spatial Interaction Models, pp. 243–280. Routledge, New York (2017)
30. Johns Hopkins University, Center for a Livable Future: Food policy networks. http://www.foodpolicynetworks.org/
31. Kalnay, E.: Atmospheric Modeling, Data Assimilation and Predictability. Cambridge University Press, Cambridge (2003)
32. Krugman, P.: Increasing returns and economic geography. J. Polit. Econ. **99**(3), 483–499 (1991)
33. Malthus, T.R.: An essay on the principle of population. Reprint 2004. Edited with an introduction and notes by Geoffrey Gilbert (1798)
34. Meiss, J.D.: Differential Dynamical Systems, second, revised edn. MM22. SIAM, Philadelphia (2017)
35. Melitz, M.: The impact of trade on intra-industry reallocations and aggregate industry productivity. Econometrica **71**(6), 1695–1725 (2003)
36. Miller, R.E., Blair, P.D.: Input-Output Analysis: Foundations and Extensions, 2nd edn. Cambridge University Press, New York (2009)
37. Miller, B.W., Breckheimer, I., McCleary, A.L., Guzmn-Ramirez, L., Caplow, S.C., Jones-Smith, J.C., Walsh, S.J.: Using stylized agent-based models for population-environment research: a case study from the Galápagos Islands. Popul. Environ. **31**, 401–426 (2010)

38. Molden, D.: Planetary boundaries: the devil is in the detail. Nat. Rep. Clim. Chang., 116–117 (2009). https://doi.org/10.1038/climate.2009.97
39. Molina, M.J.: Planetary boundaries: identifying abrupt change. Nat. Rep. Clim. Chang., 115–116 (2009). https://doi.org/10.1038/climate.2009.96
40. Montgomery County, MD: Montgomery County FoodStat Application. https://countystat.maps.arcgis.com/apps/webappviewer/index.html?id=099052a140cd4bb38e99cbeb870ebce0
41. Myers, S.S., Zanobetti, A., Kloog, I., et al.: Rising CO2 threatens human nutrition. Nature **510**(7503), 139 (2014)
42. Newman, M.: The structure and function of complex networks. SIAM Rev. **45**, 167–255 (2003)
43. Ng, M., Fleming, T., Robinson, M., et al.: Global, regional, and national prevalence of overweight and obesity in children and adults during 1980–2013: a systematic analysis for the global burden of disease study 2013. Lancet **384**(9945), 766–781 (2014)
44. Raworth, K.: A safe and just space for humanity: can we live within the doughnut. Oxfam Policy Prac. Clim. Chang. Res. **8**(1), 1–26 (2012)
45. Raworth, K.: Why it's time for 'Doughnut Economics' (2014). https://www.youtube.com/watch?v=1BHOflzxPjI. TEDxAthens
46. Raworth, K.: Doughnut Economics: Seven Ways to Think Like a 21st-Century Economist. Chelsea Green Publishing, White River Junction (2017)
47. Reggiani, A.: Network resilience for transport security: some methodological considerations. Transp. Policy **28**, 63–68 (2013)
48. Rehkamp, S., Canning, P.: Measuring embodied blue water in American diets: An EIO supply chain approach. Ecol. Econ. **147**, 179–188 (2018)
49. Rockström, J., Steffen, W., Noone, K., et al.: A safe operating space for humanity. Nature **461**(7263), 472–475 (2009)
50. Samper, C.: Planetary boundaries: rethinking biodiversity. Nat. Rep. Clim. Chang. 118–119 (2009). https://doi.org/10.1038/climate.2009.98
51. Schlesinger, W.H.: Planetary boundaries: thresholds risk prolonged degradation. Nat. Rep. Clim. Chang. 112–113 (2009). https://doi.org/10.1038/climate.2009.93
52. Smith, V.L.: Relevance of laboratory experiments to testing resource allocation theory. In: Evaluation of Econometric Models, pp. 345–377. Elsevier, Amsterdam (1980)
53. Sraffa, P., with the collaboration of Maurice H. Dobb (eds.): The Works and Correspondence of David Ricardo, vol. I. Cambridge University Press, Cambridge (1951)
54. Steffen, W., Broadgate, W., Deutsch, L., et al.: The trajectory of the Anthropocene: the great acceleration. Anthropocene Rev. **2**(1), 81–98 (2015)
55. Steffen, W., Richardson, K., Rockström, J., et al.: Planetary boundaries: guiding human development on a changing planet. Science **347**(6223), 1259855 (2015)
56. Tisue, S., Wilensky, U.: NetLogo: A simple environment for modeling complexity. Technical report, Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston (2004). https://ccl.northwestern.edu/netlogo/
57. Tufte, E.: The Visual Display of Quantitative Information, 213 pp. Graphic Press, Cheshire (1973/2001)
58. UK Government, Department for Environment, Food and Rural Affairs (DEFRA): Food Statistics Pocketbook (2017). https://www.gov.uk/government/statistics/food-statistics-pocketbook-2017
59. United Nations: The Sustainable Development Goals Report, 2017. Technical report, United Nations, New York (2017). https://unstats.un.org/sdgs/files/report/2017/TheSustainableDevelopmentGoalsReport2017.pdf
60. United Nations, The European Commission, The Food and Agriculture Organization of the United Nations, The Organisation for Economic Co-operation and Development, The International Monetary Fund, The World Bank Group: System of Environmental–Economic Accounting 2012—Central Framework. United Nations, New York (2014). https://unstats.un.org/unsd/envaccounting/seearev/seea_cf_final_en.pdf
61. US Department of Agriculture (USDA): Data. https://www.usda.gov/topics/data

62. US National Academy of Sciences: A framework for assessing effects of the food system. Technical report, IOM (Institute of Medicine) and NRC (National Research Council), Washington, DC (2015)
63. von Thünen, J.H.: The Isolated State. Perthes, Hamburg (1826). English translation. Pergamon, Oxford (1966)
64. Wiedmann, T.O., Schandl, H., Lenzen, M., et al.: The material footprint of nations. Proc. Natl. Acad. Sci. **112**(20), 6271–6276 (2015)
65. Wikipedia: System dynamics. https://en.wikipedia.org/wiki/System_dynamics
66. World Bank: World Bank Open Data. http://data.worldbank.org

Chapter 11
Dynamic Optimization, Natural Capital, and Ecosystem Services

Jon M. Conrad

Abstract This article argues that natural capital should be viewed as a stock or state variable whose evolution is described by a difference or differential equation. Ecosystem services are benefit flows produced by stocks of natural capital. To value natural capital and the ecosystem services they provide, one needs to determine their value when they are optimally managed. This requires solving a dynamic optimization problem. The steady-state optimum to such a problem can serve as a benchmark from which to estimate the losses associated with pure open access (the tragedy of the commons) or any other sub-optimal steady state. This approach is illustrated by estimating the ecosystem service from oysters that remove nutrients from Chesapeake Bay.

Keywords Chesapeake Bay · Dynamic optimization · Ecosystem services · Natural capital · Oyster culture · Socioeconomics

11.1 Introduction

Since at least the mid-1960s, economists have been aware that natural resources and natural environments often provide benefit flows beyond the net revenues generated by their harvest, extraction, or visits by users. Burton Weisbrod [24] discussed the value of a park or wilderness area which provided utility flows to visitors (users) but which also provided value to non-users who wished to preserve their option to visit that site in the future. Samuelson [16], Hartman [8], and Calish et al. [3] viewed a standing forest as providing non-timber benefits. Stocks of nonrenewable resources, such as oil and natural gas, might generate conservation or option value. For example, Vousden [23] examined how a "conservation motive" might alter the optimal time path for extraction. Clark [4] posed a problem where a renewable

J. M. Conrad (✉)
Dyson School of Applied Economics and Management, Cornell University, Ithaca, NY, USA
e-mail: jmc16@cornell.edu

H. G. Kaper, F. S. Roberts (eds.), *Mathematics of Planet Earth*, Mathematics of Planet Earth 5, https://doi.org/10.1007/978-3-030-22044-0_11

resource provided existence value that would increase the optimal steady-state stock.

Public concern over the conservation of natural resources, environmental quality, biodiversity, and the impact of climate change has grown significantly since the 1960s, as has the literature in the field of resource and environmental economics. Despite increased recognition of the importance of natural capital and the ecosystem services they provide, the distinction between these two concepts and the appropriate methodology for their valuation is not widely understood outside the field of economics. The objective of this article is to illustrate the role of dynamic optimization as way to determine the optimal stocks of natural capital and to value the ecosystem services those optimal stocks provide. The solution of a dynamic optimization problem with stocks of natural capital can accomplish at least two things. One, it can provide a benchmark from which to measure the welfare losses from open access (the tragedy of the commons) or other sub-optimal states, and two, it can suggest optimal policies for restoring stocks of natural capital to their optimal or near-optimal levels.

Determining the value of natural capital and ecosystem services via dynamic optimization has not been the approach typically taken in the empirical literature. Empirical studies estimating the value of an ecosystem service are typically static; estimating a benefit flow at a point in time and at a particular location. There are four common methodologies used to estimate the value of an ecosystem service (1) hedonic pricing, where land value may be partially determined by the benefit flow from nearby natural capital, (2) travel cost, where individuals are willing to incur additional travel cost to visit a site with higher levels of an ecosystem service, (3) contingent valuation, where people are asked their willingness-to-pay for visits to sites with visibly different ecosystem services, and (4) alternative provision, where an ecosystem service is provided by an alternative means which has a known marginal cost. This last approach was taken by Grabowski et al. [7] to determine the value of nutrient removal by oysters in Chesapeake Bay based on the lowest marginal cost for removing those nutrients by reducing the runoff from agricultural operations or by a more thorough treatment of municipal waste.

More generally, Dasgupta [6] develops a measure of wealth that is based on manufactured capital, human capital (in the form of knowledge and human health), and natural capital, which provides ecosystem services. Welfare increases over time if and only if this comprehensive measure of wealth increases over time. Economists regard capital, whatever the form, as a stock or state variable whose level may change over time as a result of management or mis-management. Positively-valued capital stocks provide a flow of positively-valued services. (Negatively-valued stocks, for example stock pollutants, might produce negatively-valued flows in the form of damage to individuals or firms.)

Economists prefer to measure changes in social welfare relative to a benchmark of efficient or optimal resource allocation. If the current management of capital is not optimal, the loss or inefficiency of that allocation might be measured relative to the welfare supported by a first-best, or optimal, allocation of resources. With capital stocks, the optimal allocation, assuming that it is unique, must be determined by

solving a dynamic optimization problem. The solution to such a problem will allow the analyst to identify "shadow prices," also called "Lagrange Multipliers." Shadow prices can be interpreted as the marginal value of a slightly larger stock of natural capital and play a critical role in determining optimally sustainable (steady-state) rates of harvest.

In Sect. 11.2 we will develop a simple dynamic model where a renewable resource, the stock of oysters in northern Chesapeake Bay, provide net revenue to watermen (harvesters), but also provide an ecosystem service in the form of improved water quality through the removal of nutrients. In Sect. 11.3, we characterize the open access equilibrium, where access to the resource and the amount harvested are unregulated. The opportunity cost of open access can be calculated based on foregone net revenue and the reduced flow of ecosystem benefits when compared to the steady-state optimum. In Sect. 11.4, (1) functional forms are specified, (2) parameters are calibrated, (3) optimal and open access steady states are computed and compared, (4) sensitivity analysis is conducted, and (5) the optimal approach to steady state is described. Section 11.5 concludes.

It turns out that oysters actually provide multiple ecosystem services. In addition to the removal of nutrients, oyster reefs may provide habitat for other valuable species, for example, blue crab, as discussed in Peterson et al. [15] and Mykoniatis and Ready [14]. Oyster reefs may also provide shoreline protection, as examined by Grabowski et al. [7]. Our simple model just considers the value of nutrient removal, but in Sect. 11.5 we will suggest how additional forms of ecosystem services might be introduced to formulate a more complete, but also a more complex, dynamic optimization problem.

11.2 A Simple Model

Let $x = x(t)$ denote the stock of oysters (in bushels) and $h = h(t)$ the harvest (also in bushels) at instant t, $\infty > t \geq 0$. The change in the oyster population is given by the differential equation $dx/dt = \dot{x} = F(x) - h$, where $F(x)$ is a net growth function. Oysters are commercially valuable, and the net revenue at instant t (in dollars) is given by $\pi = \pi(t) = \pi(x, h)$.

The ability of oysters to remove nutrients from the waters of Chesapeake Bay gives rise to an ecosystem service in the form of improved water quality. Let $w = w(t)$ be an index of water quality, where arbitrarily $100 \geq w \geq 0$. Water quality is also a state variable whose change is given by the differential equation $dw/dt = \dot{w} = G(w, x; N)$. The change in water quality depends on the current level of water quality, w, the stock of oysters, x, and the inflow of nutrients into Chesapeake Bay, $N = N(t)$. To keep the model as simple as possible, we will assume that the inflow of nutrients is fixed at a total maximum daily load (TMDL) so that $N = \bar{N} > 0$. In this case, $\bar{N}$ will be a parameter and we can drop it as an argument of $G(\bullet)$ and simply write $dw/dt = \dot{w} = G(w, x)$.

The water quality index translates into a monetary net benefit according to the function $V = V(w)$, where $V(w)$, measured in dollars, is nondecreasing and strictly concave in w. Management of the oyster population must now account for both the net revenue from harvest, $\pi(x, h)$, and the net benefit of water quality, $V(w)$, that is positively influenced by the size of the oyster stock. The dynamic optimization problem of interest seeks to

$$\begin{array}{ll} \text{maximize}_{\{h\}} & \int_0^\infty [\pi(x,h) + V(w)]e^{-\delta t}\, dt, \\ \text{subject to} & \dot{x} = F(x) - h, \\ & \dot{w} = G(w, x), \\ \text{given} & x(0) > 0,\ w(0) > 0, \end{array}$$

where $\delta > 0$ is the instantaneous rate of discount, and $x(0)$ and $w(0)$ are the initial conditions for the oyster stock and the water quality index, respectively. The current-value Hamiltonian may be written as

$$H = \pi(x, h) + V(w) + \mu_x[F(x) - h] + \mu_w G(w, x), \tag{11.2.1}$$

where μ_x and μ_w are the current-value shadow prices on the oyster stock and water quality, respectively. The Maximum Principle, specifying conditions that must be satisfied by optimal harvest, requires

$$\partial H/\partial h = \pi_h(\bullet) - \mu_x = 0, \tag{11.2.2}$$

$$\dot{\mu}_x - \delta\mu_x = -\partial H/\partial x = -[\pi_x(\bullet) + \mu_x F'(x) + \mu_w G_x(\bullet)], \tag{11.2.3}$$

$$\dot{\mu}_w - \delta\mu_w = -\partial H/\partial w = -[V'(w) + \mu_w G_w(\bullet)], \tag{11.2.4}$$

$$\dot{x} = \partial H/\partial \mu_x = F(x) - h, \tag{11.2.5}$$

$$\dot{w} = \partial H/\partial \mu_w = G(w, x), \tag{11.2.6}$$

$$\lim_{t\to\infty} e^{-\delta t}\mu_x x = 0, \tag{11.2.7}$$

$$\lim_{t\to\infty} e^{-\delta t}\mu_w w = 0, \tag{11.2.8}$$

where $\pi_x(\bullet) = \partial\pi(x, h)/\partial x$, $\pi_h(\bullet) = \partial\pi(x, h)/\partial h$, $G_x(\bullet) = \partial G(w, x)/\partial x$, $G_w(\bullet) = \partial G(w, x)/\partial w$, and where $F'(x)$ and $V'(w)$ are the first derivatives of the functions $F(x)$ and $V(w)$, respectively. Equation (11.2.2) requires that the marginal net revenue from harvest be set equal to the shadow price on the oyster population, or $\pi_h(\bullet) = \mu_x$. Equation (11.2.3) requires that the interest income from liquidating one unit of the oyster stock, $\delta\mu_x$, must equal the capital gain on the shadow price of oysters, $\dot{\mu}_x$, plus the sum of the marginal net revenue, the value of marginal net growth, and the marginal value of the contribution to water quality, or

$$\delta\mu_x = \dot{\mu}_x + \pi_x(\bullet) + \mu_x F'(x) + \mu_w G_x(\bullet). \tag{11.2.9}$$

Equation (11.2.4) places a similar requirement on any marginal reduction in water quality. The interest payment on the shadow price for water quality must be equated to the capital gain in that shadow price, plus the sum of the marginal monetary value of w, $V'(w)$ and the marginal value in its dynamics, or

$$\delta\mu_w = \dot{\mu}_w + V'(w) + \mu_w G_w(\bullet). \tag{11.2.10}$$

Equations (11.2.5) and (11.2.6) simply recover the state equations for x and w. Equations (11.2.7) and (11.2.8) are the transversality conditions that are required for convergence of the present-value integral in this infinite-horizon, dynamic optimization problem.

The benchmark for efficient management and for formulating first-best, public policy is the stationary optimum obtained by evaluating Eqs. (11.2.2)–(11.2.6) in steady state. These equations would then require

$$\mu_x = \pi_h(x, h), \tag{11.2.11}$$

$$\mu_w = \frac{V'(w)}{\delta - G_w(w, x)}, \tag{11.2.12}$$

$$\delta = F'(x) + \frac{\pi_x(x, h) + \mu_w G_x(w, x)}{\mu_x}, \tag{11.2.13}$$

$$h = F(x), \tag{11.2.14}$$

$$G(w, x) = 0. \tag{11.2.15}$$

Equations (11.2.11)–(11.2.15) constitute a five-equation system defining a steady-state optimum, $[x^*, w^*, h^*, \mu_x^*, \mu_w^*]$. By imposing strict concavity in the functions $\pi(x, h)$, $F(x)$, $V(w)$, and $G(w, x)$, the steady-state optimum will be unique and saddle-point stable with the approach being either most rapid (a most rapid approach path, or MRAP) or asymptotic.

We can make the problem of solving for the steady-state optimum a bit easier by substituting Eqs. (11.2.11) and (11.2.12) into Eq. (11.2.13),

$$\delta = F'(x) + \frac{\pi_x(\bullet)(\delta - G_w(\bullet)) + G_x(\bullet)V'(w)}{\pi_h(\bullet)(\delta - G_w(\bullet))}. \tag{11.2.16}$$

Then, Eqs. (11.2.14)–(11.2.16) become a three-equation system in $[x^*, w^*, h^*]$. With functional forms for $\pi(x, h)$, $F(x)$, $V(w)$, and $G(w, x)$ and their associated parameter values, it is relatively easy to design an algorithm to solve this three-dimensional system for $[x^*, w^*, h^*]$ and then use Eqs. (11.2.11) and (11.2.12) to calculate the shadow prices, μ_x and μ_w. We will illustrate this procedure in Sect. 11.4, but first we want to identify the pure open access equilibrium (steady state) and show how the opportunity cost of open access might be calculated based on shadow prices, net revenue, and the monetary value of water quality.

11.3 Pure Open Access

In pure open access, competitive watermen are unregulated and will harvest the oyster stock down until profit is zero, $\pi(x, h) = 0$. Denote the steady-state, open access oyster stock as x_∞. Open access will typically result in economic overfishing such that $x_\infty < x^*$, where x^* is the optimal steady-state biomass defined by Eqs. (11.2.14)–(11.2.16) above. The steady-state, open access harvest is $h_\infty = F(x_\infty)$, and again, it is typically the case that $h_\infty < h^*$. Knowing $[x_\infty, h_\infty]$, we can compute w_∞ by solving $G(w, x_\infty) = 0$. In the pure open access equilibrium, μ_x, the shadow price on the oyster stock is zero and a valuable resource has been rendered worthless. The cost of open access at instant t is measured by the loss in net revenue and reduced ecosystem services. In our simple model this loss may be calculated as

$$C_{\text{open access}} = \pi(x^*, h^*) + [V(w^*) - V(w_\infty)]. \tag{11.3.1}$$

The lesson in Eq. (11.3.1) is that we need to solve the dynamic optimization problem in order to determine the extent of the losses (foregone net revenue and ecosystem services) associated with pure open access or any other sub-optimal equilibrium.

11.4 The Cost of Open Access

In this section we (1) specify functional forms, (2) calibrate parameters, (3) solve for the steady-state optimum, (4) determine the open access equilibrium, (5) calculate the cost of open access for the base-case set of parameters, (6) conduct sensitivity analysis, and (7) describe the optimal approach to the steady-state optimum.

11.4.1 Functional Forms

There are four functions in our simple model from Sect. 11.2: $F(x)$, $\pi(x, h)$, $G(w, x)$, and $V(w)$. The following forms are adopted: $F(x) = rx(1 - x/K)$, $\pi(x, h) = (p - c/x)h$, $G(w, x) = -\gamma(\bar{N})w + \beta x$, and $V(w) = \alpha \ln(1 + w)$, where $\ln(\bullet)$ is the natural log operator. In this specification, the oyster population grows logistically with $r > 0$ being the intrinsic growth rate and $K > 0$ being the carrying capacity of northern Chesapeake Bay, in bushels of oysters.

Net revenue is linear in harvest, h, also measured in bushels of oysters, where $p > 0$ is the exvessel price/bushel, and $c > 0$ is a cost parameter. The term c/x indicates that harvest cost declines with larger oyster populations.

The function $G(w, x) = -\gamma(\bar{N})w + \beta x$ says that the rate of decline in water quality depends on the size of the nutrient inflow, $\bar{N}$, with larger inflows resulting in

larger values for γ. Because we have assumed that $\bar{N}$ is constant at the TDML, the value for γ is a constant with $1 > \gamma > 0$. Each bushel of oysters can improve the water quality index according to the parameter β, where $1 > \beta > 0$.

Finally, water quality in the northern Chesapeake Bay is a local public good with higher values for w resulting in greater welfare for boaters, swimmers, fishers, birders, and non-users who derive utility from higher quality water. $V(w) = \alpha \ln(1 + w)$, with $\alpha > 0$, implies that the dollar annual benefit from water quality improvement is strictly concave in w.

With these functional forms, Eq. (11.2.16) takes the form

$$\delta = r(1 - 2x/K) + \frac{cr(1 - x/K)(\delta + \gamma) + \alpha\beta\gamma x/(\gamma + \beta x)}{(px - c)(\delta + \gamma)}, \tag{11.4.1}$$

so x^* is the positive root of $G(x) = 0$, where G is a cubic polynomial in x,

$$G(x) = [r(1-2x/K)-\delta](px-c)(\delta+\gamma)+cr(1-x/K)(\delta+\gamma)+\alpha\beta\gamma x/(\gamma+\beta x). \tag{11.4.2}$$

For the base-case parameter values given in the next subsection, *Mathematica* computes two negative roots and one positive root. This is also the case for all parameter combinations in our sensitivity analysis. We adopt the positive root as our value for x^*.

11.4.2 Parameter Values

There are seven parameters in our functional forms: r, K, p, c, α, β, and γ. To these seven parameters we must also specify a value for δ, the instantaneous discount rate.

The oyster population in Chesapeake Bay has been studied extensively. While growth can vary significantly with salinity, dissolved oxygen, disease (MSX and Dermo), and the height and extent of an oyster reef, Kaspersky and Wieland [9] adopt the values $r = 0.239$ and $K = 5.089 \times 10^9$ market-sized oysters. There are approximately 350 market-sized oysters per bushel, so our value for carrying capacity is $K = 14{,}540{,}000$ bushels.

Table 11.1 contains the harvest (in bushels), dockside value (in dollars), and the exvessel price per bushel for the 1989–1990 through 2012–2013 seasons. The oyster season runs from October 1 through March 31 of the following year. The average price over this period was \$23.82/bu with a standard deviation of \$5.86/bu. We adopt the average price as our base-case value for p.

The calibration of the cost parameter, c, is based on the average price, $p =$ \$23.82/bu, and a guess for the open access equilibrium stock, x_∞. Net revenue is driven to zero at the pure open access equilibrium stock implying $\pi(x, h) = (p - c/x)h = 0$. If open access harvest, $h_\infty > 0$, then $x_\infty = c/p$ and $c = px_\infty$. Open access equilibrium stocks are often less than 20% of carrying

Table 11.1 Season, total harvest, dockside value, and exvessel price

Season	Total harvest (bu)	Dockside value ($)	Exvessel price ($/bu)
1989–1990	414,445	9,900,000	23.89
1990–1991	418,393	9,400,000	22.47
1991–1992	323,189	6,400,000	19.80
1992–1993	123,618	2,600,000	21.03
1993–1994	79,618	1,400,000	17.58
1994–1995	164,641	3,200,000	19.44
1995–1996	199,798	3,200,000	16.02
1996–1997	177,600	3,800,000	21.40
1997–1998	284,980	5,700,000	20.00
1998–1999	423,219	7,800,000	18.43
1999–2000	380,675	7,200,000	18.91
2000–2001	347,968	6,800,000	19.54
2001–2002	148,155	2,900,000	19.57
2002–2003	55,840	1,600,000	28.65
2003–2004	26,471	700,000	26.44
2004–2005	72,218	1,100,000	15.23
2005–2006	154,436	4,700,000	30.43
2006–2007	165,059	5,000,000	30.29
2007–2008	82,958	2,600,000	31.34
2008–2009	101,141	2,700,000	26.70
2009–2010	185,245	4,500,000	24.29
2010–2011	123,613	4,300,000	34.79
2011–2012	137,317	4,600,000	33.50
2012–2013	341,132	10,900,000	31.95

Source: Maryland oyster population status report [12]

capacity, $x_\infty/K < 0.20$. We set $x_\infty = 0.15K = 2{,}181{,}000$ bushels. This implies $c = px_\infty = \$51{,}951{,}420$ and $h_\infty = 443{,}070$ bushels.

The parameters β and γ influence the change in the water quality index according to the differential equation $\dot{w} = -\gamma w + \beta x$. We calibrate γ assuming that the water quality index 100 years ago would have been $w_{1916} = 100$ and that the water quality index today is approximately 10 ($w_{2015} \approxeq 10$). If we suppress the rate of removal of nutrients by oysters, this would imply that $10 \approxeq 100(1-\gamma)^{100}$. Solving for gamma yields $\gamma \approxeq 0.02276$. We assume further that $w_\infty = 10 = (\beta/\gamma)x_\infty$ when $x_\infty = 2{,}181{,}000$ bushels and $\gamma = 0.02276$. This implies that $\beta = 1.0436 \times 10^{-7}$. We round to $\beta = 1.04 \times 10^{-7}$, which results in a base-case $w_\infty = 9.97$, close enough to 10.

The water quality of Chesapeake Bay has also been extensively studied. Cropper and Isaac [5] review 16 studies conducted between 1985 and 2009 examining the economic value of improved water quality in Chesapeake Bay or in specific rivers or creeks that flow into Chesapeake Bay. The studies employed different valuation methodologies to estimate the benefits from improved water quality to (1) property

Table 11.2 Base-case parameter values

Parameter	Definition	Base-case value
p	Exvessel price per bushel	$23.82
x_∞	Open Access Stock	2,181,000 bushels
$c = px_\infty$	Cost parameter	$51,951,420
r	Intrinsic growth rate	0.239
K	Carrying capacity	14,540,000 bushels
α	Water quality value parameter	$21,667,907
β	Nutrient removal rate per bushel	$1.04x10^{-7}$
γ	Decay rate for water quality index	0.02276
δ	Risk-free discount rate	0.02

owners, (2) recreational users, (3) watermen (harvesters), and (4) non-users (who derive option or existence value). The valuation methods included (1) contingent valuation, (2) benefit transfer, (3) hedonic pricing, (4) travel cost, and (5) bio-economic models.

The studies that estimated benefits for the entire Bay or for households in the mid-Atlantic region are the most relevant for calibration of α. Van Houtven [22] estimates that waterfront property owners would see an increase in annual benefits of between $38.7 to $102.2 million from a marginal reduction in dissolved inorganic nitrogen (DIN). Bockstael et al. [1], [2] estimate an annual benefit of $77.1 million for a 40% reduction in the product of nitrogen and phosphorous concentrations (TNP). Krupnick [10] estimates annual benefits of $103.9 million to recreational users from a 40% reduction in TNP. Morgan and Owens [13] estimate an annual benefit of $1.25 billion for a 60% improvement in TNP. Finally, Lipton and Hicks [11] estimate an annual non-use (existence) benefit to mid-Atlantic households of $131.5 million.

We will adopt a conservative calibration for α. Suppose that the annual value of water quality when $w = 100$ is $100,000,000. Given the functional form for $V(w)$, this implies that $\alpha \ln(101) = \$100,000,000$ and that $\alpha = \$100,000,000/\ln(101) = \$21,667,907$.

Finally, we adopt a discount rate of $\delta = 0.02$. Weitzman [25] found that the subjective distribution for the risk-free rate of discount, based on a survey of 2160 economists in 48 counties, was a close fit to a gamma distribution with a modal value of $\delta = 0.02$. The base-case parameter values are summarized in Table 11.2.

11.4.3 Results

Table 11.3 presents the values for $[x^*, h^*, w^*, \mu_x^*, \mu_w^*, x^*/K, h^*/x^*, \pi^*, V(w^*)]$ at the steady-state optimum, $[x_\infty, h_\infty, w_\infty, V(w_\infty)]$ under pure open access, and the cost of open access in terms of foregone net revenue and ecosystem (water quality) services. Recall from Eq. (11.3.1) that $C_{\text{open access}} = \pi(x^*, h^*) + [V(w^*) - V(w_\infty)]$.

Table 11.3 The steady-state optimum and the cost of open access

Variable	Value
x^*	9,420,834 bushels
h^*	792,724 bushels
w^*	43.05
μ_x^*	\$18.31
μ_w^*	\$11,504,177
x^*/K	0.65
h^*/x^*	0.08414
π^*	\$14,511,192
$V(w^*)$	\$82,018,970
x_∞	2,181,000 bushels
h_∞	443,070 bushels
w_∞	9.97
π_∞	0
$V(x_\infty)$	\$51,890,107
$C_{\text{open access}}$	\$44,640,055

Movement from the pure open access equilibrium to the steady-state optimum would result in an increase in annual net benefit of \$44,640,055. This amount is the cost of open access, comprised of $\pi(x^*, h^*) = \$14{,}511{,}192$ in net revenue from harvest at the steady-state optimum and the difference in water quality ecosystem service of $[V(w^*) - V(w_\infty)] = \$30{,}128{,}863$.

11.4.4 Sensitivity Analysis

Table 11.4 provides sensitivity analysis for changes in α, β, γ, δ, K, r, and p. The base-case values for α, β, γ, δ, and p are doubled, while the base-case value of K, is halved. The base-case values from Table 11.3 are listed in Table 11.4 for comparison. We retain $x_\infty = 0.15K$. Then $\pi_\infty = 0$ requires that $c = px_\infty = 0.15pK$ and because c depends on p and K we do not vary that parameter independently.

Perhaps the first thing to note in Table 11.4 is that the ratio of the optimal stock to carrying capacity, x^*/K, is relatively insensitive to changes in α, β, γ, δ, K, r, and p. It ranges from 0.59 to 0.73. When the water quality value parameter, α, is doubled, it significantly increases the cost of open access which goes from \$44,640,055 to \$78,278,762. The water quality index is quite sensitive to changes in β, the rate at which oysters remove nutrients. In the base-case, $w^* = 43.05$, while a doubling of β increases w^* to $w^* = 86.23$.

A doubling of γ also has a significant effect on the water quality index. Recall that γ is a function of the annual nutrient loading. If that loading were to increase it would cause an increase in γ, which would cause a more rapid decline in the water

Table 11.4 Sensitivity analysis

Variable	Base-Case	$\alpha = \$43,355,813$	$\beta = 2.08x10^{-7}$	$\gamma = 0.04552$	$\delta = 0.04$	$K = 7,270,000$	$r = 0.478$	$p = \$47.64$
x^*	9,420,834	10,600,977	9,435,741	9,771,010	8,591,584	5,278,477	8,935,038	8,727,317
h^*	792,724	686,385	791,666	765,948	840,056	345,587	1,646,389	833,856
w^*	43.05	48.44	86.23	22.32	39.26	24.12	40.83	39.88
μ_x^*	$18.31	$18.92	$18.31	$18.50	$17.77	$18.90	$18.01	$35.73
μ_w^*	$11,504,177	$20,508,239	$5,809,046	$14,178,865	$8,575,824	$20,172,834	$12,114,703	$12,395,994
x^*/K	0.65	0.73	0.65	0.67	0.59	0.73	0.61	0.60
h^*/x^*	0.08414	0.06474	0.08390	0.07838	0.09777	0.06547	0.18426	0.09554
π^*	$14,511,192	$12,985,978	$14,498,727	$14,172,420	$14,930,494	$6,531,228	$29,644,315	$29,797,438
$V(w^*)$	$82,018,970	$169,120,892	$96,824,518	$68,242,631	$80,069,905	$69,849,696	$80,898,529	$80,401,175
x_∞	2,181,000	2,181,000	2,181,000	2,181,000	2,181,000	1,090,500	2,181,000	2,181,000
h_∞	443,070	443,070	443,070	443,070	443,070	221,535	886,140	443,070
w_∞	9.97	9.97	19.93	4.98	9.97	4.98	9.97	9.97
π_∞	0	0	0	0	0	0	0	0
$V(w_\infty)$	$51,890,107	$103,828,107	$65,897,956	$38,762,026	$51,890,107	$38,762,026	$51,890,107	$51,890,107
$[V(w^*) - V(w_\infty)]$	$30,128,863	$65,292,785	$30,926,562	$24,589,606	$28,179,798	$31,087,670	$29,008,422	$28,511,068
$C_{\text{open access}}$	$44,640,055	$78,278,762	$45,425,289	$43,653,025	$43,110,292	$37,618,898	$58,652,738	$58,308,505

quality index for a given oyster population. The doubling of γ causes w^* to decline from $w^* = 43.05$ in the base-case to $w^* = 22.32$.

As expected, an increase in the discount rate, δ, causes x^* to decline. Because the base-case x^*, when $\delta = 0.02$, and the value of x^* when $\delta = 0.04$ are both above the maximum sustainable yield stock of $x_{msy} = K/2$, the optimal harvest, h^*, increases from $h^* = 792{,}724$ to $h^* = 840{,}056$, even though x^* has declined from $x^* = 9{,}420{,}834$ to $x^* = 8{,}591{,}584$.

Doubling the intrinsic growth rate, r, allows for a significant increase in the level of harvest, even though x^* has declined from the base case. Harvest goes from $h^* = 792{,}724$ in the base case to $h^* = 1{,}646{,}389$ when r has doubled to $r = 0.478$. Annual net revenue from this increased harvest increases to $\pi^* = \$29{,}644{,}315$.

Finally, a doubling of the exvessel price per bushel of oysters reduces x^*, increases h^*, and results in a big bump in annual net revenue to $\pi^* = \$29{,}797{,}438$. The smaller standing stock of oysters, $x^* = 8{,}727{,}317$, lowers steady-state water quality to $w^* = 39.88$ causing a decline in value from $V(w^*) = \$82{,}018{,}970$ in the base case to $\$80{,}401{,}175$ when the oyster price has doubled.

The take-home from this sensitivity analysis would be (1) the value of water quality significantly affects the value of ecosystem services, (2) the nutrient loading and the rate of nutrient removal by oysters significantly affects the water quality index, w, (3) the intrinsic growth rate will significantly affect the level of harvest and net revenue in the oyster fishery in Chesapeake Bay, and (4) a doubling of the exvessel price will have a significant affect on net revenue.

11.4.5 Approach Dynamics

Net revenue, $\pi = \pi(t) = (p - c/x)h$, is linear in harvest, $h = h(t)$. In a model with a single state variable, $x = x(t)$, this would imply that the optimal approach to the steady-state optimum (the singular solution) would be most rapid (i.e., the most rapid approach path or MRAP). If $x(t) < x^*$, $h^* = 0$ until the stock grows to its optimal level, x^*, at which time $h(t) = h^* = rx^*(1 - x^*/k)$ for the rest of time.

With two or more state variables it may be optimal to maintain the moratorium on harvest until both state variables have reached their steady-state optimal levels. Suppose that $x(0) < x^*$ and $w(0) < w^*$. This will imply that $p - c/x - \mu_x < 0$ and $h^* = 0$. Suppose further that zero harvest allows $x = x(t)$ to reach x^* before $w = w(t)$ reaches w^*. Recall that the instantaneous net benefit was $U = (p - c/x)h + \alpha \ln(1+w)$. The change in net benefit for h constant is $dU/dt = (c/x^2)h\dot{x} + [\alpha/(1 + w)]\dot{w}$ or

$$dU/dt = (c/x^2)h[rx(1 - x/K) - h] + [\alpha/(1 + w)](-\gamma w + \beta x). \qquad (11.4.3)$$

For $h^* = 0$ or $h^* = rx^*(1 - x^*/K)$, the first term on the right-hand side of Eq. (11.4.3) will be zero. The second term on the right-hand side of Eq. (11.4.3) will be positive if $w^* > w(t) > 0$ and $x(t) \geq x^*$ and zero when $w(t) = w^*$ and $x(t) =$

x^*. In this case the optimal approach is to maintain the moratorium on harvest until $w(t)$ reaches w^*, then instantaneously harvest $[x(t) - x^*] + rx^*(1 - x^*/K)$ and then harvest $h^* = rx^*(1-x^*/K)$ thereafter. The intuition is that the value of growth in the water quality index under the extended moratorium exceeds the foregone net revenue at the steady-state optimal harvest when $w(t) < w^*$.

11.5 Conclusions and Caveats

A simple dynamic optimization model was developed to sharpen the distinction between natural capital (a stock or state variable) and ecosystem service (a flow variable). To rigorously value natural capital and ecosystem services one needs to solve a dynamic optimization problem. This is required because natural capital, like other capital stocks, must be optimally managed over time to reach its full potential value to society.

Stocks of natural resources are perhaps the most obvious forms of natural capital. Forests might provide timber, but they also provide habitat for wildlife, stabilize soils, sequester carbon, and support recreational activities. Stock of fish and shellfish can be sustainably harvested, while certain species, such as the eastern oyster, *Crassostrea virginica*, provide ecosystem services by (1) removing nutrients which would otherwise degrade water quality, (2) building reefs which provide habitat for other valuable marine species, and (3) protecting shorelines during storm surge. Our simple dynamic model only considered the ecosystem service from nutrient removal and improved water quality. Specifying plausible functional forms and calibrating the eight parameters to the dynamic optimization problem permitted the numerical calculation of a unique steady-state optimum and the pure open access equilibrium, where net revenue (profit or rent) was driven to zero. Sensitivity analysis yielded logical comparative statics and estimates for the cost of pure open access ranging between $37 million and $79 million per year.

Because of the simplicity of the dynamic model, the numerical results should be taken with a grain of salt. It would be possible to build models with multiple stocks of natural capital and multiple flows of ecosystem services. For example, adding the stock of blue crabs would introduce a third state variable and require a state equation describing the dynamics of the blue crab population. Mykoniatis and Ready [14] have the oyster population increasing the Chesapeake Bay carrying capacity for blue crab. Peterson, Grabowski, and Powers [15] identify 11 other species whose growth or carrying capacity might be enhanced by oyster reefs in the southeastern USA.

Estuarine and marine ecosystems are notoriously noisy. Instead of a single annual or present value, stochastic models of optimal capital management will yield stationary distributions for stocks and ecosystem services.

Finally, spatial considerations will almost certainly come into play, particularly if it is possible to invest in restoration, or if restoration can be enhanced through permanent or temporary marine reserves where harvest is prohibited. Sanchirico and Wilen [17, 18], Smith and Wilen [19, 20], and Smith, Sanchirico, and Wilen [21]

examine the economic role of marine reserves in a spatial-dynamic model. In such models the value of natural capital will vary by location.

These complexities do not invalidate dynamic optimization as the appropriate methodology for valuing natural capital and ecosystem services, it just makes it much more difficult. The analysis of large-scale, stochastic-dynamic-spatial models of natural capital will likely be based on extensive numerical simulation. The insight and intuition from simple models, such as the one presented here, can help in interpreting the simulation results of larger, more complex models, seeking a better approximation to reality.

References

1. Bockstael, N.E., McConnell, K.E., Strand, I.E.: Benefits from improvements in chesapeake bay water quality. Technical report, U.S. Environmental Protection Agency, Washington (1988)
2. Bockstael, N.E., McConnell, K.E., Strand, I.E.: Measuring the benefits of improvements in water quality: the chesapeake bay. Mar. Resour. Econ. **6**, 1–18 (1989)
3. Calish, S., Fight, R., Teeguarden, D.E.: How do nontimber values affect Douglas fir rotations? J. For. **76**, 217–222 (1978)
4. Clark, C.W.: Mathematical Bioeconomics: the optimal management of renewable resources. Wiley, New York (1976)
5. Cropper, M.L., Isaac, W.: The benefits of achieving the chesapeake bay tmdls (total maximum daily loads). Technical report, Resources for the Future (RFF), Washington, D.C. (2011). Discussion Paper
6. Dasgupta, P.: Nature's role in sustainable economic development. Philos. Trans. R. Soc. B **365**, 5–11 (2010)
7. Grabowski, J.H., Burmbaugh, R.D., Conrad, R.F., et al.: Economic valuation of ecosystem services provided by oyster reefs. BioScience **62**, 900–909 (2012)
8. Hartman, P.R.: The harvesting decision when a standing forest has value. Econ. Inq. **14**, 52–58 (1976)
9. Kasperski, S., Wieland, R.: When is it optimal to delay harvesting? The role of ecological services in the northern Chesapeake Bay oyster fishery. Mar. Resour. Econ. **24**, 361–385 (2009)
10. Krupnick, A.: Reducing bay nutrients: An economic perspective. Md. Law Rev. **47**, 453–480 (1988)
11. Lipton, D., Hicks, R.W.: The economic benefits of oyster reef restoration in the chesapeake bay, final report. Technical report, Chesapeake Bay Foundation, Annapolis (2004)
12. Maryland Department of Natural Resources: The 2013 fall survey. Technical report, Maryland Department of Natural Resources, Annapolis (2014)
13. Morgan, C., Owens, N.: Benefits of water quality policies: the Chesapeake Bay. Ecol. Econ. **39**, 271–284 (2001)
14. Mykoniatis, N., Ready, R.: Optimal oyster management in Chesapeake Bay incorporating sanctuaries, reserves, aquaculture, and externalities. In: Selected Papers from the 2012 AAEA Meetings (2012)
15. Peterson, C.H., Grabowski, J.H., Powers, S.P.: Estimated enhancement of fish production resulting from restoring oyster reef habitat: quantitative valuation. Mar. Ecol. Prog. Ser. **264**, 249–264 (2003)
16. Samuelson, P.A.: Economics of forestry in an evolving society. Econ. Inq. **14**, 466–492 (1976)
17. Sanchirico, J.N., Wilen, J.E.: A bioeconomic model of marine reserve creation. J. Environ. Econ. Manag. **42**, 257–276 (2001)

18. Sanchirico, J.N., Wilen, J.E.: The impacts of marine reserves on limited entry fisheries. Nat. Resour. Model. **15**, 291–310 (2002)
19. Smith, M.D., Wilen, J.E.: Economic impacts of marine reserves: the importance of spatial behavior. J. Environ. Econ. Manag. **46**, 183–206 (2003)
20. Smith, M.D., Wilen, J.E.: Marine reserves with endogenous ports: empirical bioeconomics of the California sea urchin fishery. Mar. Resour. Econ. **19** (2004)
21. Smith, M.D., Sanchirico, J.N., Wilen, J.E.: The economics of spatial-dynamic processes: applications to renewable resources. J. Environ. Econ. Manag. **57**, 104–121 (2009)
22. van Houten, G.L.: Changes in ecosystem services associated with alternative levels of ecological indicators. Technical Report EPA-452/R-09-008b, Environmental Protection Agency (EPA), RTI International, Research Triangle Park, North Carolina (2009)
23. Vousden, N.: Basic theoretical issues of of resource depletion. J. Econ. Theory **6**, 126–143 (1973)
24. Weisbrod, B.A.: Collective-consumptive services of individual consumption goods. Q. J. Econ. **78**, 457–470 (1964)
25. Weitzman, M.L.: Gamma discounting. Am. Econ. Rev. **91**, 260–271 (2001l)

Chapter 12
Quantitative Models for Infrastructure Restoration After Extreme Events: Network Optimization Meets Scheduling

Thomas C. Sharkey and Sarah G. Nurre Pinkley

Abstract This chapter focuses on the recovery of critical infrastructure systems from large-scale disruptive events and shows how optimization can help guide decision makers in the restoration process. The operation of an infrastructure system is modeled as a network flow problem, which can be used to assess the impact of the disruption on the services provided by the system. To restore the disrupted services, decision makers must schedule the repair operations by allocating scarce resources such as work crews and equipment over time. An overview of the relevant areas of network flows and scheduling is followed by a discussion of how techniques from network optimization and scheduling can be integrated to quantitatively model infrastructure restoration.

Keywords Disruption · Extreme event · Infrastructure · Network flow · Optimization · Restoration · Scheduling

12.1 Critical Infrastructures and Extreme Events

Critical infrastructure systems provide key services to a community and help to ensure the safety, comfort, and well-being of its citizens. Examples of critical infrastructure systems include electrical power systems, transportation systems, telecommunications, water supply systems, and wastewater systems. An important responsibility of the managers of these systems is to ensure their services are restored efficiently after they have been disrupted by an extreme event. Extreme events such as hurricanes, earthquakes, and tsunamis can cause catastrophic damage

T. C. Sharkey (✉)
Department of Industrial and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY, USA
e-mail: sharkt@rpi.edu

S. G. N. Pinkley
Department of Industrial Engineering, University of Arkansas, Fayetteville, AR, USA
e-mail: snurre@uark.edu

H. G. Kaper, F. S. Roberts (eds.), *Mathematics of Planet Earth*, Mathematics of Planet Earth 5, https://doi.org/10.1007/978-3-030-22044-0_12

to the components of an infrastructure system and severely disrupt their performance and their ability to provide services. For example, Hurricane Sandy, which struck the East Coast of the USA in late October 2012, caused damage estimated at $65 billion [39] and reduced the power load in Manhattan by 30% and on Long Island by 70%, see Fig. 12.1. As another example, Hurricane Matthew, which struck the southeastern USA in October 2016, caused peak outage levels of 10% in Florida, 7% in Georgia, 33% in South Carolina, 14% in North Carolina, and 7% in Virginia [38].

Given their importance to society, increasing the *resilience* of infrastructure systems is an important direction for public policy. The presidential policy directive on critical infrastructure resilience [40] defines infrastructure resilience as the infrastructure's "*ability to withstand and rapidly recover from unexpected, disruptive events*." Therefore, the effective restoration of disrupted services after an extreme event plays an important role in the resilience of an infrastructure system.

Figure 12.2 provides a conceptual curve of infrastructure performance before and after an extreme event. After the event, the performance degrades until the full impact of the event is reached. The system then begins to restore services until it returns to a performance level at or above the level prior to the event. The *time to recover* is the length of time between the first impact of the event and the time when performance is fully restored. The *restoration performance* is the amount of restored services, integrated over a finite planning horizon. The restoration performance is equivalent to the (weighted) average length of a disruption to a customer. In other words, the time to recover focuses on the longest time a customer is without service, while the restoration performance focuses on the average time a customer is without services. The quantitative methods discussed in this chapter can be applied to

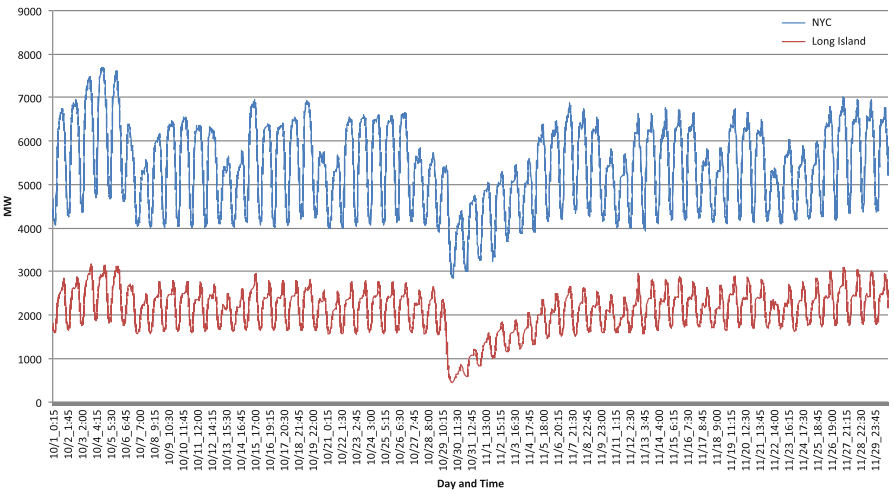


Fig. 12.1 The impact of Hurricane Sandy on the power load curves for the New York City and Long Island areas

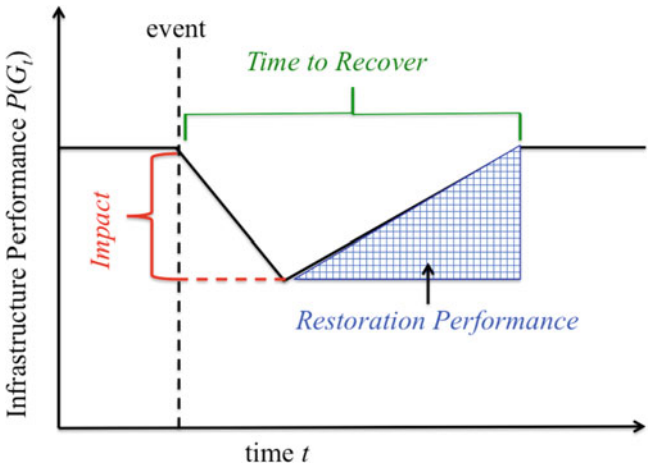


Fig. 12.2 Conceptual curve of infrastructure resilience with respect to a disruptive event. (A similar curve was proposed in [12])

analyze both the time to recover and the restoration performance of an infrastructure system.

Outline of the Chapter Section 12.2 introduces the basic framework for modeling infrastructure systems providing services as flows on networks. Infrastructure restoration requires an integrated approach to network optimization and scheduling. Section 12.3 provides an overview of the classical results of network optimization, discusses the limitations of modeling infrastructure systems using network flows, and surveys the advances that have been made to better capture the physical characteristics of infrastructure operations. Section 12.4 provides an overview of the classical results of network scheduling. Section 12.5 contains a survey of the literature on network restoration and details of the integration of network optimization and scheduling. Section 12.6 provides extensions of the work discussed in Sect. 12.5, including approaches to model interdependent infrastructure systems. The final Sect. 12.7 summarizes the findings and conclusions of the chapter.

12.2 Modeling the Performance of Infrastructures

To quantitatively measure the performance of an infrastructure system, we use the concept of "flow" across a network [3]. A network or graph, G, is composed of a set of nodes, N, and a set of arcs, A, connecting nodes, $G = (N, A)$. (Nodes and arcs are also called vertices and edges, respectively, and the terms can be used interchangeably.) An arc $(i, j) \in A$ provides a relationship between node i and node j. Throughout this chapter, we assume that the arcs are *directed*, meaning that

flow can only move along arc (i, j) from node i to node j. We use the notation G_t to indicate that the infrastructure varies with time.

From a modeling perspective, nodes represent components that can either generate services in an infrastructure system (*supply nodes*), alter the routes of these services (*transshipment nodes*), or consume services (*demand nodes*). Arcs then connect these different nodes and move the services between them. (This is the "flow" in the network.) For example, in the electric power system, power plants would be supply nodes, substations would be transshipment nodes, and anything ranging from households to factories to hospitals to malls would be demand nodes; power lines would be the arcs in the network and electricity the flow.

When an infrastructure system is impacted by a disruptive event, the goal is often to meet as much demand as possible, regardless of the delivery cost. This process can be modeled using the *maximum flow problem*, which will be discussed in Sect. 12.3.

After a disruptive event, damaged components of an infrastructure system must be repaired and services restored. The repair requires scarce resources such as work crews and equipment, each of which can only repair one component at a time. Therefore, an infrastructure manager must create a *schedule* that describes the times when a resource is repairing a component. The scheduling focuses on processing a set of *jobs* (which would correspond to damaged components) on a set of *machines* (corresponding to the work crews available to make these repairs), while minimizing a certain objective function [31]. A schedule provides completion times for each job (which we refer to as C_k). Traditional objectives in scheduling include minimizing the total weighted completion time of the jobs, $\sum_{k=1}^{K} w_k C_k$, where w_k characterizes the priority of job k, or minimizing the makespan, $C_{\max} = \max_{k=1,\ldots,K} C_k$. The total weighted completion time objective provides the (weighted) average time a job is completed; this time is similar to the restoration performance in Fig. 12.2. The makespan objective is similar to the time to recover in Fig. 12.2. Note that in infrastructure restoration, completing a certain job does not guarantee that services are restored to any customer. For example, if multiple power lines leading into a neighborhood are damaged, it is necessary to complete this set of jobs in order to restore the services.

12.3 The Maximum Flow Problem

The first objective after a disruption of an infrastructure system is often to meet as much demand for its services as possible, without concern for the cost of delivering these services. The *maximum flow problem* is a network optimization problem that seeks to determine the most one can get out of a particular network.

Definition 1 (Maximum Flow Problem) Consider a network $G = (N, A)$, where each arc $(i, j) \in A$ has a capacity u_{ij}. Let s be a source node and t sink node. The *maximum flow problem* seeks to determine the largest amount of flow that can

flow from s to t while respecting the capacities of the arcs and maintaining flow conservation at all nodes $i \in A\backslash\{s, t\}$.

Flow conservation is the concept that the amount of flow leaving node i is equal to the amount of flow coming into node i. More precisely, let $A(i)$ denote the adjacency list of node i, $A(i) = \{j : (i, j) \in A\}$, and let x_{ij} be the amount of flow on arc (i, j). Flow conservation implies that

$$\sum_{j \in A(i)} x_{ij} - \sum_{j:i \in A(j)} x_{ji} = 0. \tag{12.3.1}$$

If flow conservation holds for all $i \in A\backslash\{s, t\}$, then any flow that leaves source node s must arrive at the sink node t. It is possible to model the maximum flow problem as a linear program [3], but it is not necessary to present this model for the purposes of this chapter.

The source node s in the maximum flow problem can be viewed as the supply node and the sink node t as the demand node. One potential difficulty in modeling infrastructure performance as a maximum flow problem is the fact that infrastructure systems often have multiple supply nodes and multiple demand nodes. However, this difficulty can be overcome through a network expansion technique [29].

In particular, let $G = (N, A)$ represent an infrastructure system, where $S \subset N$ are its supply nodes and $D \subset N$ its demand nodes. Each supply node $i \in S$ has a supply level s_i, which is the maximum amount of supply that can be generated from it, and each demand node $i \in D$ has a demand level d_i. The network seeks to determine the maximum amount of demand that can be met by sending flow from the set of supply nodes to the set of demand nodes, while respecting the capacity levels of the supply nodes and the arcs in the network and making sure that a demand node does not receive more than its requested demand level. This problem can be modeled as a maximum flow problem in an expanded network $G' = (N', A')$, where $N' = N \cup \{s, t\}$ and $A' = A \cup \{(s, i) : i \in S\} \cup \{(i, t) : i \in D\}$. Taking $u_{si} = s_i$ for $i \in S$ ensures that the flow out of i does not exceed its supply; similarly, taking $u_{it} = d_i$ for $i \in D$ ensures that flow into i does not exceed its demand. The maximum flow from s to t in G' then provides the amount of demand that can be met in the infrastructure system.

In the expanded network G', the maximum flow cannot exceed $\sum_{i \in D} d_i$, since the capacity of the *s-t cut* that separates t from the rest of the network is $\sum_{i \in D} u_{it}$. An $s - t$ cut in a network $G = (N, A)$ is a partition of the nodes into two sets N_1 and N_2, where s belongs to N_1, t belongs to N_2, $N_1 \cup N_2 = N$, and $N_1 \cap N_2 = \emptyset$. The *capacity* of the cut is $\sum_{(i,j) \in A: i \in N_1, j \in N_2} u_{ij}$.

Definition 2 (Minimum Cut Capacity Problem) Consider a network $G = (N, A)$ where each arc $(i, j) \in A$ has a capacity u_{ij}. Let s be a source node and t sink node. The *minimum cut capacity problem* seeks to determine the *s-t* cut (N_1^*, N_2^*) that has the minimum capacity of all cuts.

The following theorem establishes the weak duality of the maximum flow problem and the minimum cut capacity problem.

Theorem 1 *The flow from s to t in any feasible flow in the network is less than or equal to the capacity of any s-t cut in the network.*

Proof Given the fact that $s \in N_1$ and $t \in N_2$ for any s-t cut, flow conservation implies that each unit of flow must make its way from a node in N_1 to a node in N_2. Each time this unit of flow passes along an arc (i, j) with $i \in N_1$ and $j \in N_2$, it takes up one unit of capacity in the cut. Since the flow is feasible, no arc can have an amount of flow exceeding its capacity. □

A strong duality relationship between these two problems is established by applying a very intuitive algorithm to find the maximum flow in the network. The algorithm is similar to the idea of finding a directed path (defined as a sequence of nodes s-i_1-i_2-...-i_k-t from s to t with $(s, i_1), (i_1, i_2), \ldots (i_k, t) \in A$, where all arcs along the path have unused capacity, and then pushing as much flow as possible along this path.

However, this idea does not allow for altering flow already present in the network. Therefore, it is necessary to introduce the concept of the *residual capacity* of an arc. The residual capacity of arc (i, j), based on a flow x in the network, is defined as the maximum amount of flow that can be sent from i to j either (a) by using the unused capacity $u_{ij} - x_{ij}$ of arc (i, j) or (b) by reversing flow x_{ji} on arc (j, i). Let $r_{ij} = u_{ij} - x_{ij} + x_{ji}$. The *residual network* based on flow x is $G(x) = (N, A(x))$, where $A(x) = \{(i, j) : r_{ij} > 0\}$. Thus, the arc set $A(x)$ is defined to be all arcs with positive residual capacity.

This construction leads to the *Augmenting Path Algorithm* (see below, Algorithm 1), also referred to as the *Ford–Fulkerson Algorithm* [21], to solve the maximum flow problem. The Augmenting Path Algorithm finds an s-t path in the residual network (implying that the path has a residual capacity) and pushes as much flow along this path as possible. This is equivalent to altering the flow along the arcs and reverse arcs on a similar path in the original network (the δ calculation in Algorithm 1). Note that δ units of flow are pushed from the source node to the sink node along this path, thus increasing the maximum flow by δ. The algorithm then updates the residual capacities of the arcs in the path and removes those arcs from the residual network whose capacity drops to zero. The algorithm finds not only the maximum flow but identifies the minimum capacity cut in the network as well.

Theorem 2 *The Augmenting Path Algorithm determines the maximum flow and the minimum cut in the network. Therefore, strong duality exists between the maximum flow problem and the minimum cut capacity problem.*

Proof Consider the residual network at the termination of the Augmenting Path Algorithm. Let $N_1 = \{i$: there exists a directed path from s to i in $G(x)\}$ and $N_2 = \{i$: there does not exist a directed path from s to i in $G(x)\}$. Since there is no directed path from s to t, (N_1, N_2) defines a valid s-t cut in the original network.

Algorithm 1 Augmenting path algorithm

```
Input: Network G = (N, A), source node s, sink node t, and capacities u_ij for (i, j) ∈ A
Set max_flow = 0.
for (i, j) ∈ A do
    Set x_ij = 0
    Set r_ij = u_ij
    Put (i, j) into A(x)
end for
while There exists a directed s − t path in G(x) do
    Find a directed s − t path, P, in G(x)
    Set δ = min_{(i,j)∈P} r_ij
    Set max_flow = max_flow +δ
    for (i, j) ∈ P do
        r_ij = r_ij − δ
        if r_ij = 0 then
            Remove (i, j) from A(x)
        end if
        r_ji = r_ji + δ
        Put (i, j) in A(x)
    end for
end while
Return max_flow
```

Consider an arc (i, j) with $i \in N_1$ and $j \in N_2$. It must be the case that $r_{ij} = 0$. The proof is by contradiction. If $r_{ij} > 0$, the directed path from s to i combined with the arc (i, j) would be a directed path from s to j. Since $r_{ij} = u_{ij} - x_{ij} + x_{ji}$ and the flow must respect the arc capacity, it must be the case that $u_{ij} - x_{ij} \geq 0$ and $x_{ji} \geq 0$, which implies that $u_{ij} - x_{ij} = 0$ and $x_{ji} = 0$.

Therefore, the current amount of flow that goes from N_1 to N_2 is equal to $\sum_{(i,j):i\in N_1, j\in N_2} u_{ij}$, which is the cut capacity of (N_1, N_2). Because of flow conservation, this flow must have been generated at s and must be absorbed by t. Therefore, the flow level of x is equal to the capacity of the cut (N_1, N_2). By Theorem 1, the flow level of x is the maximum flow and the capacity of (N_1, N_2) is the minimum cut. □

The idea behind the Augmenting Path Algorithm is important for prioritizing those infrastructure components needing repair at restoration. Algorithm 1 does not specify which s-t path to select at each step of the algorithm; in fact, the selection could impact the running time of the method [3]. If a breadth-first search is applied to find the s-t path, then one obtains the Edmonds–Karp algorithm [19], which has a polynomial running time of $\mathcal{O}(|N||A|^2)$.

It is important to note that the modeling of infrastructures with the maximum flow problem is an abstraction, since it is assumed that the only constraints governing flow are conservation at the nodes and capacities at the arcs. This type of model is applicable to supply chain networks, where physical goods are moving through factories, warehouses, distribution centers, and stores.

A major application of the models proposed in this section is supply chain restoration, for example, for restoring stores to distribute food, water, and medical supplies. However, the operation of different infrastructures is constrained and dictated by intrinsic physical laws, so one must be careful when applying the insights obtained by analyzing a system in the abstract manner discussed in this chapter.

For example, the physics of power flow are often captured through the *AC flow* model, which relates voltages (represented through complex numbers), angles at the nodes of the network, and flows on the arcs of the network. Since the AC flow model is nonlinear, a linear DC flow model is often used instead. An excellent discussion of the relationship between these two models is given in [9]. There have been significant advances in linear approximations [16], semi-definite programming relaxations [18, 27], and convex quadratic relaxations [17, 25] for the AC flow model. These advances provide much better approximations of the actual operations of the power grid and should be applied when the focus is on the details of the infrastructure system.

The power grid is not the only system where additional, often nonlinear, constraints should be incorporated into the models. For natural gas networks, the Weymouth equation captures the relationship between the pressure at the nodes and gas flows on the arcs. This equation is nonconvex; convex relaxations are presented in [10]. An example of an optimization problem in the context of the Weymouth equation can be found in [32].

For water distribution networks, the Darcy–Weisbach equations [11] or Hazen–Williams equations [41] capture the relationship between the flow of water and pressure throughout the network. The equations are, once again, nonconvex. Zhang et al. [42] provide effective algorithms to examine water distribution networks that specifically capture the Hazen–Williams equations.

In general, the inclusion of infrastructure-specific constraints helps to obtain better approximations, albeit at the cost of increased computation times. Nevertheless, insights can often be obtained about effective restoration policies through the use of simpler models of infrastructure systems.

12.4 Dynamic Allocation of Work Crews

The repair of an infrastructure system after an extreme event requires managers to allocate work crews over time. In scheduling notation [31], each damaged component of the system can be viewed as a job k (where K is the set of all jobs) with a certain weight w_k (emphasizing the importance of the job) and a processing time p_k (the duration of the repair operation). It is often the case that the processing time would also depend on the work crew; however, for ease of presentation, we will not discuss this case.

Each work crew can be thought of as a machine m (where M is the set of all machines). A feasible schedule *assigns* each of the jobs in K to a machine and then

orders the set of jobs assigned to machine m. The completion time of job k, denoted by C_k, is the time the machine needs to finish the job and is the sum of its processing time and the processing times of all jobs ordered before it on the machine to which it is assigned.

In the notation of [24], a scheduling problem is indicated by three entries, $\alpha|\beta|\gamma$, where α corresponds to the machine environment, β to special characteristics, and γ to the objective of the problem. For example, $1||\sum w_k C_k$ refers to the problem of scheduling jobs on a single machine in order to minimize the total weighted completion time. The problem $Pm|\text{prec}|C_{\max}$ would refer to the problem of scheduling jobs with precedence constraints on m parallel identical machines in order to minimize the makespan.

Recall that the restoration performance (Fig. 12.2) of an infrastructure can be viewed as the weighted average time that disrupted services are restored to customers, which is similar to the total weighted completion time objective. Therefore, we will examine the problem $1||\sum w_k C_k$ in order to provide insight into algorithms that can provide high-quality solutions to infrastructure restoration problems.

A greedy approach to this problem would be to schedule the jobs in non-increasing order of their weight-to-processing-time ratio, w_k/p_k. This ratio essentially provides the per-unit time increase in the objective by delaying the scheduling of job k. In other words, the ratio measures the impact on the objective of delaying the processing of the job. It therefore makes sense to schedule jobs with higher ratios earlier in the schedule. This greedy algorithm, often referred to as the *weighted shortest-processing time* (WSPT) rule, solves this scheduling problem [36].

Theorem 3 *The schedule resulting from applying the WSPT rule is optimal to the* $1||\sum w_k C_k$ *problem.*

Proof The proof is by contradiction.

Suppose the WSPT schedule is not optimal. Then there are two jobs, say k_1 and k_2, where $w_{k_1}/p_{k_1} > w_{k_2}/p_{k_2}$ and job k_2 is scheduled before k_1 in the optimal schedule $\mathscr{S}$.

Suppose that job k_2 is processed, followed by jobs $\ell_1, \ldots, \ell_\sigma$, and k_1. If $w_{k_1}/p_{k_1} > w_{\ell_\sigma}/p_{\ell_\sigma}$, then redefine $k_2 = \ell_\sigma$, so k_1 and k_2 are adjacent. Otherwise, redefine $k_1 = \ell_\sigma$. Then there is one job less between k_1 and k_2. Repeat the process until there are no jobs between k_1 and k_2.

The fact that k_1 and k_2 are now adjacent in $\mathscr{S}$ makes it easier to analyze the impact of *swapping* the positions of k_1 and k_2. Recall that $\mathscr{S}$ schedules k_2, followed immediately by k_1. Let $\mathscr{S}'$ be the schedule obtained from $\mathscr{S}$ by swapping the positions of k_1 and k_2, where k_1 is scheduled before k_2.

Let C_k is the completion time of job k under $\mathscr{S}$ and C'_k the completion time of job k under $\mathscr{S}'$, and let $s = C_{k_2} - p_{k_2}$ be the start time of k_2 in $\mathscr{S}$, which also ends up being the start time of k_1 in $\mathscr{S}'$. For all jobs $k \neq k_1, k_2$, we have $C_k = C'_k$. This is easily seen for jobs that have $C_k \leq s$, since we don't alter any of the schedule before S. For jobs that have $C_k > s$, note that the pair of jobs k_1 and k_2 are completed at the same time in both schedules. For $\mathscr{S}$, start k_2 at s, finish it at $s + p_{k_2}$, immediately start k_1, and then complete k_1 at $s + p_{k_2} + p_{k_1}$. For $\mathscr{S}'$,

the only change is in the order of the jobs; therefore, start k_1 at s and finish k_2 at $s + p_{k_1} + p_{k_2}$. Hence, a comparison of the total weighted completion times of $\mathscr{S}$ and $\mathscr{S}'$ can focus completely on the weighted completion times of k_1 and k_2.

Since $\mathscr{S}$ is optimal, it must be the case that

$$w_{k_1} C_{k_1} + w_{k_2} C_{k_2} \leq w_{k_1} C'_{k_1} + w_{k_2} C'_{k_2}. \tag{12.4.1}$$

Substitution of the actual completion times of the jobs in each schedule yields the inequality

$$w_{k_1}(s + p_{k_1} + p_{k_2}) + w_{k_2}(s + p_{k_2}) \leq w_{k_1}(s + p_{k_1}) + w_{k_2}(s + p_{k_1} + p_{k_2}). \tag{12.4.2}$$

Canceling out common terms, we obtain the inequality $w_{k_1} p_{k_2} \leq w_{k_2} p_{k_1}$ or, equivalently, $w_{k_1}/p_{k_1} \leq w_{k_2}/p_{k_2}$, in contradiction with the initial assumption $w_{k_1}/p_{k_1} > w_{k_2}/p_{k_2}$. Therefore, the optimal schedule must follow the WSPT rule. □

The proof of Theorem 3 follows a path similar to a number of optimality proofs in the area of scheduling. In particular, the swapping of jobs that violate the properties of a decision rule (sometimes referred to as a dispatching rule) is quite common, and then one looks at a comparison of the objectives between the schedule that violates the rule and the swapped schedule [31].

The WSPT rule can easily be adapted to situations with parallel identical machines. In this setting, there are m machines that can process the jobs and have identical properties. In particular, job k requires p_k amount of time to be processed on any of the machines. The WSPT rule would prioritize the jobs and then, whenever a machine completes a job, the first job on this priority list which has not been processed or is not being processed would be started on the available machine.

Most scheduling problems are quite difficult to solve to optimality and, therefore, relatively simple decision rules like the WSPT rule cannot always provide the optimal solution. In particular, many scheduling problems are NP-hard or, equivalently, their decision versions are NP-complete [22]. Consequently, there are currently no solution methods that are polynomial in terms of the input size of the problem. (For scheduling problems, the input size would be the number of jobs, the number of machines, $\log \sum p_k$, and $\log \sum w_k$.) For example, the problem of minimizing the makespan of the jobs on two parallel identical machines, $P2||C_{\max}$, is NP-hard. Dispatching rules can play an important role in approaching NP-hard scheduling problems, since they are intuitive, efficient, and often provide high-quality solutions. In fact, a dispatching rule inspired by the WSPT rule often results in high-quality solutions for infrastructure restoration problems.

12.5 Infrastructure Restoration: Network Optimization Meets Scheduling

Modeling infrastructure restoration involves aspects of both network optimization and scheduling, as indicated by the term *integrated network design and scheduling* (INDS) [29].

In the maximum flow INDS problem, we are given an *initial* network $G = (N, A)$ with a source node s and sink node t and arc capacities u_{ij}. The network will operate over a planning horizon with T time periods, which can be thought of as the length of the restoration process. There is a set of arcs A', which can be *installed* into the network by a set of parallel identical machines, denoted by M. This set can be viewed as the set of components damaged by the extreme event. Note that focusing on installing just arcs into the network is without loss of generality. In particular, a damaged node i can be split into two nodes, i_1 and i_2, with an arc (i_1, i_2), and all $(j, i) \in A$ being replaced with (j, i_1) and all $(i, j) \in A$ with (i_2, j). Each arc $(i, j) \in A'$ has a capacity u_{ij} and a processing time p_{ij}. The schedule of arcs in A' on the machines generates their completion times.

Let $G_t = (N, A_t)$ where $A_t = A \cup \{(i, j) \in A' : C_{ij} \leq t\}$, i.e., the network at time t is the original network plus all of the arcs that have been completed by t. The objective in this problem is to maximize $\sum_{t=1}^{T} P(G_t)$, where $P(G_t)$ is the maximum flow level in the network at time t. Figure 12.3 provides a visualization of this objective function and the improvement in the performance of the network as arcs are installed into it. Determining the schedule of a single machine that maximizes $\sum_{t=1}^{T} P(G_t)$ (with $P(G_t)$ representing the maximum flow at time t) is NP-hard [29]. For example, consider the network in Fig. 12.4. Here, $G = (\{1, 2, 3\}, \emptyset)$, $s = 1$, and $t = 3$. We consider a planning horizon of $T = 9$

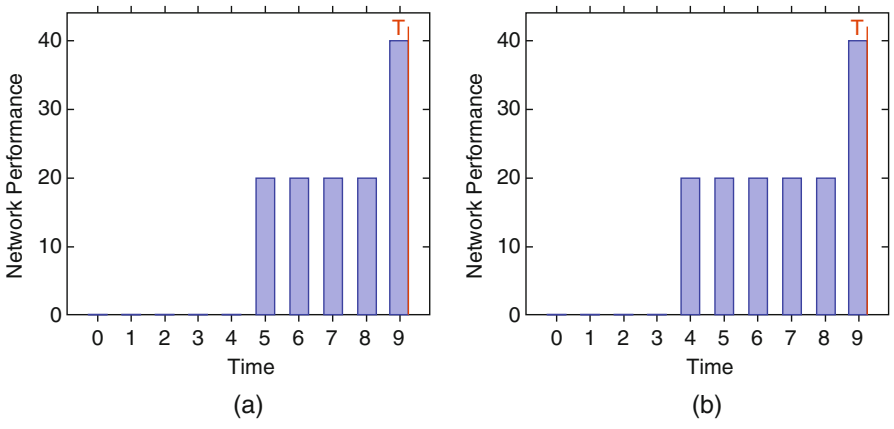


Fig. 12.3 Example of the cumulative objective function, where we to seek to maximize the sum of the performance of the network at each time until T. (**a**) Arc-based approach. (**b**) Path-based approach

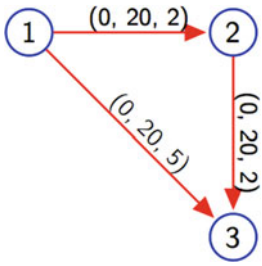


Fig. 12.4 Example INDS network. Each arc is labeled with (x_{ij}, u_{ij}, p_{ij})

time periods. We have $A' = \{(1, 2), (1, 3), (2, 3)\}$, where all arc capacities are 20, $p_{12} = p_{23} = 2$, and $p_{13} = 5$. Assume that a single machine is available to process the arcs.

Recall that the WSPT rule examines the impact on the objective function of delaying (or processing) a certain task. Given that the objective of the INDS problem is to maximize flow, one may initially suggest that the numerator of a ratio for an arc (job) could be the improvement in the maximum flow resulting from the installation of this arc. For our example, the individual installation of arc $(1, 2)$ or $(2, 3)$ would not increase the maximum flow. For arc $(1, 3)$, its installation would increase the flow by 20 and, therefore, it would have a ratio of $20/p_{13} = 20/5 = 4$. We would process arc $(1, 3)$ first, and then either $(1, 2)$ or $(2, 3)$ in any order. The arc-based approach in Fig. 12.3 shows the performance of the network over time for this solution. We see an increase in the maximum flow of 20 at time 5 (after completing $(1, 3)$) and an additional increase at time 9 (after completing arcs $(1, 2)$ and $(2, 3)$), for a total level of performance of 120 over the restoration horizon. However, by processing arcs $(1, 2)$ and $(2, 3)$ before arc $(1, 3)$, we can achieve our first increase of 20 at time 4 ($p_{12} + p_{23}$), with a total level of performance of 140, which is demonstrated in the path-based approach in Fig. 12.3.

The choice to process arcs $(1, 2)$ and $(1, 3)$ first in this example results from the fact that completing them together installs an *augmenting path* from 1 to 3 with capacity 20, which only takes 4 units of time to complete. This means that the adaptation of the WSPT rule to our maximum flow INDS problem should view "jobs" as augmenting paths that could be installed into the network, rather than on the individual contributions of each arc. Therefore, the proposed dispatching rule will integrate concepts from both network optimization (augmenting paths) and scheduling (the WSPT rule).

The *augmenting path dispatching rule* for the maximum flow INDS problem will greedily select the augmenting path P that has the largest ratio of residual capacity (which is the increase in flow resulting from installing the path) to the sum of the processing times of the *uninstalled* arcs on the path. For ease of this initial discussion, we assume that there is a single machine available to process jobs. Let G_t be the network at time t and suppose that the machine becomes available to process another set of jobs. We can first determine the maximum flow in the network G_t, which provides the current performance level. Furthermore, we can determine the residual network $G_t(x)$ associated with this flow by setting $r_{ij} = u_{ij} - x_{ij} + x_{ji}$

and only putting arcs into $A_t(x)$ that have $r_{ij} > 0$. Recall that $G_t(x)$ *cannot* have a path from s to t, since it would otherwise contradict that we have determined the maximum flow in the network.

We can *extend* this residual network to include all arcs that can still be installed into the network—that is, for any arc $(i, j) \in A' \backslash A_t$ we set its residual capacity $r_{ij} = u_{ij}$ and put it into the arc set $A'_t(x)$ of the extended residual network $G'_t(x)$. Any s-t path in the extended residual network represents a path that could be installed into G_t in order to increase the performance (maximum flow) in the network. Define $\mathscr{P}$ as the set of all directed s-t paths in $G'_t(x)$. Since $G_t(x)$ does not contain a directed s-t path, each path $P \in \mathscr{P}$ must contain at least one uninstalled arc with non-zero processing time in $A' \backslash A_t$. Our augmenting path dispatching rule will seek to determine

$$\max_{P \in \mathscr{P}} \frac{\min_{(i,j) \in P} r_{ij}}{\sum_{(i,j) \in P : (i,j) \in A' \backslash A_t} p_{ij}}. \tag{12.5.1}$$

The numerator in Eq. (12.5.1) is equal to the residual capacity of the path and is equivalent to the δ parameter of path P in Algorithm 1. The denominator in Eq. (12.5.1) is equal to the sum of the processing times of the *uninstalled* arcs in path P. Therefore, this problem determines the path P that has the largest flow per unit processing time ratio, which is similar to the weight per unit processing time ratio of the WSPT rule. The difficulty, though, with determining the optimal path P^* for (12.5.1) is that there could be an exponential number of paths based on the inputs to the INDS problem. For the traditional WSPT rule, we are able to calculate the w_k/p_k ratios in polynomial time. Therefore, more work needs to be done in analyzing the optimization problem (12.5.1) for it to be solved in polynomial time.

An approach to solve the optimization problem (12.5.1) that relies on an observation about solving the problem, given that we know the optimal numerator, is discussed by the authors in [30] and [29]. In particular, suppose that we knew that $\delta^* = \min_{(i,j) \in P^*} r_{ij}$ was the residual capacity of the path P^* that optimizes (12.5.1) without knowing P^*. We know that for any $(i, j) \in P^*$, it must be the case that $r_{ij} \geq \delta^*$, which is to say that the residual capacity of any arc on the path is greater than or equal to δ^*. Given that we know the numerator of Eq. (12.5.1), we would seek to make the denominator as small as possible in order to increase the ratio. Therefore, P^* would be the *shortest processing time* path from s to t containing arcs with residual capacities greater than or equal to δ^*. More precisely, we define the network $G'_t(x, r) = (N, A'_t(x, r))$ where $A'_t(x, r) = \{(i, j) \in A'_t(x) : r_{ij} \geq r\}$. Then P^* is the shortest processing time path from s to t in $G'_t(x, \delta^*)$. Since all processing times are nonnegative, Dijkstra's algorithm [3] can be used to determine this path.

However, we do not know δ^*, but we can use the above observation to find the best path for any fixed residual capacity. In particular, we can determine the shortest processing time path in $G'_t(x, r)$ from s to t for any r to determine how quickly we can install a path with a residual capacity greater than or equal to r. For any path P, its residual capacity must equal the residual capacity of some arc on the path, which

implies that there are only $\mathcal{O}(|A| + |A'|)$ possible values for δ^* to take in $G'_t(x)$. Therefore, we can solve the optimization problem (12.5.1) by repeatedly solving shortest path problems in $G'_t(x, r)$ for all relevant residual capacity levels (i.e., the value is equal to the residual capacity of some arc in $G'_t(x)$). For each of the residual capacity levels, it is straightforward to determine its ratio for the problem (12.5.1) and then select the path that has the highest ratio of the residual capacity level and the length of the path from s to t.

For the single machine environment, once we solve (12.5.1), we process the uninstalled arcs in P^* on the machine. For a parallel identical machine environment (i.e., it requires p_{ij} time to process arc (i, j) on any of the machines), we keep a queue of arcs needing to be processed and solve a slightly different version of (12.5.1). When a machine becomes available, if there is an arc in the queue, we process it on the machine. If there is no arc in the queue, we populate it by solving (12.5.1), where all arcs installed *or currently being processed by the machines* are considered "installed." In other words, we look at the best path to install after finishing the current installations (but we won't wait to begin processing this path). We then populate the queue with the optimal path by solving this modified version of (12.5.1).

The augmenting path dispatching rule performs quite well in practice. For example, Nurre and Sharkey [29] examined its computational performance on an infrastructure network resembling the power grid of lower Manhattan. This network has $|N| = 1603$ nodes and $|A| + |A'| = 2621$ arcs. Before applying the network expansion technique discussed in Sect. 12.3, it had 14 supply nodes and 134 demand nodes (which represented aggregate customers; for example, city blocks). Nurre and Sharkey [29] created instances of the INDS problem by damaging a certain percentage of the 2621 arcs. Table 12.1 shows the performance of the augmenting path dispatching rule and solving an integer programming formulation of the maximum flow INDS problem with CPLEX 12.0.

Table 12.1 Computational results comparing the augmenting path dispatching rule with solving an integer programming formulation using CPLEX 12.0 over five instances

		Dispatching rule		CPLEX 12.0	
Machines	Percentage	Time (s)	Gap (%)	Time (s)	Gap (%)
1	25	28.75	0.72	12335.11	0.34
	50	31.97	0.07	5178.08	0.03
	75	19.33	0.40	14400.00	0.40
2	25	32.13	2.61	14400.00	0.66
	50	68.10	0.15	7363.34	0.03
	75	43.22	0.39	12100.49	0.27
3	25	29.33	0.61	14400.00	0.51
	50	139.59	1.05	14400.00	0.46
	75	57.45	0.51	14400.00	0.35

Adapted from [29]

The dispatching rule is capable of solving problems in under 3 min, which would be sufficient to support real-time decision making in infrastructure restoration. The optimality gap of a solution (either determined by the dispatching rule or the best found solution for the integer program) for this class of problems is calculated from the formula

$$\frac{\text{Upper Bound on Objective} - \text{Objective of Solution}}{\text{Upper Bound on Objective}}. \tag{12.5.2}$$

Even with a 4-hour time limit, CPLEX 12.0 is not capable of improving significantly upon the near-optimal solution returned by the heuristic, which is demonstrated by the fact that the gap does not significantly change between the dispatching rule and the solution found by CPLEX 12.0. Therefore, the integration of ideas from network optimization and scheduling can help to provide near-optimal solutions to these problems in a way that standard optimization approaches cannot.

The objective function of maximizing the cumulative maximum flow ($\sum_{t=1}^{T} P(G_t)$) would focus on optimizing the restoration performance in Fig. 12.2. We could also define a maximum flow INDS *makespan* problem that would model the time to recover in Fig. 12.2. In particular, we wish to find the minimum amount of time to achieve a certain level of maximum flow, say F'. The objective of the maximum flow INDS makespan problem is then to minimize $\bar{T}$ such that $P(G_{\bar{T}}) \geq F'$. Figure 12.5 visualizes this objective function. This version of the problem is NP-complete [29], even for a single machine, and the augmenting path

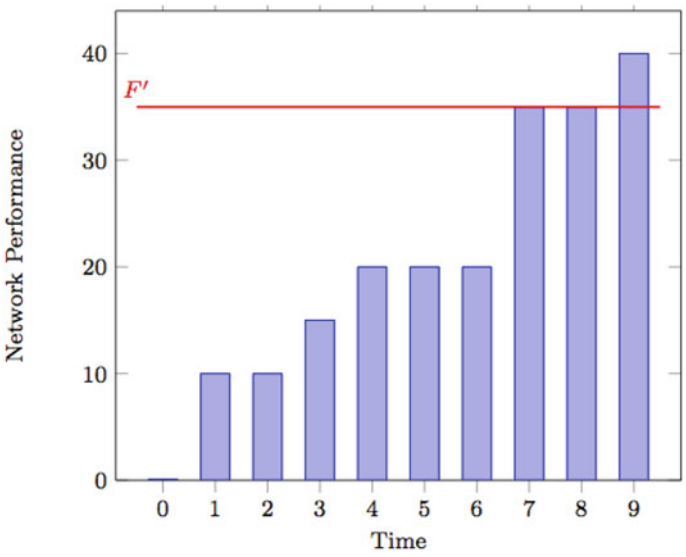


Fig. 12.5 Example of the makespan objective, where we seek to minimize the time to meet or exceed a threshold F'. In this example, $F' = 35$ and the makespan is $T = 9$

Table 12.2 Computational results comparing the augmenting path dispatching rule with solving an integer programming formulation of the maximum flow INDS makespan problem using CPLEX 12.0 over five instances

		Dispatching rule		CPLEX 12.0	
Machines	Percentage	Time (s)	Gap (%)	Time (s)	Gap (%)
1	25	21.03	4.88	1.61	0.00
	50	112.35	6.99	1347.62	0.00
	75	130.07	7.11	13759.31[a]	1.33[a]
2	25	18.93	6.70	6155.98	0.98
	50	83.55	10.81	14400.00	3.44
	75	136.73	11.05	14400.00	3.17
3	25	16.68	10.01	8677.14	2.34
	50	75.89	11.46	14400.00	3.50
	75	121.09	12.04	14400.00	3.15

Adapted from [29]

[a]Two of the five instances ran into memory errors. The solution and time right before running into the memory errors were captured and averaged into the values displayed

dispatching rule can be applied to it as well. Table 12.2 demonstrates the results of this application to problems focused on recovering 100% of the disrupted services.

The dispatching rule does not perform quite as well for the makespan problems, which can be explained by the fact that the objective functions are much smaller for this class of problems. Therefore, "errors" made by the dispatching rule in selecting which arcs to process have less chance to correct themselves than in the problem with the cumulative objective. However, given that the dispatching rule can help infrastructure managers examine both the restoration performance and the time to recover after a disruptive event, it is a powerful approach to providing solutions to infrastructure restoration problems.

In recent years, there has been a significant amount of literature studying network restoration problems. We highlight some of this literature here and refer the reader to [14] for a review.

Incremental network design problems [7, 20, 26] are special cases of INDS problems, where exactly one component can be installed per time period. Baxter et al. [7] and Kalinowski et al. [26] examine incremental network design problems, where the performance metrics are related to the shortest path in the network and the max flow in the network, respectively. They provide approximation guarantees for intuitive greedy algorithms for these problems. Engel et al. [20] show that the incremental network design problem, where the minimum spanning tree in the network is the performance of it at time t, can be solved with a greedy algorithm. Goemans and Unda [23] consider approximation guarantees of greedy algorithms for a general class of incremental optimization problems. Researchers have also considered problems associated with clearing debris and re-opening of roads in order to restore connectivity between supply and demand nodes within the transportation infrastructure [4–6, 8, 15]. The work of Averbakh [4] and Averbakh

and Pereira [5, 6] focuses on objectives related to the *recovery time* of a node, which is when a path is established from a supply node to it; Çelik et al. [15] consider a similar problem, but specifically model the fact that the restoration planner may have incomplete information about the resources required to repair arcs in the network.

12.6 Extensions

The maximum flow INDS problems discussed in Sect. 12.5 focus on modeling the recovery of an infrastructure system from a disruptive event that is focused on maximizing the amount of services it provides to its customers (or, equivalently, minimizing the amount of unmet demand in the system). Once services are restored to all customers, it may be that the infrastructure is then concerned with the *cost* of meeting the demand of the customers, which would be a different type of network optimization problem, the *minimum cost flow problem*.

12.6.1 Infrastructure Performance Metric: Minimizing Cost Flow

Definition 3 (Minimum Cost Flow Problem) Consider a network $G = (N, A)$ where each node has a supply level $b(i)$ and each arc $(i, j) \in A$ has a capacity u_{ij} and a cost c_{ij}. The *minimum cost flow problem* seeks to determine the flow (where x_{ij} is the flow on arc (i, j)) in the network that respects all capacities and maintains flow balance (i.e., the outflow of a node minus the inflow of the node is equal to $b(i)$) that has the smallest cost $\sum_{(i,j)\in A} c_{ij}x_{ij}$.

In the minimum cost flow problem, the concept of *flow conservation* at the nodes is replaced by the *flow balance*. If $b(i) > 0$, node i is considered a supply node, and if $b(i) < 0$, node i is considered a demand node. The identity

$$\sum_{j\in A(i)} x_{ij} - \sum_{j:i\in A(j)} x_{ji} = b(i) \tag{12.6.1}$$

implies that supply nodes "generate" flow (similar to the source node in the maximum flow problem) and demand nodes "absorb" flow (similar to the sink node in the maximum flow problem).

The residual network plays an important role in analyzing this problem. In particular, we examine *directed cycles* in the residual network. We define the cost of a directed cycle C in the network as $c(C) = \sum_{(i,j)\in C} c_{ij}$ (where if arc (i, j) has cost c_{ij}, then its backwards arc (j, i) has cost $-c_{ij}$). The cost of a cycle represents the change in the objective that would result in pushing one unit of flow along the cycle (which would maintain flow balance at all nodes). The following result is

a generalization of the augmenting path optimality conditions associated with the maximum flow problem; the proof is given in [3].

Theorem 4 *The feasible flow x in a minimum cost flow problem is optimal if and only if its residual network does not contain a negative cycle.*

Nurre and Sharkey [29] discuss how the framework for the augmenting path dispatching rule can be extended to minimum cost flow INDS problems. In particular, if $\mathscr{C}$ is the set of all cycles in the extended residual network, then we seek to find the negative cycle C which satisfies the maximization problem

$$\max_{C \in \mathscr{C}} \frac{\left| c(C) \min_{(i,j) \in C} r_{ij} \right|}{\sum_{(i,j) \in C : (i,j) \in A' \setminus A_t} p_{ij}}. \tag{12.6.2}$$

The numerator is the absolute value of the product of the decrease in the cost of pushing one unit of flow along cycle C and the minimum residual capacity of an arc in the cycle (i.e., how much flow could actually be pushed along the cycle). Therefore, the numerator provides the impact to the minimum cost flow of installing the necessary arcs to have cycle C available in the network. We then seek to determine the cycle that maximizes the decrease in the minimum cost flow per unit processing time. Nurre and Sharkey [29] show that the optimal solution of (12.6.2) can be determined by solving *minimum cost to time ratio* problems [3] in the extended residual network $G'_t(x^*, r)$ for each possible residual capacity. Therefore, the dispatching rule of Sect. 12.5 can be extended to other performance metrics.

12.6.2 Modeling Multiple Interdependent Infrastructure Systems

From a community perspective, it may be important for the *set* of infrastructure systems which provide services to the community to come back online rather than any one single infrastructure. This would then imply that the resilience and the recovery of the *community* would need to measure the performance across its set of infrastructure systems. However, modeling the performance of a set of infrastructure systems is more complicated than simply *individually* modeling each of the systems as a network. This is due to the *interdependencies* between the operations of the infrastructures. For example, a subway station needs power to operate within the subway system. Therefore, if services are not provided to the subway station node in the power network, it cannot operate in the subway network. Another example is that a hospital requires power and potable water to provide health services to its patients. If the hospital node in the power network or the hospital node in the water network does not receive a proper level of services, it will not be able to

properly function in its own network. Therefore, it is necessary to account for the interdependencies between the networks to understand their overall performance.

We can formalize the approach to model the performance of a set of interdependent infrastructure systems. Let $\mathscr{I}$ be the set of infrastructure systems, each of which is represented by a network $G_\ell = (N_\ell, A_\ell)$. We refer to the interdependent infrastructure network as $G = \bigcup_{\ell \in \mathscr{I}} G_\ell$. The performance of the set of infrastructures, $P(G)$, is often a weighted sum of the performances of the individual infrastructure networks, $P(G) = \sum_{\ell \in \mathscr{I}} w_\ell P_\ell(G_\ell)$. Alternative performance measures of the set of infrastructure networks could focus on the *functionality* of certain key nodes, such as hospitals, police stations, shelters, and emergency response headquarters, based on the set of services received across infrastructures or to examine how well the community has recovered based on the services provided by the infrastructures. These performance measures would better capture the fact that the role of infrastructures is to support the community. An important distinction in all these measures is that the performance of infrastructure network ℓ depends on services being provided to nodes in other infrastructures.

In particular, let I_{ℓ_1,ℓ_2} be the set of dependencies from infrastructure ℓ_1 to infrastructure ℓ_2. For ease of presentation, we assume that each entity in this set is a pair of nodes $(i_1, i_2) \in I_{\ell_1,\ell_2}$, where $i_1 \in N_{\ell_1}$ and $i_2 \in N_{\ell_2}$. Note that the general technique can be extended to situations where a dependency exists between a node in ℓ_1 and an arc in ℓ_2. For $i_2 \in N_{\ell_2}$ to be able to operate in network G_{ℓ_2}, an appropriate level of service must be met at node $i_1 \in N_{\ell_1}$, represented as the demand of i_1, d_{i_1}. In addition to flow variables, we define variable v_{i_1} to be the level of service provided to node $i_1 \in N_{\ell_1}$ and a *binary* variable y_{i_1,i_2} that indicates whether or not enough services were met at i_1 for i_2 to properly function. We then specifically model the interdependency of (i_1, i_2) with the following set of constraints:

$$\sum_{j \in A_{\ell_1}(i_1)} x_{i_1 j} - \sum_{j:i \in A_{\ell_1}(j)} x_{j i_1} = -v_{i_1} \tag{12.6.3}$$

$$d_{i_1} y_{i_1,i_2} \leq v_{i_1} \tag{12.6.4}$$

$$\sum_{j \in A_{\ell_2}(i_2)} x_{i_2 j} \leq M y_{i_1,i_2} \tag{12.6.5}$$

$$\sum_{j:i \in A\ell_2(j)} x_{j i_2} \leq M y_{i_1,i_2} \tag{12.6.6}$$

$$y_{i_1,i_2} \in \{0, 1\}. \tag{12.6.7}$$

Constraint (12.6.3) replaces the flow balance constraint of i_1 and says that the outflow minus the inflow at i_1 is equal to the negative of the amount of services provided to that node. Constraint (12.6.4) ensures that if services were met at i_1 (i.e., $y_{i_1,i_2} = 1$), then the services met at i_1 are greater than or equal to the demand of i_1. Constraints (12.6.5) and (12.6.6) ensure that node i_2 is only operational in ℓ_2 (i.e., flow moves in/out of it) if services were met at i_1. In other words, if services

were not met at i_1 ($v_{i_1} < d_{i_1}$), then constraint (12.6.4) forces $y_{i_1,i_2} = 0$ and thus the right-hand side of constraints (12.6.5) and (12.6.6) must be zero. This would imply that all flow variables into i_2 and out of i_2 must be zero.

The above discussion focuses on a specific class of interdependencies, so-called *input* interdependencies [28]. In general, Lee et al. [28] discuss how similar integer programming formulations can be created to model the different classes of interdependencies between the operations of infrastructure systems. This, in turn, creates their *interdependent layered network* model, which can capture the performance of a set of interdependent infrastructure systems. Cavdaroglu et al. [13] have examined INDS problems to model the restoration of interdependent infrastructure systems in the context where the only interdependencies between the systems are in their operations and there is a centralized decision maker (such as an emergency manager of a county) controlling the restoration efforts of all infrastructures.

After an extreme event there may be interdependencies between the restoration jobs associated with the different infrastructure systems. For example, after Hurricane Sandy, repairs needed to be done to subway stations and subway lines. After these were completed, it was necessary to run test trains to ensure the safety and quality of the repairs prior to opening the station and/or line. Therefore, there were two restoration jobs in the subway system for a particular station, namely the repair job and the test train job. However, if power was disrupted to the subway station, then it needed to be restored prior to running the test train job. In other words, the test train job could not begin until after power was restored to the subway track. In scheduling terms [31]), this means that there was a *precedence constraint* between the power restoration and the test train job. A precedence constraint (usually expressed as a directed arc between two jobs, $A \rightarrow B$) between job A and job B implies that job B cannot start until job A is complete. Another example of this type of relationship is when trees bring down power lines onto a road. First, a safety inspection job must be done by the power company to ensure the street is safe to enter, then the road can be cleared of debris (often by the Department of Public Works), and then the downed power lines can be repaired (by the power company). This means there are precedence constraints between the inspection job and the debris clearance job as well as between the clearance job and the repair job. We refer the reader to [35] for a description of the different classes of restoration interdependencies observed after Hurricane Sandy.

In these examples, the precedence constraints exist between jobs in *different* infrastructures. In many cases, there is no centralized decision maker coordinating the restoration efforts between the infrastructures. In other words, the multiple infrastructures will be forming their restoration efforts independently of one another, often with little communication among them. Figure 12.6 provides a comparison of a centralized decision-making environment with precedence constraints (left) and a decentralized decision-making environment with precedence constraints for interdependent infrastructure restoration (right). In the decentralized setting, we can view each infrastructure as a *player* in the restoration scheduling *game* and analyze the problem in a game-theoretic framework.

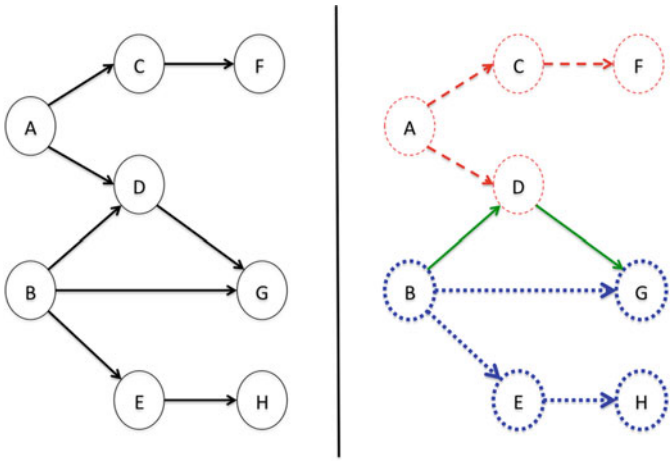


Fig. 12.6 A restoration scheduling environment with tasks $A - H$ and precedence constraints with a centralized decision maker (left) and with two infrastructures (red and blue) forming their restoration efforts (right)

An important concept in game theory is the concept of an *equilibrium* solution. An equilibrium solution is one in which no player has an incentive to change their decisions (actions) assuming that all the other players' decisions remain fixed. In other words, for our interdependent infrastructure restoration game, a solution is an equilibrium solution if no infrastructure can improve its restoration objective by altering its restoration schedule. The importance of an equilibrium solution in the restoration game is that, if the emergency manager of an area can "suggest" the restoration schedules to infrastructure managers, then no infrastructure manager would have an incentive to act differently than the suggested schedule. Therefore, it can be possible to measure the price of the decentralized decision-making process. In particular, the *price of anarchy* [33] measures the differences in the objective of the worst equilibrium solution and the optimal centralized solution. Abeliuk et al. [1, 2] have begun to examine issues around interdependent network restoration games. The analysis of the equilibrium solutions to interdependent network restoration games is an important area of future work.

Sharkey et al. [34] have examined the price of decentralized decision making for interdependent infrastructure restoration from an empirical perspective. They found that *information sharing* between infrastructures can reduce this price by between 20% and 60%, while still maintaining the autonomy of the infrastructures in their decision-making process. In particular, information sharing occurs when each infrastructure announces its tentative restoration schedule to all other infrastructures. Each infrastructure can then alter their tentative schedule based upon this information. In the example of Fig. 12.6, it may be that the weight of job G is much higher than the weights of job E and H ($w_G \gg w_E + w_H$). If the blue infrastructure did not know the red infrastructure's restoration schedule, it

may observe that, after completing B, G cannot be done until the red infrastructure completes D. Therefore, not knowing when the red infrastructure starts D, it could move on to E or hold back resources, expecting to complete G as soon as possible. However, if the blue infrastructure knows that the red infrastructure will complete D as soon as possible, the blue infrastructure could *determine* whether it makes sense to hold back resources for G or complete E first—that is, it could complete E before D is completed by the red infrastructure. Sharkey et al. [34] observed that there are situations where the overall restoration performance could increase after infrastructures update their plans after information sharing. Smith et al. [37] provide conditions under which the restoration plans resulting from information sharing will never converge—that is, there will always be an infrastructure that would prefer to adapt its restoration plans after receiving the latest round of information. An interesting area of future work is to formalize the potential gains obtained by information sharing between the infrastructures.

12.7 Summary and Conclusions

In this chapter, we focused on quantitative models for infrastructure restoration after a large-scale, disruptive event. We showed how network optimization techniques can be used to model the *operations* of an infrastructure system, while the *restoration efforts* of the system after the event can be modeled within a scheduling framework. We reviewed relevant network optimization and scheduling results, which form the basis for algorithms to provide near-optimal solutions to infrastructure restoration problems. The integration of both network optimization and scheduling was critical to create these algorithms. We then discussed how the models can be extended to interdependent infrastructure restoration and pointed out that game theory could be an important area to be applied to modeling community restoration efforts.

References

1. Abeliuk, A., Aziz, H., Berbeglia, G., et al.: Interdependent scheduling games. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI), New York, pp. 2–9. Elsevier, New York (2016)
2. Abeliuk, A., Berbeglia, G., Van Hentenryck, P.: One-way interdependent games. In: Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems, Paris, France, pp. 1519–1520 (2014)
3. Ahuja, R.K., Magnanti, T.L., Orlin, J.B.: Network Flows: Theory, Algorithms, and Applications. Prentice-Hall, Englewood Cliffs (1993)
4. Averbakh, I.: Emergency path restoration problems. Discret. Optim. **9**(1), 58–64 (2012)
5. Averbakh, I., Pereira, J.: The flowtime network construction problem. IIE Trans. **44**(8), 681–694 (2012)
6. Averbakh, I., Pereira, J.: Network construction problems with due dates. Eur. J. Oper. Res. **244**(3), 715–729 (2015)

7. Baxter, M., Elgindy, T., Ernst, T.A., et al.: Incremental network design with shortest paths. Eur. J. Oper. Res. **238**(3), 675–684 (2014)
8. Berktas, N., Kara, B.Y., Karasan, O.E.: Solution methodologies for debris removal in disaster response. EURO J. Comb. Optim. **4**(3-4), 403–445 (2016)
9. Bienstock, D., Mattia, S.: Using mixed-integer programming to solve power grid blackout problems. Discret. Optim. **4**(1), 115–141 (2007)
10. Borraz-Sanchez, C., Bent, R., Backhaus, S., et al.: Convex relaxations for gas expansion planning. INFORMS J. Comput. **28**(4), 645–656 (2016)
11. Brown, G.O.: The history of the Darcy-Weisbach equation for pipe flow resistance. In: Proceedings of the 150th Anniversary Conference of ASCE, Washington, D.C., USA, pp. 34–43 (2002)
12. Bruneau, M., Chang, S., Eguchi, R., et al.: A framework to quantitatively assess and enhance the seismic resilience of communities. Earthq. Spectra **19**(4), 733–752 (2003)
13. Cavdaroglu, B., Hammel, E., Mitchell, J.E., et al.: Integrating restoration and scheduling decisions for disrupted interdependent infrastructure systems. Ann. Oper. Res. **203**(1), 279–294 (2013)
14. Çelik, M.: Network restoration and recovery in humanitarian operations: framework, literature review, and research directions. Surv. Oper. Res. Manag. Sci. **21**(2), 47–61 (2015)
15. Çelik, M., Ergun, O., Keskinocak, P.: The post-disaster debris clearance problem under incomplete information. Oper. Res. **63**(1), 65–85 (2015)
16. Coffrin, C., Van Hentenryck, P.: A linear-programming approximation of AC power flows. INFORMS J. Comput. **26**(4), 718–734 (2014)
17. Coffrin, C., Hijazi, H., Van Hentenryck, P.: The QC relaxation: a theoretical and computational study on optimal power flow. IEEE Trans. Power Syst. **31**(4), 3008–3018 (2016)
18. Coffrin, C., Hijazi, H., Van Hentenryck, P.: Strengthening the SDP relaxation of AC power flows with convex envelopes, bound tightening, and valid inequalities. IEEE Trans. Power Syst. **32**(5), 3549–3558 (2017)
19. Edmonds, J., Karp, R.: Theoretical improvements in algorithmic efficiency for network flow problems. J. ACM **19**(2), 248–264 (1972)
20. Engel, K., Kalinowski, T., Savelsbergh, M.W.P.: Incremental network design with minimum spanning trees. J. Graph Algorithms Appl. **21**(4), 417–432 (2017)
21. Ford, L., Fulkerson, D.: Maximal flow through a network. Can. J. Math. **8**, 399–404 (1956)
22. Garey, M., Johnson, D.: Computers and Intractability. W.H. Freeman and Company, New York (1979)
23. Goemans, M.X., Unda, F.: Approximating incremental combinatorial optimization problems. In: Proceedings of APPROX/RANDOM 2017, pp. 1–6. Dagstuhl, Wadern (2017)
24. Graham, R.L., Lawler, E.L., Lenstra, J.K., et al.: Optimization and approximation in deterministic sequencing and scheduling: a survey. Ann. Discrete Math. **5**, 287–326 (1979)
25. Hijazi, H., Coffrin, C., Van Hentenryck, P.: Convex quadratic relaxations for mixed-integer nonlinear programs in power systems. Math. Program. Comput. **9**(3), 321–367 (2017)
26. Kalinowski, T., Matsypura, D., Savelsbergh, M.W.P.: Incremental network design with maximum flows. Eur. J. Oper. Res. **242**(1), 51–62 (2015)
27. Lavaei, J., Low, S.: Zero duality gap in optimal power flow problem. IEEE Trans. Power Syst. **27**(1), 92–107 (2012)
28. Lee, E.E., Mitchell, J.E., Wallace, W.A.: Restoration of services in interdependent infrastructure systems: a network flows approach. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **37**(6), 1303–1317 (2007)
29. Nurre, S.G., Sharkey, T.C.: Integrated network design and scheduling problems with parallel identical machines: complexity results and dispatching rules. Networks **63**, 306–326 (2014)
30. Nurre, S.G., Cavdaroglu, B., Mitchell, J.E., et al.: Restoring infrastructure systems: an integrated network design and scheduling problem. Eur. J. Oper. Res. **223**(3), 794–806 (2012)
31. Pinedo, M.L.: Scheduling: Theory, Algorithms, and Systems, 4th edn. Springer, New York (2012)

32. Romo, F., Tomasgard, A., Hellemo, L., et al.: Optimizing the Norwegian natural gas production and transport. Interfaces **39**(1), 46–56 (2009)
33. Roughgarden, T.: Selfish Routing and the Price of Anarchy. MIT Press, Boston (2005)
34. Sharkey, T.C., Cavdaroglu, B., Nguyen, H., et al.: Interdependent network restoration: on the value of information-sharing. Eur. J. Oper. Res. **244**(1), 309–321 (2015)
35. Sharkey, T.C., Nurre, S.G., Nguyen, H., et al.: Identification and classification of restoration interdependencies in the wake of Hurricane Sandy. J. Infrastruct. Syst. **22**(1), 04015, 007 (2016)
36. Smith, W.: Various optimizers for single-stage production. Naval Res. Log. Q. **3**(1-2), 59–66 (1956)
37. Smith, A.M., Gonzalez, A.D., Dueñas-Osorio, L., et al.: Interdependent network recovery games. Risk Anal. (2017). https://doi.org/10.1111/risa.12923
38. U.S. Department of Energy: Hurricane Matthew situation reports. Tech. rep., US DOE Office of Electricity Delivery and Energy Reliability, Washington, DC (2016)
39. U.S. Department of Housing and Urban Development: Hurricane Sandy rebuilding strategy. Tech. rep., US HUD Hurricane Sandy Rebuilding Task Force, Washington DC (2013)
40. White House, Office of the Press Secretary: Presidential Policy Directive: critical infrastructure security and resilience. Tech. rep., The White House, Office of the Press Secretary (2013)
41. Williams, G.S., Hazen, A.: Hydraulic Tables. Wiley, New York (1914)
42. Zhang, W., Chung, G., Pierre-Louis, P., et al.: Reclaimed water distribution network design under temporal and spatial growth and demand uncertainties. Environ. Model. Softw. **49**, 103–117 (2013)

Chapter 13
The Internet of Things and Machine Learning, Solutions for Urban Infrastructure Management

Ernesto Arandia, Bradley J. Eck, Sean A. McKenna, Laura Wynter, and Sebastien Blandin

Abstract Urban infrastructure management requires the ability to reason about a large-scale complex system: What is the state of the system? How can it be compactly represented and quantified? How is the system likely to evolve? Reasoning calls for predictive modeling, feedback, optimization, and control. With an understanding of the system state and its likely evolution, how should resources be allocated or policies changed to produce a better outcome? By leveraging data from the *Internet of Things*, it becomes feasible to perform online estimation, optimization, and control of such systems to help our cities function better. This involves taking traditional applications of mathematical sciences into a large-scale, online, and adaptive setting. We focus in this chapter on two particular applications that are important to effectively manage a city: transportation and municipal water services.

Keywords Algorithms · Control · Data · Estimation · Internet of Things · Machine learning · Prediction · Smart city · Urban computing

13.1 Introduction

The *Internet of Things* (IoT) allows for a degree of connectivity that has not been witnessed before. Everyday consumer objects as well as industrial components are being equipped with sensors and communication devices that are becoming ever smaller and more powerful. From home automation to supply chains to city

E. Arandia · B. J. Eck · S. A. McKenna
IBM Research, Dublin, Ireland
e-mail: ernestoa@ie.ibm.com; bradley.eck@ie.ibm.com; seanmcke@ie.ibm.com

L. Wynter (✉) · S. Blandin
IBM Research, Singapore, Singapore
e-mail: lwynter@sg.ibm.com; sblandin@sg.ibm.com

H. G. Kaper, F. S. Roberts (eds.), *Mathematics of Planet Earth*, Mathematics of Planet Earth 5, https://doi.org/10.1007/978-3-030-22044-0_13

infrastructure, IoT has the power to transform our lives. Harnessing this power to improve the management of urban infrastructure is the focus of this chapter.

While many aspects of a city are touched by IoT and can be improved by embracing this new technology, we concentrate on two key areas of city operations—transportation networks and municipal water supply services—and in particular on how ubiquitous sensing powered by IoT enables improved operations via the use of machine learning. Transportation and water services both rely on efficient real-time sensing and control of a city-wide network; hence, they are archetypes of challenging large-scale monitoring and operations applications.

Urban infrastructure management requires the ability to reason about a large-scale complex system: What is the state of the system, and how is it likely to evolve? With an understanding of the system state and its likely evolution, how should resources be allocated or policies changed to produce a better outcome? While prediction, reasoning, optimization, and control have been studied for decades, the Internet of Things and machine learning require embracing a new paradigm, as these functions must now be performed online and leverage vast amounts of data. They must also, importantly, quantify uncertainty in observations and models and adapt to the way in which the system changes over time via learning processes. When managing an urban infrastructure, the analytics must be designed to continuously (e.g., online) understand, reason, and learn about the system and use those capacities to better monitor, control, and manage it. Thus, feedback and adaptive learning are important components of the system. Figure 13.1 shows a conceptual illustration of the main functions of such a system.

The particularity of this domain is that it takes the traditional applications of the mathematical sciences into a large-scale, online, and adaptive learning setting. Much work has been done in recent years; for example, variants of subgradient and stochastic optimization methods that permit online learning via the incorporation of new data in successive iterations are proposed in [12, 29, 130]; a faster version of model predictive control (MPC)—a common optimal control technique, where a finite but rolling time horizon is defined and the next time slot is optimized while taking into account its effect on subsequent time slots—that can be used online is described in [122]; and an online method for the optimal control of continuous-time linear systems can be found in [59].

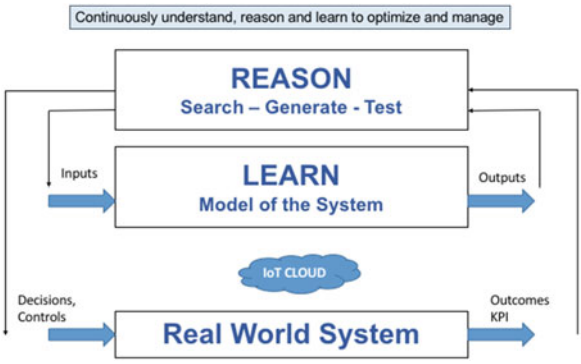


Fig. 13.1 Key functions of a system leveraging the Internet of Things for urban infrastructure management

Real-time data are key to the effectiveness of these online systems. The Internet of Things enables the measurement and systematic improvement of the state of an urban network in real time, thus opening the field of *urban computing* [132]. An example of a visually interesting collection of real-time urban data and metrics can be found on a city of London website, http://citydashboard.org/london/, which collates data on weather, train service quality, bike-share programs, air pollution, the river water level, and even the general mood of the city [66].

In this chapter, we review advances in the monitoring and control of urban infrastructure, focusing in particular on transportation networks and water supply services. In the transportation domain, online sensing means that problems traditionally solved offline with low-frequency batch data can now be reformulated and studied as online problems. Examples are rail transport scheduling [17]; vehicle routing [36]; real-time road traffic prediction—a key input in most online navigation and traffic control systems [83]; real-time arrival time prediction for an urban bus network [134]; and real-time train service-level estimation [109].

In municipal water networks, the Internet of Things enables online sensing of flows, pressures, tank levels, valve settings, and water quality parameters, which are now routinely monitored in large municipal water systems. Mathematical, statistical, and machine learning algorithms applied to these new data feeds enable capabilities such as real-time state estimation for leak localization [39], early warning of contamination events [67, 81], and temperature-adjusted residence times for estimation of water aging [32]. Smart meters at network connection points recording water use with relatively high frequency and publishing that information through Internet connections can be considered an archetype for IoT devices. They are increasingly being used by water utilities to better manage drinking-water networks [96] and identify customer usage patterns [80], among other applications.

Outline of the Chapter The remainder of the chapter is organized as follows. Section 13.2 describes the general framework. In Sect. 13.3, we discuss key scientific methodologies underlying transportation applications, both model-based and data-driven; specifically, state-space modeling (Sect. 13.3.1), estimation (Sect. 13.3.2), and control (Sect. 13.3.3). In Sect. 13.4, we give an overview of the main challenges in effectively managing municipal water supplies for cities. We discuss two examples: a data-driven estimation problem, namely municipal water-demand prediction (Sect. 13.4.1); and a control problem, namely pump scheduling for distribution networks (Sect. 13.4.2). We conclude in Sect. 13.5 with a perspective on future research at the frontier of the Internet of Things and machine learning.

13.2 General Framework

To set the stage, we briefly describe the state-space formulation, which has served as the dominant framework for the modeling of monitoring and control problems.

Consider a system with true state at time t denoted by Ψ_t, and let y_t be the vector of all available observations from the system up to time t. The *estimation* problem

is concerned with the computation of an *optimal estimate* of Ψ for a predefined loss function on the state-space model. The solvability of the estimation problem, which is a necessary condition for the proper understanding of the system of interest and subsequent applications, depends on the properties of the available data. While low-latency and high-volume IoT data opens an era of *data-intensive scientific discoveries* [54], it is not without challenges. IoT data typically consist of noisy measurements whose properties vary over time, necessitating novel formulations of the estimation problem.

Consider the following general discrete-time state-space model:

$$x_{t+1} = f(x_t, u_t) + w_t, \tag{13.2.1}$$

where x denotes the state variables and u the control variables. Time dependence is indicated by the subscript t. The function f characterizes the possibly nonlinear state model, the random variable $w \sim \mathcal{N}(0, W)$ is a white noise term which fully characterizes modeling errors. In this setting, the true state Ψ is assumed to follow the dynamics f without additional noise [61]. Measurements are modeled by the observation equation,

$$y_t = g(\Psi_t) + v_t, \tag{13.2.2}$$

where the function g characterizes observations y of the true state Ψ, and $v \sim \mathcal{N}(0, V)$ is a white noise term, which accounts for measurement errors assumed uncorrelated with modeling errors. The model parameters f, g, V, and W are estimated from data, using either a *model-based* approach or a *data-driven* approach. Historically, the estimation of these parameters has been data-constrained, and model-based approaches guided by *Occam's principle* have prevailed. In the types of analytics we explore here, the functional relations f and g need not be derived from physical principles or calibrated from small datasets. These functions can be fully data-driven and highly complex as, for example, when they are derived using neural networks [72].

We propose to analyze the key advances brought by the Internet of Things and machine learning in this framework through a review of various recent works, illustrating how traditional modeling, estimation, and control of urban networks have evolved. This exposition is instantiated on transportation and water supply settings, but in a framework general enough to allow the transfer of ideas of interest to other aspects of urban infrastructure, from energy to telecommunications to social networks.

13.3 Urban Transportation

Transportation networks are fundamental to the effective functioning of every city in the world. Goods must be transported to serve producers and consumers, people must be able to travel efficiently to accomplish professional, personal, and social

objectives. The need for efficient transportation applies to every person in the city and every business, and the development of increasingly efficient transportation methods to mitigate spreading congestion phenomena [104] is by nature a large-scale network problem. While public transport service is often centralized, with one or a small number of operators, the use of public transport services by the public brings it to the realm of decentralized systems, in that choices of commuters are made independently by many individuals, to the extent that mobility on demand is now competitive with traditional transport services. Road networks are another example of a highly decentralized system, in spite of the fact that traffic-signal control is, in some cases, centralized. Hence, transportation is by nature both critically important to manage and scientifically challenging to analyze, optimize, and control.

Transportation science has existed for well over 50 years and can be traced back to the pioneering work of Beckman et al. [11] in 1956, among others. These authors showed that the way transportation network users choose their route could be represented as an equilibrium problem, with an equivalent convex optimization formulation. The work that emerged from these early works gave rise to many more complex behavioral models, stochastic and dynamic extensions, applications to networks with both private transport and commercial vehicles, and numerous other variations. The scientific work encompassed and continues to advance the state of the art in convex and non-convex optimization, variational inequalities, and game theory; see the classic reference [90] for more details. The resulting models and algorithms form the basis of the majority of software in the market today to help transportation planners and engineers determine the best way to upgrade transport infrastructure.

Until recently, models relied on relatively little data to calibrate their parameters. Mobility surveys carried out by transport agencies have been the main source of information for modeling commuter or public transport passenger demand. The state of the network itself has historically been known only imprecisely and to a limited extent. For this reason, classical approaches to modeling transportation networks deduced the utilization of the system through physical models of how people travel in general. These models make assumptions on user behavior to estimate how synthetic users who follow various forms of rational behavior would arrange themselves on the network. This type of approach has proven useful for modeling scenarios such as medium-term network optimization.

The Internet of Things, however, turns this scenario upside down and opens up significant new opportunities and challenges. Whereas in the past, movements across a transportation network could only be deduced from first principles, today, in many cases, we can obtain nearly pervasive data on the system utilization [23]. This paradigm shift empowers transport providers with the ability to understand much more precisely the behavior and thus the needs of their customers, as well as the quality of the service they are offering with respect to those needs.

Developing machine learning solutions for IoT data also means that problems that in the past were addressed for a typical commuter at the network-wide level can now be solved for small groups of individuals, and that network management can be

tuned to specific user behavior and needs, culminating in mobility on demand. More formally, while historically the estimation and control problem has been expressed for the network state x as in (13.2.1) through a macroscopic framework, it is now conceivable to have finer and finer estimation problems, where x and f characterize the state of an agent (e.g., a car, a pedestrian) and its dynamics, allowing applications such as adaptive traffic control [121, 124], self-driving vehicles [87], or on-demand ride-sharing [2].

The remainder of this section is organized as follows. In Sect. 13.3.1, we present background information on state-space modeling for urban transportation. Section 13.3.2 describes classical approaches and recent progress related to estimation methods. Given these results, Sect. 13.3.3 discusses a few topics related to control of urban transportation networks in ways enabled by IoT data and machine learning.

13.3.1 State-Space Modeling

State-space models characterize mathematical relations between the state of different components of the system. The simplest state-space models are linear and time-invariant (13.2.1), but more complex nonlinear, time-varying, stochastic models have reached sufficient maturity for applications. State-space models are typically categorized according to the method underlying their design; *model-based* approaches propose to derive mathematical equations from first principles, and these models usually exhibit interpretable properties, whereas *data-driven* approaches are primarily guided by some metric evaluated on data, and usually exhibit good accuracy properties in the presence of data.

Model-Based Approaches Urban mobility has historically been modeled at three different scales. At a *microscopic scale*, vehicles are considered to behave independently by reacting to stimuli from neighboring vehicles according to a dynamical model. Traffic dynamics, for example, can be modeled as a set of coupled ordinary differential equations. At a *mesoscopic scale*, vehicles are considered as a large set of atomic elements with individual behavior following macroscopic laws or relations. Traffic dynamics can be modeled using gas-kinetic models [95] or cellular automata [20].

At a *macroscopic* scale, vehicles are considered to behave as a continuum medium. Traffic dynamics are modeled as a distributed system, using *partial differential equations* (PDEs) inspired from hydrodynamics theory [42, 73, 98]. Consequently, in this framework, the effect of network-wide route choices is not conveniently accounted for. One of the strengths of macroscopic models resides in the level of complexity they capture at a relatively low analytical and computational cost, and with limited data requirements for calibration. This has motivated the use of macroscopic models in particular for real-time online estimation and corridor management; see [82] for *Metanet* and [10] for *Mobile Millennium*. Furthermore, the mathematical theory of hydrodynamics brings a solid mathematical structure to macroscopic models.

As macroscopic models have become more exhaustive in the level of details they can include, microscopic models have benefited from increasing data availability and have become widely used. This is particularly true of pedestrian modeling [25], which is critical for public transport applications. Indeed, IoT data allows for identifying commuter movements with far greater granularity, such as identifying the movements of residents versus tourists. The authors of [127] use Singapore public transport system farecard data, distinguishing rechargeable-card transactions from single-ride transactions, to identify tourist movements across the city. Another source of IoT data, cellphone transaction records, also allows for identifying tourist trips from those of residents, via the telecommunication code used by mobile devices in roaming mode versus those attached to a local operator [93].

One of the biggest changes in modeling and managing public transportation networks enabled through the prevalence of IoT data is in real-time analytics. Whereas estimating origin-to-destination flows used to rely on surveys and, consequently, were imprecise and could not cover more than a small fraction of the population, it is now possible to accurately describe commuters' origin-to-destination movements using IoT data with a fraction of the effort and far greater coverage [118]. Telecommunications data has been used for this purpose with some success [92, 93]. Importantly, these estimates can be generated and updated in near real time, allowing for a far greater degree of responsiveness in adapting public transport-related services to commuter demand.

Interestingly, the IoT era has enabled the modeling and estimation of complex phenomena related to movement patterns, including the generation and management of traffic externalities, such as pollution emissions [101], energy consumption [115], and logistics and fleet delivery [9]. Additionally, the coupling of multiple complex networks such as the road network and the smart grid has received much attention in recent years [123, 129] under the more general umbrella of cyber-physical systems.

Data-Driven Approaches Data-driven traffic models dispense with much of the underlying physical relations described above in favor of learning the relations between variables of interest directly from the IoT data itself. In [83], the authors propose a space-time autoregressive integrated moving average, or STARIMA, model to represent the evolution of the traffic state on a road network,

$$X_t - \sum_{i=1}^{p} \phi_i \Phi B^i X_t = a_t + \sum_{j=1}^{q} \theta_j \Phi B_j a_t,$$

where X_t is the state vector of traffic flow across the links of the road network at time t; a_t is the error term, Φ a spatial correlation matrix, and B the backshift operator so that $B^d X_t = X_{t-d}$; p and q are the order of the autoregressive and moving average terms, respectively. While the model is data-driven, physical characteristics are taken into account via the choice of spatiotemporal locations used in the spatial correlation matrix. The parameters of the model, ϕ, θ, the variance

σ^2, and the spatial correlation matrix Φ are estimated from IoT data. A fully data-driven extension of the model of [83] was presented in [63], which eliminated the explicit description of interacting network links used in Φ through a lasso-based procedure [116] for automatic variable selection.

Given the lack of regularity of IoT data sources, both in spatial and temporal coverage, numerous researchers have investigated alternative methods for modeling urban traffic, such as support-vector regression (SVR) [58] and deep learning [77]. A comparison of time-series and supervised learning techniques such as SVR and neural network-based approaches was provided in [74].

IoT data are useful in improving the models and methods used for infrastructure planning, enabling, for example, the derivation of a real-time actual timetable of a public transport service. In [56], the authors develop a method to use telecommunications records to derive the actual timetable of a regional train service, by identifying transitions of cellphone users in geo-localized regions intersecting the regional train stations, and detecting bursts over time of the aggregation of those transitions. The method is shown to work reasonably well in spite of the very low spatial and temporal frequency of the cellphone records.

In a similar vein, [109] developed a technique for determining the true timetable of an urban metro system, this time using WiFi data available by passively monitoring the WiFi "probe requests" produced by mobile devices. This method has the advantage of using data that are universally accessible to the general public, neither requiring special software on the mobile devices nor access to proprietary hardware such as the WiFi access points. The authors propose a technique based on spectral clustering which serves to identify not only the actual timetable of the metro service, but can also be used as a low-latency approach for detecting incidents and delays. Indeed, perturbations, incidents, and other events can be detected in near real-time and used as input in models to determine the best response both for public transport passengers and for the operators themselves. Figure 13.2 illustrates the results of the spectral clustering technique on the passive WiFi data to estimate a metro timetable in normal conditions (on the left) and in the presence of an interruption in service (on the right). Pedestrian movements were also estimated

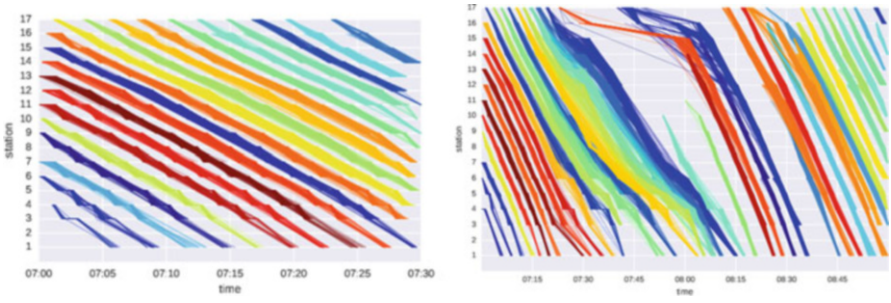


Fig. 13.2 Estimated metro timetable under normal conditions (left: Normal conditions) and a service disruption (right: Disrupted conditions)

using machine learning techniques on WiFi data by the authors of [71]. In [40], the authors estimated travel times between pairs of locations on a motorway using WiFi signals.

An interesting first step towards the comprehensive use of IoT data for modeling public transport was taken in [41]. The authors integrate numerous sources of data from the UK in order to create a multi-modal, weighted, directed, temporal, and multilayer network of the national public transport system. Their publicly available network couples the different transport modes including airports, ferry docks, rail, metro, coach, and bus stations.

13.3.2 Estimation Methods

A significant amount of research has focused on integrating information from data with prior knowledge. For real-time applications, it is natural to consider a variant of the estimation problem where data are made available sequentially. In this context, *sequential estimation algorithms* often rely on *Bayes' rule* and a computationally explicit optimality criterion (e.g., *Gauss–Markov* theorem for *minimum mean-squared error* (MMSE) estimation). In the case of additive noise, one of the best known sequential estimation algorithms is the seminal *Kalman filter* (KF) [62].

State Estimation The Kalman filter sequentially computes the best linear unbiased estimate (BLUE) at time $t+1$ from the BLUE estimate at time t, as follows:

$$\text{Forecast:} \begin{cases} x_{t+1|t} = A_{t+1}\, x_{t|t}, \\ \Sigma_{t+1|t} = A_{t+1}\, \Sigma_{t|t}\, A_{t+1}^T + W_{t+1}, \end{cases} \tag{13.3.1}$$

$$\text{Analysis:} \begin{cases} x_{t+1|t+1} = x_{t+1|t} + K_{t+1} \left(y_{t+1} - C_{t+1}\, x_{t+1|t} \right), \\ \Sigma_{t+1|t+1} = \Sigma_{t+1|t} - K_{t+1}\, C_{t+1}\, \Sigma_{t+1|t}, \\ \text{where} \quad K_{t+1} = \Sigma_{t+1|t}\, C_{t+1}^T \left(C_{t+1}\, \Sigma_{t+1|t}\, C_{t+1}^T + V_{t+1} \right)^{-1}. \end{cases} \tag{13.3.2}$$

The forecast step (13.3.1) consists in propagating the mean and covariance of the state through the linear model (13.2.1). The analysis step (13.3.2) amounts to the computation of the conditional mean of the state given the observations, for the linear observation model (13.2.2) and jointly Gaussian statistics. The conditional covariance is computed similarly. From a Bayesian perspective, the Kalman filter sequentially computes the posterior distribution of the state, based on the prior distribution given by the state-space model.

The estimation problem can alternatively be formulated as a signal recovery problem. In the framework of *compressed sensing*, an unknown sparse matrix $M \in \Re^{m \times t}$ is to be recovered from a random subset of its entries. The matrix in this case may represent the traffic level, in terms of speed, on the m links of the

road network at each time step over some n time periods. Clearly, some of the m links would possess IoT sensor data in some time periods, but the matrix is likely to have a large number of gaps. Mitrovic et al. [84] propose, for example, using CX decomposition in an online manner at each time step for this purpose. Related methods are proposed in [97, 133]; see also [102] for an application to OD-matrix estimation. As the sources and variety of real-time IoT data on and related to the traffic state increase, we can expect considerable improvements in the area of traffic state estimation.

The use of multiple approaches to prediction depending on the characteristics of the data raises the question of how to perform information fusion. The ensemble Kalman filter from [34] was recently applied to the problem of highway traffic state estimation [52], using fixed infrastructure and GPS-enabled mobile devices in [125, 126]. It was shown that accurate estimates can be provided even with low sampling rates (e.g., 1%). The dependency between data volume and accuracy was further quantified numerically in [89].

Another method for fusion of fixed-location sensor data and mobile IoT traffic data was proposed in [105] and is illustrated in Fig. 13.3. As shown in Fig. 13.3, during calibration periods, actual link speed observations on critical links are collected. Together with the GPS data received during the same period, these data points are stored as prediction candidates. The GPS records received in real time can then be used to determine which prediction candidate is most appropriate.

Revealing the Tail While the physics of traffic and traffic phenomena such as *phantom jams* [38] are well understood, and specific modes of traffic can be monitored [112], precise understanding and quantification of the impact of rare events such as traffic incidents is much more limited. Paradoxically, rare and unpredictable traffic incidents, having higher adverse impact potential for traffic conditions, are by nature the main focus of traffic operators. IoT data allows for

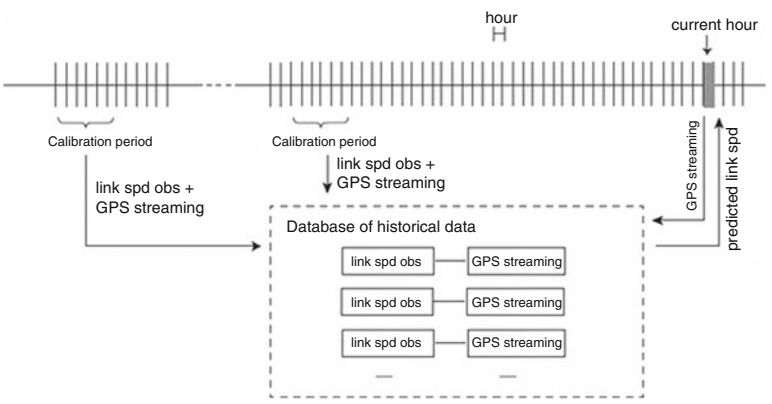


Fig. 13.3 An example fusion framework for urban road traffic data from IoT devices

the development of parsimonious models and robust data-driven understanding for incident-type conditions as well.

Consider a general model of the evolution of the local impact of a road traffic incident as a piecewise affine model, as illustrated in Fig. 13.4. A piecewise affine model, while parsimonious by nature, captures the two key properties of interest for traffic management, namely duration of incident and increase in occupancy—that is, a traffic jam. The primary difficulty in modeling and estimating properties of the tail of a distribution is the small number of observations available. In the case of traffic incidents, incidents of different nature occurring at different spatiotemporal locations generally have different impact. Given the small number of observations compared to the dimensionality of the state space (incidents in a typical dataset represent fewer than 0.01% of all observations), achieving high accuracy of the estimation models is a challenge.

A parsimonious parameterization of the data-constrained space of traffic incidents includes the initial occupancy $alpha$, change points $knot_1$, $knot_2$, $knot_3$, and $knot_4$, and slope $beta_1$ and $beta_3$. The incident impact prediction problem is concerned with the prediction of the impact of an incident characterized by these parameters, given incident features such as number of lanes closed and type of incident [51]. A variety of recent works has proposed social media and text analysis for such purpose, or for the real-time analysis of large-scale public events [91, 94].

In [51], three approaches are considered for predicting the incident profile parameter: decision tree (DT), multi-variate decision tree (MVDT), and neural

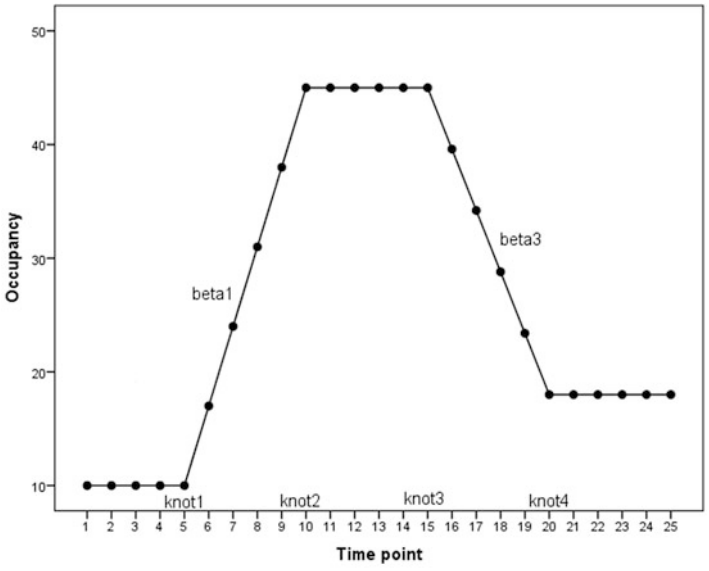


Fig. 13.4 A demonstration of a regression model on time and occupancy data; $alpha$ is the initial occupancy; $knot_1$, $knot_2$, $knot_3$, and $knot_4$ are the four change points; $beta_1$ and $beta_3$ are the slopes

network (NN). Correlation results show that the methods presented perform much better for variables directly related to incident impact such as occupancy variables, compared to indirect impact variables such as temporal parameters. Numerical results indicate that the NN model achieves the smallest prediction error.

13.3.3 Control

The last component of the real-time urban transport decision support system shown in Fig. 13.5 is the control layer. A machine learning-based system for managing the road network must go beyond real-time state-space modeling, estimation and prediction, and needs to include a means to (optimally) control the system as a function of the estimated (and predicted) state. Control plans include measures such as real-time adjustment of traffic signals as well as driver information and routing recommendations. Traffic-signal control has traditionally been based on predefined signal plans or local adaptive methods. While network-level adaptive signal control is clearly preferable in terms of optimality of the control plan, computational requirements have often precluded their real-time use in practice. Given that traffic congestion is a network phenomenon, significant attention has been devoted to finding suitable decompositions of the traffic-signal control problem so as to allow it to run efficiently in real time, as in [99, 113, 135].

Starting with California corridors [19], projects have focused on providing platforms to support the study of the problem of network-wide, or corridor-wide, estimation and control [26]. Recent projects have proposed practical solutions for managing traffic at a network level [55, 70]. The multi-year DECON project [14, 18]

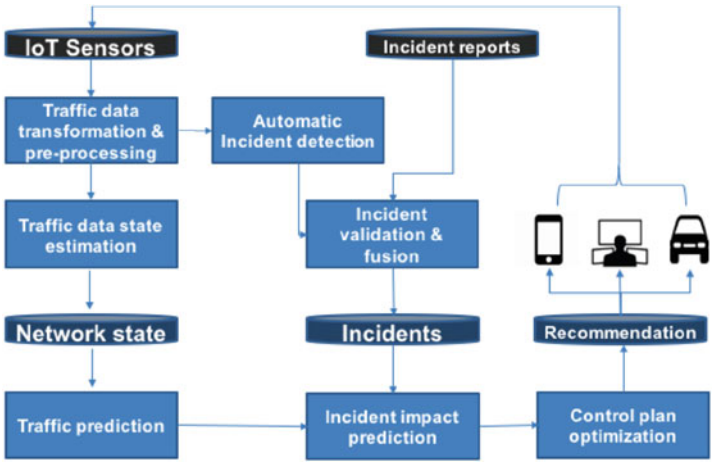


Fig. 13.5 A component flow for a real-time urban transport information and decision support system

aimed at demonstrating the benefits of machine learning solutions for road transport management. The system focuses on managing the impact of traffic incidents, with localized spatiotemporal properties.

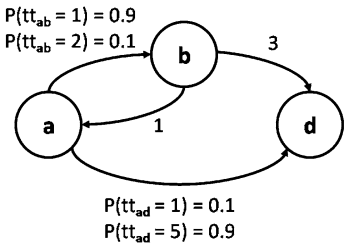
Adaptive traffic control methods [88], able to respond to current local traffic conditions, have been shown to significantly improve on traditional time-based control systems in terms of local congestion, in particular on highways [120], where traffic is more predictable and data more widely available. Recently, inspired by stability results on telecommunications networks [114], the benefits of decentralized adaptive traffic control approaches for network-wide stability have been more thoroughly studied, in particular on arterial networks, and it has been shown [121, 124] that this class of methods, while completely decentralized and operating at isolated intersections, are throughput-optimal, in the sense that there is a vanishing probability of having infinite queues. This work has been subsequently extended to include more realistic constraints, see [47, 48].

These advances in decentralized feedback control applied to traffic-signal actuation have revived the question of the stability of a transportation network, dating back to the work of Smith [107, 108]. Recent research has focused on the impact of IoT, smarter devices, and adaptive decision-making [16] on the stability of transportation networks traversed by human or actuated vehicles. The stability properties of associated equilibria, in the context of normal conditions or incidents, have been thoroughly characterized in [21, 22].

The common use of node-based routing rates in modeling traffic flow and in studying network traffic stability illustrates the tight connection between the macroscopic scale of network traffic flow and the microscopic scale of individual drivers and driving vehicles. In the context of IoT and in particular mobile sensing [68] and mobile routing [15], significant attention has been devoted to online adaptive routing.

For the user, online adaptive routing means that the route is not specified at the onset of the journey. An example of such an approach is the *stochastic on-time arrival problem* introduced in [35], which focuses on identifying the policy, i.e., adaptive turn-by-turn directions, maximizing the probability of reaching the target destination given a deadline. This adaptive formulation improves significantly on the traditional path-based initial solution, since random traffic events can cause the optimal route to contain cycles; see Fig. 13.6. Recent work [86] has shown that this problem can be solved much faster than expected, although it is still far from being solvable at speeds required for continent-scale routing [8]. Another issue with the promise of online routing lies in the lack of robustness of the proposed policy to the uncertainty in travel-time estimates. Typical estimation methods being only able to accurately provide average travel time, the need for a complete travel-time distribution can be problematic. The work of [37, 57] addresses this problem and proposes robust counterparts to the online stochastic routing problem.

Ideally, personalized recommendations should be made to users. Recommendations need not always be prescriptive, but may instead aim to nudge the user to take better actions. Congestion pricing, introduced in Singapore several decades ago [46], is one such means of encouraging users to make the socially better route or



Path	Travel-time	Probability
$\{(a,b),(b,d)\}$	4	0.9
$\{(a,d)\}$	1	0.1
$\{(a,b),(b,a),(a,d)\}$	4	0.01

Fig. 13.6 A simple network with an optimal routing policy that may contain a loop. Links (b, d) and (b, a) have deterministic travel times of, respectively, 3 and 1 time units. Link (a, b) has a travel time of 1 with probability 0.9 and a travel time of 2 with probability 0.1. Link (a, d) has a travel time of 5 with probability 0.9 and a travel time of 1 with probability 0.1. Adapted from [100]

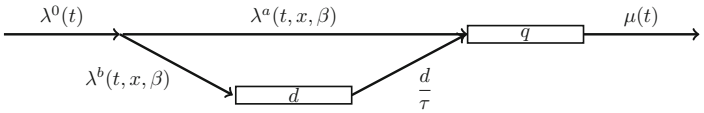


Fig. 13.7 Scheme of a pedestrian queue with diversion

mode choices. Recent work has proposed the concept of token-based pricing [46], inspired by telecommunications networks, and actualized the concept of cordon-based pricing [131]. A novel approach towards the use of incentives for influencing transportation choices in a city are the so-called tradable credit schemes. Tradable credits describe a type of fictitious currency, in general provided with no or minimal charge to eligible participants, along with a market created for their exchange. The idea of issuing credits to all eligible individuals and allowing trading among the individuals to encourage a certain outcome is appealing in that it leverages the psychological benefit of providing positive rewards to all individuals.

Studies of tradable credits have gained momentum in recent years in the transportation literature, following the seminal paper of Yang and Wang [128], which introduced the concept for road network congestion mitigation. In [69], the authors describe a tradable credits scheme for the choice of using one's private car versus using public transport and show that it admits an equivalent potential game, so that equilibria can be provably reached using simple learning algorithms such as best response dynamics. With IoT-based time-distance-place road charging being deployed over the coming years, such novel transport management strategies are becoming realistic, and the choice to drive or not can be linked to a fixed fee toll, or indeed to a tradable credits scheme. Furthermore, when public transport payment uses the same smart card as the payment of the road usage (via tolls or tradable credits), the usage of the two may be linked.

Incentives for pedestrians are discussed in [27]. The authors present an approach for optimally determining the minimal incentives needed to encourage pedestrians to perform an alternative activity and thus to defer joining a queue when the queue is crowded, as illustrated in Fig. 13.7.

13.4 Municipal Water Distribution

Water demand (consumption) drives a variety of operational decisions for water utilities covering asset management, source water production, treatment volumes, tank levels, and valve settings, as well as pressures within the distribution network [13]. Estimating daily water consumption 24–48 h in advance to determine the volume to treat or purchase depends on an accurate demand forecast [28]. Higher temporal resolution forecasts, e.g., hourly, may be necessary to take advantage of electricity rate structures for managing pumping activities [3].

A conservative estimate of the total annual cost of non-revenue water (NRW) to utilities worldwide is USD14 billion [65]. Energy can comprise 30–40% of a utility's total operating costs [117]. Therefore, reducing NRW and energy consumption are key objectives, and a finely tuned demand-response plan is a powerful strategy to achieve them. Exploitation of currently existing data acquisition systems and the development of dynamic forecasting methods will continue to grow [5, 6].

13.4.1 Water Demand Forecasting

Forecasting demands within drinking-water networks has been an active area of research for at least 40 years. Initial work focused on employing statistical models including multiple regression and time-series analysis to provide these forecasts. Over the past 15 years, forecasting algorithms built on machine learning or artificial intelligence (AI) approaches have become popular, with artificial neural networks (ANN), support vector regression (SVR), and fuzzy logic algorithms all being successfully applied to demand forecasting. Currently, a research emphasis on hybrid forecasting techniques that combine statistical time-series tools with machine learning tools is underway with the goal of combining the best attributes of traditional time-series techniques and improvements accessible through AI approaches. Further background on the evolution of demand forecasting and their comparison on example problems can be found in [1, 4, 7, 53].

This section presents a water-usage forecasting method based on seasonal autoregressive integrated moving average (SARIMA) models [106] and assimilates demand measurements using a Kalman filter to produce online forecasts and estimates of uncertainty. Unlike ANNs or other "black box" models, SARIMA can be cast in state-space form. Parametric structures suitable for water demands with temporal resolutions ranging from sub-hourly to daily are identified, and offline and online forecasting approaches are discussed. The offline mode is suitable for utility operations such as sizing daily water production, while the online mode is often better suited for operations such as scheduling pumps. The forecast horizon is fixed to 24 h for consistency with the daily planning of water utilities.

SARIMA Model A SARIMA model denoted by ARIMA$(p, d, q)(P, D, Q)$ is described in [106]. It is compactly formulated as

$$\Phi_P(B^s)\phi(B)\nabla_s^D\nabla^d x_t = \delta + \Theta_Q(B^s)\theta(B)\epsilon_t. \tag{13.4.1}$$

Here, x_t represents the measured water-demand time series and ε_t a random error process; t is the time index. B is the backshift operator, $B^k x_t = x_{t-k}$ for $k = 0, 1, 2, \ldots$. $\Phi_P(B^s)$ is the seasonal autoregressive polynomial, $\Theta_Q(B^s)$ the seasonal moving average polynomial, $\phi(B)$ the ordinary autoregressive polynomial, and $\theta(B)$ the ordinary moving average polynomial, which are defined by the expressions

$$\begin{aligned}
\Phi_P(B^s) &= 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps},\\
\Theta_Q(B^s) &= 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \cdots + \Theta_Q B^{Qs},\\
\phi(B) &= 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p,\\
\theta(B) &= 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q.
\end{aligned}$$

P, Q, p, and q are the orders of the respective polynomials, and s is the seasonal period. The model involves the seasonal differencing operator $\nabla_s^D = (1 - B^s)^D$ and the ordinary differencing operator $\nabla^D = (1 - B)^d$. Lastly, $\delta = \mu(1 - \phi_1 - \ldots - \phi_p)(1 - \Phi_1 - \ldots - \Phi_P)$ is the intercept, where μ is the mean of the demand time series.

State Space Model To facilitate online updating with a Kalman filter, the SARIMA model is cast as state-space model decomposed into observation and state equations,

$$\mathbf{y}_t = \mathbf{A}^{\mathrm{T}}\mathbf{u}_t + \mathbf{H}^{\mathrm{T}}\mathbf{z}_t + \mathbf{w}_t, \tag{13.4.2}$$

$$\mathbf{z}_t = \mathbf{F}\mathbf{z}_{t-1} + \mathbf{v}_t. \tag{13.4.3}$$

where $\mathbf{y}_t$ is the vector of observations, $\mathbf{z}_t$ the state vector consisting of unobserved variables, and $\mathbf{u}_t$ a vector of predetermined variables that possibly are lagged values of $\mathbf{y}_t$. $\mathbf{A}$ is a predetermined matrix, $\mathbf{F}$ and $\mathbf{H}$ are parameter matrices, $\mathbf{w}_t$ and $\mathbf{v}_t$ are observation and model error terms, respectively, assumed to be distributed as white noise with covariance matrices W and R [49]. Equation (13.4.2) has the form of a linear regression model, and Eq. (13.4.3) is written as a first-order vector autoregressive model [49].

To obtain the state-space form, a SARIMA$(p, d, q)(P, D, Q)_s$ model of the univariate water-demand series $\mathbf{y}_t$ is written as an equivalent ARIMA$(p + sP, q + sQ)$ [103] for a transformed variable $\mathbf{y}_t^*$,

$$\mathbf{y}_t = \mathbf{H}^{\mathrm{T}}\mathbf{z}_t, \tag{13.4.4}$$

$$\mathbf{z}_t = \mathbf{F}\mathbf{z}_{t-1} + \mathbf{v}_t, \tag{13.4.5}$$

where $\mathbf{A}$, $\mathbf{u}_t$, and $\mathbf{R}$ are set to zero as no exogenous variables are considered in the predictions here and measurement noise is considered to be negligible. To apply the

Kalman filter, the matrices $\mathbf{F}$, $\mathbf{H}$, $\mathbf{v}$, and $\mathbf{W}$ need to be computed using the previously estimated parameters of the SARIMA model.

The Kalman filter allows updating the state vector $\mathbf{z}_t$ every time there is a new observation, $\mathbf{y}_t$ [50]. The filter consists of a sequence of steps, where $\hat{\mathbf{z}}_t$ is linearly estimated from known values of $\hat{\mathbf{z}}_{t-1}$ and $\mathbf{y}_t$. An initial step is required where the prior state $\mathbf{z}_1$ is computed from prior information or assumed to be zero in the absence of such information. The variance of the prior state, $P_{1|0}$, is also calculated using estimates of $\mathbf{F}$ and $\mathbf{W}$ and a prior estimate of $P_{1|0}$,

$$\hat{\mathbf{z}}_{t+1|t} = \mathbf{F}\hat{\mathbf{z}}_{t|t-1} + \mathbf{F}\mathbf{P}_{t|t-1}\mathbf{H}(\mathbf{H}^{\mathrm{T}}\mathbf{P}_{t|t-1}\mathbf{H}^{-1}(\mathbf{y}_t - \mathbf{H}^{\mathrm{T}}\hat{\mathbf{z}}_{t|t-1})), \tag{13.4.6}$$

where $\hat{\mathbf{z}}_{t|t-1}$ is the forecast of the true state $\mathbf{z}_t$ based on a linear function of the observations $\mathbf{y}_1, \ldots, \mathbf{y}_{t-1}$ and $\mathbf{P}_{t|t-1}$ is the variance of this forecast. The forecast variance $\mathbf{P}_{t+1|t}$ is updated through

$$\mathbf{P}_{t+1|t} = \mathbf{F}\left[\mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1}\mathbf{H}(\mathbf{H}^{\mathrm{T}}\mathbf{P}_{t|t-1}\mathbf{H}^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{P}_{t|t-1})\right]\mathbf{F}^{\mathrm{T}} + \mathbf{W}, \tag{13.4.7}$$

Throughout the remainder of the chapter, we will use the term "offline" to refer to forecasts that are obtained by means of parameter re-estimation and "online" to mean forecasts that result from updating the model parameters (in state-space form) by applying the Kalman filter.

Water Demand Data The data examined here are a set of flow rate measurements generated by 550+ individual telemetry instruments distributed over 190 district-metered areas (DMAs) in the city of Dublin, Ireland. The time series covers a period of 19 months, from January 2010 through July 2011, with a temporal resolution of 15 min. The measurements from individual DMAs were aggregated to obtain a total-system demand series. The mean value (144.11 ML/d) represents roughly 30% of Dublin's total consumption; however, from a modeling perspective the data correspond to a high level of aggregation and qualitatively have the characteristics of a total-system demand series. The data were divided into a training set (60%) and a validation set (40%); see [4] for more details.

An extensive model selection process was conducted using offline training and validation [4]; performance measures included the auto and partial correlation functions of the residuals and various information criteria summary measures of prediction error.

The results of the model selection process indicate that a seasonal moving average process is most suitable; thus, the seasonal autoregressive and seasonal moving average orders were assigned values $P = 0$ and $Q = 1$, respectively. From further detailed analysis, the autoregressive and moving average orders were assigned values $p = 0$ and $q = 4$. The final general structure of the model is

$$(1 - B^s)(1 - B)x_t = \delta + (1 + \Theta_1 B^s)\left(1 + \sum_{i=1}^{4} \theta_i B^i\right)\epsilon_t, \tag{13.4.8}$$

where s depends on the temporal resolution and the seasonal correlation period; the vector of parameters to estimate is $\Theta = (\Theta_1, \theta_1, \theta_2, \theta_3, \theta_4, \sigma, \delta)^{\mathrm{T}}$. The model designations are assigned as the product of the sampling frequency, inverse of the sample interval, and the seasonal period of the model; see Table 13.1.

Forecasting Results In addition to the model parameters and the demand measurements, the forecasting algorithm requires values for the length of the training window, τ, and the forecast horizon, h. The length τ was selected through experimentation with values ranging from 7 to 28 days. A 7-day window represents the minimum length of data required to fit a model with a 1-week seasonal period. Even though the models with a daily period require smaller windows, the same lower limit was used for all models to facilitate a comparison between periods. The forecast horizon h was set to 24 h.

Increases in τ resulted in increased prediction error for hourly and sub-hourly sample data; however, for daily sample data, increasing τ decreased the prediction errors. This behavior is due to increased noise in the data collected with the shorter sample intervals. For the final application, we used $\tau = 7$ days for forecasting data where the sample interval was 1 h or less, and $\tau = 28$ days in cases where the sample interval was more than 1 day.

A sample of the online forecasts for the sub-hourly (15 min) and daily models is presented in Fig. 13.8. Each plot shows a data segment with the 1-day-ahead forecasts and the uncertainty bands (95% confidence level) for the corresponding data source and resolution. The figure illustrates how the different models respond to the data characteristics. For instance, the top panel displays good agreement between data and forecast at the level of total-system demand and when the temporal resolution is high. The bottom panel also shows good forecast behavior for a coarser temporal resolution of 1 day. In general, daily models perform better than higher-resolution models due to noise reduction, or smoothing, of the demand signal through temporal aggregation.

Using a set of SARIMA parameters estimated during the online forecasting stage, the Kalman filter was applied to generate online forecasts for each model structure (Table 13.1). The model parameters were not recalculated in online mode; hence, the online computational performance was substantially higher, as much as a 97% reduction in runtime. In most cases, the quality of the forecasts was considerably

Table 13.1 Model performance results

Model	Sample interval (h)	Seasonal period (days)	RMSE (ML/d)		MAPE (%)	
			Offline	Online	Offline	Online
S-96	0.25	1	9.57	3.04	4.21	1.49
S-672	0.25	7	9.31	2.52	3.55	1.43
S-24	1.0	1	8.80	3.21	3.51	1.49
S-168	1.0	7	9.30	4.95	3.53	1.83
S-1	24.0	1	2.48	3.08	1.29	1.96
S-7	24.0	7	2.36	2.33	1.27	1.10

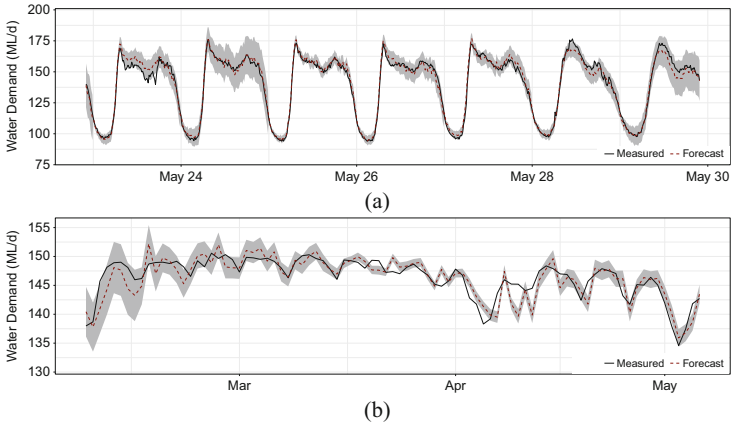


Fig. 13.8 Sample of the demand forecasts by the models with sub-hourly and daily resolutions; (**a**) S-672 model and data, (**b**) S-7 model and data. The grey bounds indicate the forecast uncertainty (95% confidence level)

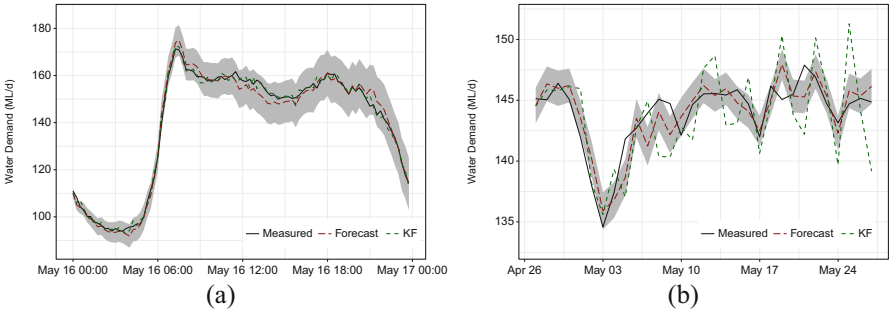


Fig. 13.9 Comparison of the forecasts before and after applying the Kalman filter; (left) S-672, (right) S-7

improved as well. An example comparison is presented in Fig. 13.9, where it is clear that for the S-672 model the Kalman filter noticeably increases the prediction accuracy; for the S-7 model the improvements are smaller.

RMSE and MAPE results are shown in Table 13.1. MAPE is considered the most relevant for cross comparison as it normalizes the errors at every measurement. The median percentage error for all runs is below 2.5%. The results in Table 13.1 show that tailoring the seasonal component to the temporal resolution of the data can improve results, with a 7-day period producing MAPE values as low or lower than the shorter 1-day period.

Addition of the online Kalman filter to the predictions reduces the RMSE and MAPE values for the hourly and sub-hourly sample data. Additionally, the Kalman filter homogenizes the MAPE values to be below 2.0% across all temporal intervals and seasonal period values. Using the online Kalman filter increases the MAPE

value for the S-1 model relative to the offline results. This increase is due to the lack of any updating of the SARIMA parameters. For daily forecasting, a frequent (e.g., daily) update of the SARIMA parameters is recommended.

13.4.2 Pump Scheduling Under Uncertainty

Water utilities have optimized pump schedules to take advantage of day/night electricity pricing plans for several decades. As intermittent renewable energy sources such as solar and wind power provide an increasingly large share of the available electricity, energy providers are moving to dynamic pricing schemes, where the electricity price is forecast 24 h in advance with relatively fine temporal resolution (e.g., 30 min). Water utilities are uniquely positioned to take advantage of dynamic pricing schemes using their existing infrastructure for pumping and storage to respond to changing costs for power. Optimization of the pump schedules under uncertainty in the forecasted energy prices is necessary to minimize electricity prices.

Techniques to minimize pumping costs in water systems have been the target of considerable research. Since the early work in the mid-1970s, both mathematical optimization and metaheuristic approaches to the problem have been proposed. Mathematical programming methods include dynamic programming [110], linear programming [60], and mixed-integer linear programming [75]. Metaheuristic approaches to optimization have also been applied, including hierarchical decomposition [111], genetic algorithms [119], ant colony optimization [76], and simulated annealing [79]. The most recent approaches, termed hybrid methods, combine mathematical programming techniques with heuristics to arrive at good solutions quickly. For example, [45] solves a linear programming relaxation and then uses this solution to start a greedy algorithm. An approach combining mixed-integer programming and hydraulic simulation is presented in [85]. A theoretical lower bound for pump scheduling costs is obtained by [44] by applying a quadratic approximation for pipe friction [31] and relaxing dynamic constraints.

Most previous research considers the energy price to be given as an input. The main mechanism of including uncertainty in electricity prices has been through a maximum demand charge. Under a maximum demand tariff, the total energy cost depends on both the time of consumption and the maximum power demand over a billing period such as 1 month. The maximum power demand depends in turn on the water demand. An optimization method for this problem is proposed in [78]. Their method finds a maximum power demand and uses this to constrain the daily operational schedule.

The present work focuses on optimizing the daily pumping schedule, but with the consideration that the energy price fluctuates at 30-min intervals. These fluctuations occur when renewable but variable power sources such as wind energy connect to the grid [43]. Historical data on actual prices are used to condition stochastic sampling of daily energy price trajectories, using covariance decomposition methods.

From this ensemble of realizations, electricity price profiles are classified into a handful of scenario classes. The optimal pumping schedule for each price class is then computed. Once the pumping schedule is known, the probability distribution of cost for that schedule is evaluated using Monte Carlo methods.

In the present section, we summarize the developments reported in [30] for an operational technique for generating pump schedules and quantifying the uncertainty in the costs of these schedules. Given this information, a system operator can pump according to a desired level of risk. The overall technique proposed here is comprised of several steps:

1. Collect a sample of observed historical price trajectories that represent the time of interest.
2. Expand the sample size by creating an ensemble that fully describes the possible price scenarios. In this case, 1000 random samples of price profiles are generated and are statistically similar to the observations.
3. Classify the ensemble members into 10 clusters and identify the medoid price trajectory of each cluster.
4. Compute the optimal pumping schedule for each cluster using the medoid price scenario.
5. Estimate the probability density function of daily pumping costs for each schedule using the set of random samples.
6. Compute the desired objective value from this bootstrapped probability density function.

Electricity Price Scenario Simulation Expansion of the size of the observed price scenario dataset to a full ensemble that quantifies uncertainty and can be used in pump schedule optimization is necessary. The price scenario simulation process must be capable of preserving the distribution of prices at every time step across the set of ensembles as well as preserving the temporal correlation of every scenario across all time steps. The price scenario simulation method proceeds through a series of four steps:

1. Data Whitening. Temporal correlation in the observed price scenarios is removed by centering and decorrelating the price scenario: $Y = W(x - \mu)$, where μ is the vector of row means of x and the whitening matrix $W = \Delta^{-1/2}U^{\mathrm{T}}$ is computed from an eigen-decomposition of the sample covariance matrix, C. If C is symmetric positive definite, Δ is the diagonal matrix of eigenvalues and U is the corresponding matrix of eigenvectors.
2. Positive Definite Covariance. The covariance matrix defined by the observations will not necessarily be positive definite, and, due to noise in the data, smaller eigenvalues can be near zero. A reduction in the dimensionality by removing the lowest eigenvalues (summing to less than 5% of the total spectral energy) eliminates these issues and guarantees a positive definite covariance matrix for use in the inverse whitening transform.
3. Simulation and Back Transformation. Generate random uncorrelated samples by selecting a value, y_i, at random and replacing that value with a Gaussian

deviate drawn from a distribution with mean y_i and standard deviation chosen as $1.06\sigma\, N^{-0.20}$, where N is the number of samples and σ is the sample standard deviation [33]. In this case, $\sigma = 1$ due to the whitening transform (Step 1). The simulated vector is then transformed back to the original observation space using the inverse whitening transform with the updated matrices calculated in Step 2.

4. Final Check. The new simulated price scenario, $\hat{x}$, is checked for feasibility. Here, simulated scenarios with values outside the range (min-max) of the observed prices are discarded.

Testing of this price scenario simulation process demonstrates that the generated scenarios reproduce the observed distributions at each time step and also reproduce the measured temporal correlation. Clustering of the simulated price scenarios is done using the partition around medoids (PAM) algorithm developed in [64] with $k = 10$ medoids and a Euclidean dissimilarity metric.

Optimal Pump Scheduling The optimization problem considered here is to schedule pumps to minimize electricity costs. The formulation given here considers constant-speed centrifugal pumps. The technique of piecewise linearization is used to model nonlinear pump and pipe hydraulics in a mixed-integer linear program (MILP).

A water distribution network comprises N_n nodes connected by N_l links. A subset y of the nodes are tanks, and a subset p of the links are pumps. The decision variables for each time t in the planning horizon are the hydraulic head at each node, $h_{i,t}$; the flow rate through each link, $q_{ij,t}$; and the schedule for each pump, $s_{p,t}$. The objective function is defined as the total cost of electricity for operating N_p pumps over N_t time periods, each with duration Δt,

$$\min \sum_{t=1}^{N_t} \sum_{p=1}^{N_p} \frac{\gamma\, C_t}{\eta_p} \Delta h_{p,t} q_{p,t}\, \Delta t. \tag{13.4.9}$$

The cost of electricity in €/kWh, C_t, may be different for each time period. The total head delivered by a pump is $\Delta h_{p,t}$, and the volumetric flow rate is $q_{p,t}$. The pump efficiency is γ_p, and the specific weight of water is γ. The minimization is carried out under the hydraulic and operational constraints on the network, including energy conservation on every link (pipe) between nodes and mass conservation at every node, with formulations taking into account the particular shape and geometry for each node that is a tank.

Several operational constraints apply to the solution of this problem. The head (and thus level) in each tank should be at least as high at the end of the planning period as the beginning,

$$h_{y,N_t} \geq h_{y,1}. \tag{13.4.10}$$

The flow rate for each pump is positive while operating or zero when off,

$$0 < q_{p,t} \leq s_{p,t} Q_p^u. \tag{13.4.11}$$

The number of times a pump can be started within a given time period is limited to N_s,

$$\sum_{t=2}^{N_t} (s_{t,p} - s_{t-1,p}) \leq N_s. \tag{13.4.12}$$

The optimization problem (13.4.9), subject to the constraints mentioned here, is solved using CPLEX [24] to find the minimum cost pumping schedules that satisfy the physical requirements of the system. Further details are provided in [30].

Application The methods outlined above were applied to a small network, where electricity costs dominate operational expenses and dynamic pricing tariffs are available. The study considered pumping operations of the network for a "typical" day in May 2013. A hydraulic simulation model was developed and calibrated. The system includes two pumps drawing from a single reservoir, each pumping into a storage tank approximately 50 m above the reservoir. The storage tanks are connected, and system demands are allocated to a single node connected to the second storage tank. The hydraulic model was applied each day for 7 days, and simulated tank levels closely approximated measured levels.

Dynamic electricity prices for 68 days from April to June 2013 were obtained from the electricity market. Prices fluctuate on a 30-min basis according to supply and demand. Over the study period, energy prices ranged from 5 to 262 €/mWh, with an average value of 63 €/mWh. Price levels were correlated to the time of day but showed considerable fluctuation.

The actual electricity price profiles were used to randomly generate a sample of 1000 similar price profiles using the methods described above. The distribution of electricity price within each of the 48 half-hour periods was compared between the simulated and observed electricity prices. The generated price profiles were similar to observed ones in terms of median value, inter-quartile range, and extreme values. The simulated electricity price profiles were grouped into ten clusters, and the price profile nearest the medoid of the cluster was identified in Fig. 13.10.

An optimal pumping schedule was computed for each of the 10 medoid price profiles. The resulting schedules (Fig. 13.11) are constrained to use a maximum of 6 pump starts per day and tend to emphasize pumping at night, when prices are generally lower. Between the schedules there are fluctuations in pump start and stop times according to the different prices. Uncertainty in the daily pumping cost was explored through calculation of the cost of pumping schedules A–J for every one of the 1000 randomly generated price profiles.

Considering the foregoing analysis, the question remains as to which pump schedule should be selected. The answer depends on the criterion of the system operator and the desired risk profile. Different criteria will produce different

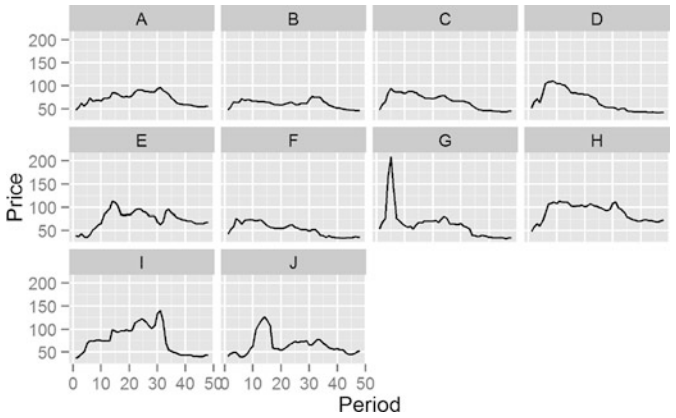


Fig. 13.10 Medoid price trajectories for ten clusters [30, used with permission]

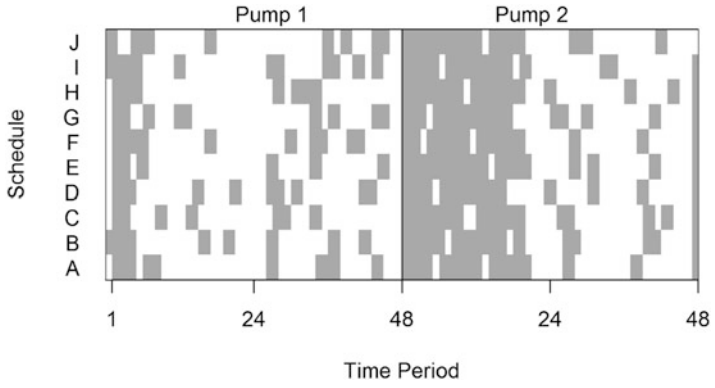


Fig. 13.11 Optimal pump schedules (grey: on, white: off) for 10 medoid price trajectories ([30], used with permission)

Table 13.2 Values of optimal schedules (€/day) over different measures of pump schedule performance

Measures	Schedule	Value	Next best
Lowest average cost	C	671	A 673
Lowest Std. error of mean	A	2.66	C 2.69
Lowest median cost	C	667	A 668
Lowest inner quartile range	A	103	C 104
Lowest maximum cost	A	971	H 987

schedules; see Table 13.2. For example, the schedule with the lowest average cost (schedule C) differs from that with the lowest maximum cost (schedule A).

Conclusion This work demonstrates how combining a series of techniques can provide a method for scheduling pumps when electricity prices are unknown.

Exploitation of the observed covariance between historical energy prices enables efficient simulation of possible energy price scenarios that can be used to quantify price uncertainty. Optimization of pump schedules across the entire ensemble of simulated price scenarios is not feasible. By clustering energy price profiles into groups, only a modest number of pump scheduling optimizations are required. Once schedules are obtained, it is possible to evaluate the cost of many possible price profiles and estimate their cost distribution.

Future work on this problem could examine the role of ensembles of price predictions in selecting an optimal schedule. Optimization approaches can utilize the ensemble of scenarios to identify pumping schedules that target cost minimization against the most likely and/or the worst case scenarios. More complex approaches, such as robust optimization, utilize multiple scenarios to identify "best worst-case" scenarios.

13.5 Perspectives

In this chapter, we demonstrated how the Internet of Things along with machine learning can help cities improve the management of their infrastructure. The basic ingredients include an urban-scale network, a state-space model, and techniques for estimation—prediction and optimization—control. We considered two basic components of a city's infrastructure in detail, namely the transportation network and the municipal water supply system. We discussed the similarities of these two types of networks in the context of monitoring and operations, and illustrated the general nature of current applications developed for these networks. The current explosion of IoT data has motivated innovative research explorations, which offer enhanced models, more accurate estimates, and fine-grained control. Edge devices such as smart meters and smart phones have played the role of intermediary between the world of bits and the world of atoms. For the two types of networks considered, personalization remains a key goal of the revolution brought by IoT, where devices will able to effectively inform and guide individuals in their travels and manage their individual consumption of services at home. Other applications include, for instance, the energy domain, wherein the onset of electric vehicles, smart home appliances, and on-site generation of electricity through rooftop photovoltaic panels will create a greater coupling between transportation, activities, and energy.

One important perspective for research is that with increased volumes of sensor data through IoT devices, it is crucial to quantify uncertainty and to assess its impact on forecasting and decision making. New and better mathematical tools are necessary for both. Since mathematical models will be embedded in complex decision-making systems, the accuracy and reliability of those models will be critical for the efficiency and safety of the systems that comprise them.

For multimedia data in particular, the advent of deep learning is revolutionizing the sensing abilities of machines, which are now able to hear, see, and more generally acquire unstructured data at the human performance level. For higher-

level applications such as self-driving cars, where automated context understanding has not yet reached the human performance level, significant effort will be required, in particular in the modeling and analysis of uncertainty and the design of robust adaptive methods for online control.

In the future, we expect to see the emergence of networks of sensing and control agents, operating at the edge, alternating between decentralized and centralized, with coordination depending on context. To capture the complexity of the real world and safely deploy virtual agents will require fundamental innovations in modeling, estimation, and control. In particular, the impact of adverse tail events, which are very difficult to manage using data-driven methods, is likely to become a significant concern. Similarly, extending the ability of multi-scale models to adapt and specialize seems to hold valuable lessons for the modeling of intelligence and the subsequent generation of innovative research.

References

1. Adamowski, J., Fung Chan, H., Prasher, S.O., et al.: Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada. Water Resour. Res. **48**(1) (2012)
2. Alonso-Mora, J., Samaranayake, S., Wallar, A., et al.: On-demand high-capacity ride-sharing via dynamic trip-vehicle assignment. Proc. Natl. Acad. Sci. **114**(3):462–467 (2017)
3. Aly, A.H., Wanakule, N.: Short-term forecasting for urban water consumption. J. Water Resour. Plan. Manag. **130**(5), 405–410 (2004)
4. Arandia, E., Ba, A., Eck, B., McKenna, S.: Tailoring seasonal time series models to forecast short-term water demand. J. Water Resour. Plan. Manag. **142**(3), 04015067 (2016)
5. Arandia, E., Eck, B., McKenna, S.: The effect of temporal resolution on the accuracy of forecasting models for total system demand. Procedia Eng. **89**, 916–925 (2014)
6. Arandia, E., Uber, J., Boccelli, D., et al.: Modeling automatic meter reading water demands as nonhomogeneous point processes. J. Water Resour. Plan. Manag. **140**(1), 55–64 (2014)
7. Ashy, J., Ormsbee, L.E.: Short-term water demand forecast modeling techniques: conventional methods versus AI. J. Amer. Water Works Assoc. **94**(7), 64–72 (2002)
8. Bast, H., Carlsson, E., Eigenwillig, A., et al.: Fast routing in very large public transportation networks using transfer patterns. In: Algorithms–ESA 2010, pp. 290–301. Springer, Berlin (2010)
9. Baudel, T., Dablanc, L., Alguiar-Melgarejo, P., et al.: Optimizing urban freight deliveries: from designing and testing a prototype system to addressing real life challenges. Transp. Res. Proc.**12**, 170–180 (2016)
10. Bayen, A., Butler, J., Patire, A.: Mobile millennium final report. Technical report, California Center for Innovative Transportation, Institute of Transportation Studies, University of California, Berkeley, Research Report, UCB-ITS-CWP-2011-6 (2011)
11. Beckmann, M., McGuire, C.B., Winsten, C.B.: Studies in the economics of transportation. Technical report, Transportation Research Board of the National Academies (1956)
12. Beloglazov, A., Buyya, R.: Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers. Concurr. Comput. Pract. Exper. **24**(13), 1397–1420 (2012)
13. Billings, R.B., Jones, C.V.: Forecasting Urban Water Demand, 2nd edn. American Water Works Association, Denver (2008)

14. Blandin, S., Gopal, V., Thirumalai, K., et al.: Prediction of duration and impact of non-recurrent events in transportation networks. In: TRISTAN VIII. San Pedro de Atacama (2013)
15. Borokhov, P., Blandin, S., Samaranayake, S., et al.: An adaptive routing system for location-aware mobile devices on the road network. In: 2011 14th International IEEE Conference on, Intelligent Transportation Systems (ITSC) (2011). https://doi.org/10.1109/ITSC.2011.6083021
16. Boyer, S., Blandin, S., Wynter, L.: Stability of transportation networks under adaptive routing policies. Transp. Res. Proc. **7**, 578–597 (2015). https://doi.org/10.1016/j.trpro.2015.06.030
17. Cacchiani, V., Huisman, D., Kidd, M., et al.: An overview of recovery models and algorithms for real-time railway rescheduling. Transp. Res. B Methodol. **63**, 15–37 (2014)
18. Cheong, J., Blandin, S., Thirumalai, K., et al.: The development of a traffic decongestion engine with predictive analytic and decision support capabilities. In: 13th ITS Asia Pacific Forum (2014)
19. Choe, T., Skabardonis, A., Varaiya, P.: Freeway performance measurement system: an operational analysis tool. Transp. Res. Rec. **1811**, 67–75 (2002)
20. Chowdhury, D., Santen, L., Schadschneider, A.: Statistical physics of vehicular traffic and some related systems. Phys. Rep. **329**(4–6), 199–329 (2000)
21. Como, G., Lovisari, E., Savla, K.: Throughput optimality and overload behavior of dynamical flow networks under monotone distributed routing. IEEE Trans. Control Netw. Syst. **2**(1), 57–67 (2015)
22. Como, G., Savla, K., Acemoglu, D., et al.: Robust distributed routing in dynamical networks, Part I: locally responsive policies and weak resilience. IEEE Trans. Autom. Control **58**(2), 317–332 (2013)
23. Cottrill, C., Pereira, F., Zhao, F., et al.: Future mobility survey: experience in developing a smartphone-based travel survey in Singapore. Transp. Res. Rec. J. Transp. Res. Board **2354**, 59–67 (2013)
24. CPLEX, IBM ILOG: CPLEX Optimization Studio (2013). http://www-03.ibm.com/software/products/us/en/ibmilogcpleoptistud/
25. Cristiani, E., Piccoli, B., Tosin, A.: Multiscale Modeling of Pedestrian Dynamics, vol. 12. Springer, Cham (2014)
26. De Wit, C.C., Morbidi, F., Ojeda, L.L., et al.: Grenoble traffic lab: an experimental platform for advanced traffic monitoring and forecasting [applications of control]. IEEE Control. Syst. **35**(3), 23–39 (2015)
27. Desfontaines, L., Wynter, L.: Optimal decentralized queuing system with diversion: using incentives to influence behavior. In: 2016 IEEE 55th Conference on Decision and Control (CDC), pp. 1912–1919 (2016). https://doi.org/10.1109/CDC.2016.7798544
28. Donkor, E.A., Mazzuchi, T.A., Soyer, R., et al.: Urban water demand forecasting: review of methods and models. J. Water Resour. Plan. Manag. **140**(2), 146–159 (2014)
29. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. **12**(Jul), 2121–2159 (2011)
30. Eck, B., McKenna, S., Akrhiev, A., et al.: Pump scheduling for uncertain electricity prices. In: ASCE World Environmental and Water Resources Congress (2014)
31. Eck, B., Mevissen, M.: Quadratic approximations for pipe friction. J. Hydroinf. **17**(3), 462–472 (2015)
32. Eck, B., Saito, H., McKenna, S.: Temperature dynamics and water quality in distribution systems. IBM J. Res. Dev. **50**(5/6), 7:1–7:8 (2016)
33. Epanechinkov, V.: Nonparametric estimates of a multivariate probability density. Theor. Probab. Appl. **14**, 153–158 (1969)
34. Evensen, G.: Data Assimilation: The Ensemble Kalman Filter. Springer, Berlin (2009)
35. Fan, Y.Y., Nie, Y.: Optimal routing for maximizing travel time reliability. Netw. Spat. Econ. **3**(6), 333–344 (2006)
36. Ferrucci, F., Bock, S., Gendreau, M.: A pro-active real-time control approach for dynamic vehicle routing problems dealing with the delivery of urgent goods. Eur. J. Oper. Res. **225**(1), 130–141 (2013)

37. Flajolet, A., Blandin, S., Jaillet, P.: Robust adaptive routing under uncertainty. Oper. Res. **66**(1), 210–229 (2017)
38. Flynn, M.R., Kasimov, A.R., Nave, J.C., et al.: Self-sustained nonlinear waves in traffic flow. Phys. Rev. E **79**(5), 056113 (2009)
39. Fusco, F., Eck, B., McKenna, S.: Bad data analysis with sparse sensors for leak localisation in water distribution networks. In: Proceedings of the 22nd International Conference on Pattern Recognition (ICPR), pp. 3642–3647 (2014)
40. Fuxjaeger, P., Ruehrup, S., Weisgrab, H., et al.: Highway traffic flow measurement by passive monitoring of Wi-Fi signals. In: 2014 International Conference on Connected Vehicles and Expo (ICCVE), pp. 396–401 (2014). https://doi.org/10.1109/ICCVE.2014.7297578
41. Gallotti, R., Barthelemy, M.: The multilayer temporal network of public transport in Great Britain. Sci. Data **2**, 140056 (2015)
42. Garavello, M., Piccoli, B.: Traffic Flow on Networks. American Institute of Mathematical Sciences, Springfield (2006)
43. Garcia-Gonzalez, J., Moraga, R., Matres, L., et al.: Stochastic joint optimization of wind generation and pumped-storage units in an electricity market. IEEE Trans. Power Syst. **23**(2), 460–468 (2008)
44. Ghaddar, B., Naoum-Sawaya, J., Kishimoto, A., et al.: A Lagrangian decomposition approach for the pump scheduling problem in water networks. Eur. J. Oper. Res. **241**, 490–501 (2015)
45. Giacomello, C., Kapelan, Z., Nicolini, M.: Fast hybrid optimization method for effective pump scheduling. J. Water Resour. Plan. Manag. **139**(2), 175–183 (2013)
46. Goh, M.: Congestion management and electronic road pricing in Singapore. J. Transp. Geogr. **10**, 29–38 (2002)
47. Gregoire, J., Frazzoli, E., de la Fortelle, A., et al.: Back-pressure traffic signal control with unknown routing rates. IFAC Proc. Vol. **47**(3), 11,332–11,337 (2014)
48. Gregoire, J., Qian, X., Frazzoli, E., et al.: Capacity-aware backpressure traffic signal control. IEEE Trans. Control Netw. Syst. **2**(2), 164–173 (2015)
49. Hamilton, J.: Time Series Analysis. Princeton University Press, Princeton (1994)
50. Harvey, A.: Forecasting Structural Time Series Models and the Kalman Filter. Cambridge University Press, Cambridge (1989)
51. He, Y., Blandin, S., Wynter, L., et al.: Analysis and real-time prediction of local incident impact on transportation networks. In: 2014 IEEE International Conference on Data Mining Workshop (ICDMW), pp. 158–166. IEEE, Piscataway (2014)
52. Herrera, J.C., Work, D., Ban, X., et al.: Evaluation of traffic data obtained via GPS-enabled mobile phones: the mobile century field experiment. Transp. Res. C **18**, 568–583 (2009)
53. Herrera, M., Torgo, L., Izquierdo, J., et al.: Predictive models for forecasting hourly urban water demand. J. Hydrol. **387**(1–2), 141–150 (2010)
54. Hey, T., Tansley, S., Tolle, K. (eds.): The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, Redmond (2009)
55. Hoogendoorn, S., Landman, R., van Kooten, J., et al.: Integrated network management Amsterdam: towards a field operational test. In: Transportation Research Board 93rd Annual Meeting, 14-2755 (2014)
56. Horn, C., Kern, R.: Deriving public transportation timetables with large-scale cell phone data. Procedia Comput. Sci. **52**, 67–74 (2015). http://dx.doi.org/10.1016/j.procs.2015.05.026
57. Jaillet, P., Qi, J., Sim, M.: Routing optimization under uncertainty. Oper. Res. **64**(1), 186–200 (2016)
58. Jeong, Y.S., Byon, Y.J., Castro-Neto, M.M., et al.: Supervised weighting-online learning algorithm for short-term traffic flow prediction. IEEE Trans. Intell. Transp. Syst. **14**(4), 1700–1707 (2013). https://doi.org/10.1109/TITS.2013.2267735
59. Jiang, Y., Jiang, Z.P.: Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics. Automatica **48**(10), 2699–2704 (2012)
60. Jowitt, P.W., Germanopoulos, G.: Optimal pump scheduling in water supply networks. J. Water Resour. Plan. Manag. **118**(4), 406–422 (1992)

61. Kaipio, J., Somersalo, E.: Statistical and Computational Inverse Problems. Springer, New York (2005)
62. Kalman, R.: A new approach to linear filtering and prediction problems. J. Basic Eng. **82**(1), 35–45 (1960)
63. Kamarianakis, Y., Shen, W., Wynter, L.: Real-time road traffic forecasting using regime-switching space-time models and adaptive lasso. Appl. Stoch. Model. Bus. Ind. **28**(4), 297–315 (2012)
64. Kaufman, L., Rousseeuw, P.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York (1990)
65. Kingdom, B., Liemberger, R., Marin, P.: The challenge of reducing non-revenue water (NRW) in developing countries. Water Supply and Sanitation Sector Board Discussion Paper Series, Paper no 8, The World Bank, Washington D.C. (2006)
66. Kitchin, R.: The real-time city, big data and smart urbanism. GeoJournal **79**(1), 1–14 (2014)
67. Koch, M.W., McKenna, S.A.: Distributed sensor fusion in water quality event detection. J. Water Resour. Plan. Manag. **137**(1), 10–19 (2010)
68. Krause, A., Horvitz, E., Kansal, A., et al.: Toward community sensing. In: Proceedings of the 7th International Conference on Information Processing in Sensor Networks, pp. 481–492. IEEE Computer Society, Washington D.C. (2008)
69. Lahlou, S., Wynter, L.: A Nash equilibrium formulation of a tradable credits scheme for incentivizing transport choices: from next-generation public transport mode choice to HOT lanes. Transp. Res. B Methodol. **101**, 185–212 (2017)
70. Landman, R., Hoogendoorn, S., Westerman, M., et al.: Design and implementation of integrated network management in the Netherlands. In: TRB 89th Annual Meeting Compendium of Papers DVD (2010)
71. Le, T.V., Song, B., Wynter, L.: Real-time prediction of length of stay using passive Wi-Fi sensing. In: Communications (ICC), 2017 IEEE International Conference on, pp. 1–6. IEEE, Piscataway (2017)
72. LeCun, Y., Bengio, Y. , Hinton, G.: Deep learning. Nature **521**, 7553 (2015)
73. Lighthill, M., Whitham, G.: On kinematic waves, II: a theory of traffic flow on long crowded roads. Proc. R. Soc. Lond. **229**(1178), 317–345 (1956)
74. Lippi, M., Bertini, M., Frasconi, P.: Short-term traffic flow forecasting: an experimental comparison of time-series analysis and supervised learning. IEEE Trans. Intell. Transp. Syst. **14**(2), 871–882 (2013). https://doi.org/10.1109/TITS.2013.2247040
75. Little, K., McCrodden, B.: Minimization of raw water pumping costs using MILP. J. Water Resour. Plan. Manag. **115**(4), 511–522 (1989)
76. Lopez-Ibanez, M., Prasad, T., Paechter, B.: Ant colony optimization for optimal control of pumps in water distribution networks. J. Water Resour. Plan. Manag. **134**(4), 337–346 (2008)
77. Lv, Y., Duan, Y., Kang, W., et al.: Traffic flow prediction with big data: a deep learning approach. IEEE Trans. Intell. Transp. Syst. **16**(2), 865–873 (2015). https://doi.org/10.1109/TITS.2014.2345663
78. McCormick, G., Powell, R.: Optimal pump scheduling in water supply systems with maximum demand charges. J. Water Resour. Plan. Manag. **129**(5), 372–379 (2003)
79. McCormick, G., Powell, R.: Derivation of near-optimal pump schedules for water distribution by simulated annealing. J. Oper. Res. Soc. **55**(7), 728–736 (2004)
80. McKenna, S.A., Fusco, F., Eck, B.J.: Water demand pattern classification from smart meter data. Procedia Eng. **70**, 1121–1130 (2014)
81. McKenna, S.A., Klise, K.A., Wilson, M.P.: Testing water quality change detection algorithms. In: Proceedings of 8th Water Distribution Systems Analysis (WDSA) Symposium. ASCE, Reston (2006)
82. Messmer, A., Papageorgiou, M.: Metanet: A macroscopic simulation program for motorway networks. Traffic Eng. Control **31**(9), 466–470 (1990)
83. Min, W., Wynter, L.: Real-time road traffic prediction with spatio-temporal correlations. Transp. Res. Part C Emerg. Technol. **19**(4), 606–616 (2011)

84. Mitrovic, N., Asif, M.T., Dauwels, J., et al.: Low-dimensional models for compressed sensing and prediction of large-scale traffic data. IEEE Trans. Intell. Transp. Syst. **16**(5), 2949–2954 (2015)
85. Naoum-Sawaya, J., Ghaddar, B., Arandia, E., et al.: Simulation-optimization approaches for water pump scheduling and pipe replacement problems. Eur. J. Oper. Res. **246**, 293–306 (2015)
86. Niknami, M., Samaranayake, S.: Tractable pathfinding for the stochastic on-time arrival problem. In: International Symposium on Experimental Algorithms, pp. 231–245. Springer, Cham (2016)
87. Paden, B., Čáp, M., Yong, S.Z., et al.: A survey of motion planning and control techniques for self-driving urban vehicles. IEEE Trans. Intell. Veh. **1**(1), 33–55 (2016)
88. Papageorgiou, M., Hadj-Salem, H., Blosseville, J.M.: Alinea: A local feedback control law for on-ramp metering. Transp. Res. Rec. **1320**(1), 58–67 (1991)
89. Patire, A.D., Wright, M., Prodhomme, B., et al.: How much gps data do we need? Transp. Res. Part C Emerg. Technol. **58**, 325–342 (2015)
90. Patriksson, M.: The Traffic Assignment Problem: Models and Methods. Courier Dover Publications, Mineola (2015)
91. Pereira, F.C., Rodrigues, F., Ben-Akiva, M.: Using data from the web to predict public transport arrivals under special events scenarios. J. Intell. Transp. Syst. **19**(3), 273–288 (2015)
92. Pinelli, F., Nair, R., Calabrese, F., et al.: Data-driven transit network design from mobile phone trajectories. IEEE Trans. Intell. Transp. Syst. **17**(6), 1724–1733 (2016). https://doi.org/10.1109/TITS.2015.2496783
93. Poonawala, H., Kolar, V., Blandin, S., et al.: Singapore in motion: insights on public transport service level through farecard and mobile data analytics. In: Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, pp. 589–598. ACM, New York (2016). https://doi.org/10.1145/2939672.2939723
94. Pozdnoukhov, A., Kaiser, C.: Space-time dynamics of topics in streaming text. In: Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-based Social Networks, pp. 1–8. ACM, New York (2011)
95. Prigogine, I., Herman, R.: Kinetic Theory of Vehicular Traffic. Elsevier, New-York (1971)
96. Purcell, M., Barry, M., Eck, B.: Using smart water meters in (near) real-time on the iWIDGET system. In: Proceedings of the 11th International Conference on Hydroinformatics (HIC2014) (2006)
97. Ran, B., Song, L., Zhang, J., et al.: Using tensor completion method to achieving better coverage of traffic state estimation from sparse floating car data. Plos One **11**(7), 1–16 (2016). https://doi.org/10.1371/journal.pone.0157420
98. Richards, P.: Shock waves on the highway. Oper. Res. **4**(1), 42–51 (1956)
99. Rinaldi, M., Himpe, W., Tampère, C.M.J.: A sensitivity-based approach for adaptive decomposition of anticipatory network traffic control. Transp. Res. Part C Emerg. Technol. **66**, 150–175 (2016)
100. Samaranayake, S., Blandin, S., Bayen, A.: A tractable class of algorithms for reliable routing in stochastic networks. Procedia. Soc. Behav. Sci. **17**, 341–363 (2011)
101. Samaranayake, S., Glaser, S., Holstius, D., et al.: Real-time estimation of pollution emissions and dispersion from highway traffic. Comput. Aided Civ. Inf. Eng. **29**(7), 546–558 (2014)
102. Sanandaji, B.M., Varaiya, P.: Compressive origin-destination estimation. Transp. Lett. **8**(3), 148–157 (2016)
103. Sävås, F.N.: Forecast comparison of models based on SARIMA and the Kalman filter for inflation. Master's thesis, Uppsala University (2013)
104. Schrank, D., Eisele, B., Lomax, T.: TTI's 2012 urban mobility report. Texas A&M Transportation Institute. The Texas A&M University System **4** (2012)
105. Shen, W., Wynter, L.: Real-time road traffic fusion and prediction with GPS and fixed-sensor data. In: 2012 15th International Conference on Information Fusion, pp. 1468–1475. IEEE, Piscataway (2012)

106. Shumway, R.H., Stoffer, D.S.: Time Series Analysis and its Applications. Springer, New York (2000)
107. Smith, M.J.: The existence, uniqueness and stability of traffic equilibria. Transp. Res. B Methodol. **13**(4), 295–304 (1979)
108. Smith, M.J., Liu, R., Mounce, R.: Traffic control and route choice; Capacity maximization and stability. Transp. Res. Proc. **7**(1), 556–577 (2015). https://doi.org/10.1016/j.trpro.2015.06.029
109. Song, B., Wynter, L.: Real-time public transport service-level monitoring using passive WiFi: a spectral clustering approach for train timetable estimation. Preprint, arXiv:1703.00759 (2017)
110. Sterling, M., Coulbeck, B.: A dynamic programming solution to optimization of pumping costs. Proc. Inst. Civil Eng. **59**(2), 813–818 (1975)
111. Sterling, M., Coulbeck, B.: Optimisation of water pumping costs by hierarchical methods. Proc. Inst. Civil Eng. **59**(2), 787–797 (1975)
112. Sun, X., Munoz, L., Horowitz, R.: Highway traffic state estimation using improved mixture Kalman filters for effective ramp metering control. In: Proc. of the 42nd IEEE Conference on Decision and Control, vol. 6. IEEE, Piscataway (2003)
113. Tang, X., Blandin, S., Wynter, L.: A fast decomposition approach for traffic control. IFAC Proc. Vol. **47**(3), 5109–5114 (2014)
114. Tassiulas, L., Ephremides, A.: Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. IEEE Trans. Autom. Control **37**(12), 1936–1948 (1992)
115. Thiagarajan, A., Ravindranath, L., LaCurts, K., et al.: VTrack: Accurate, energy-aware road traffic delay estimation using mobile phones. In: Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems, pp. 85–98. ACM, New York (2009)
116. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Methodol. **58**(1), 267–288 (1996)
117. U.S. Environmental Protection Agency: Energy efficiency for water and wastewater utilities. Technical report, U.S. Environmental Protection Agency (2015). URL http://water.epa.-gov/infrastructure/sustain/energyefficiency.cfm. Accessed 11 Mar 2015
118. Vaccari, A., Liu, L., Biderman, A., et al.: A holistic framework for the study of urban traces and the profiling of urban processes and dynamics. In: 2009 12th International IEEE Conference on Intelligent Transportation Systems (ITSC), pp. 1–6. IEEE, Piscataway (2009)
119. van Zyl, J.E., Savic, D., Walters, G.: Operational optimization of water distribution systems using a hybrid genetic algorithm. J. Water Resour. Plan. Manag. **130**(2), 160–170 (2004)
120. Varaiya, P.: Reducing highway congestion: an empirical approach. Eur. J. Control. **11**(4–5), 301–309 (2005)
121. Varaiya, P.: A universal feedback control policy for arbitrary networks of signalized intersections. Published online (2009). URL http://paleale.eecs.berkeley.edu/~varaiya/papers$\delimiter"026E30F$\ps.dir/090801-Intersectionsv5.pdf
122. Wang, Y., Boyd, S.: Fast model predictive control using online optimization. IEEE Trans. Control Syst. Technol. **18**(2), 267–278 (2010)
123. Waraich, R.A., Galus, M.D., Dobler, C., et al.: Plug-in hybrid electric vehicles and smart grids: investigations based on a microsimulation. Transp. Res. Part C Emerg. Technol. **28**, 74–86 (2013)
124. Wongpiromsarn, T., Uthaicharoenpong, T., Wang, Y., et al.: Distributed traffic signal control for maximum network throughput. In: 15th International IEEE Conference on Intelligent Transportation Systems (ITSC), pp. 588–595. IEEE, Piscataway (2012)
125. Work, D., Blandin, S., Tossavainen, O.P., et al.: A traffic model for velocity data assimilation. Appl. Math. Res. eXpress **1**, 1–35 (2010)
126. Wright, M., Horowitz, R.: Fusing loop and GPS probe measurements to estimate freeway density. IEEE Trans. Intell. Transp. Syst. **17**(12), 3577–3590 (2016)

127. Xue, M., Wu, H., Chen, W., et al.: Identifying tourists from public transport commuters. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1779–1788. ACM, New York (2014)
128. Yang, H., Wang, X.: Managing network mobility with tradable credits. Transp. Res. B Methodol. **45**(3), 580–594 (2011)
129. Yilmaz, M., Krein, P.T.: Review of battery charger topologies, charging power levels, and infrastructure for plug-in electric and hybrid vehicles. IEEE Trans. Power Electron. **28**(5), 2151–2169 (2013)
130. Zeiler, M.D.: ADADELTA: An adaptive learning rate method. Preprint, arXiv:1212.5701 (2012)
131. Zheng, N., Waraich, R.A., Axhausen, K.W., et al.: A dynamic cordon pricing scheme combining the macroscopic fundamental diagram and an agent-based traffic model. Transp. Res. A Policy Pract. **46**(8), 1291–1303 (2012)
132. Zheng, Y., Capra, L., Wolfson, O., et al.: Urban computing: concepts, methodologies, and applications. ACM Trans. Intell. Syst. Technol. **5**(3), 38 (2014)
133. Zheng, Z., Su, D.: Traffic state estimation through compressed sensing and Markov random field. Transp. Res. B Methodol. **91**, 525–554 (2016). http://dx.doi.org/10.1016/j.trb.2016.06.009
134. Zhou, P., Zheng, Y., Li, M.: How long to wait?: Predicting bus arrival time with mobile phone based participatory sensing. In: Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services, pp. 379–392. ACM, New York (2012)
135. Zhou, Z., De Schutter, B., Lin, S., et al.: Two-level hierarchical model-based predictive control for large-scale urban traffic networks. IEEE Trans. Control Syst. Technol. **25**(2), 496–508 (2017). https://doi.org/10.1109/TCST.2016.2572169

Index

H. G. Kaper, F. S. Roberts (eds.), *Mathematics of Planet Earth*, Mathematics of Planet Earth 5, https://doi.org/10.1007/978-3-030-22044-0