CLIMATE MODELLING

Philosophical and Conceptual Issues

Edited by
Elisabeth A. Lloyd
and Eric Winsberg

Climate Modelling

Elisabeth A. Lloyd • Eric Winsberg
Editors

Climate Modelling

Philosophical and Conceptual Issues

palgrave
macmillan

Editors
Elisabeth A. Lloyd
Indiana University Bloomington
Bloomington, IN, USA

Eric Winsberg
Department of Philosophy
University of South Florida
Tampa, FL, USA

ISBN 978-3-319-65057-9 ISBN 978-3-319-65058-6 (eBook)
https://doi.org/10.1007/978-3-319-65058-6

Library of Congress Control Number: 2017963873

Cover illustration: artpartner-images.com / Alamy Stock Photo

Printed on acid-free paper

This Palgrave Macmillan imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

We have both been fascinated by models for our entire careers. Climate models are especially interesting, because they are the largest and most complex of models and also, in some sense, the most mysterious. The systems are completely filled with nonlinear equations and unpredictability, yet some climate models are valued for their predictive capacities. Others are appreciated for their abilities to represent causal forces within climate systems and their interactions, and yet others represent those systems simply, elegantly, and yet powerfully.

There are numerous philosophical questions involving representation, grounding, and reality itself that arise when using climate models, as well as conceptual issues concerning the models as tools themselves. Yet there is no book or collection available that addresses these issues. We have aimed to collect a set of essays here that discusses these and other philosophical and conceptual questions about climate models. We asked some of the best philosophers and some of the best modelers to contribute to the book, and they agreed, to our delight.

Our book is intended to be enjoyed by policy-makers, climate scientists, and philosophers alike, as well as the general public. Some essays, such as those concerning policy and robustness, in parts 2 and 3 of the book, are very accessible. There are sections of part 1 that are more technical, such as the Santer et al. paper, but that is explained in Lloyd's essay and in Santer et al.'s "Fact Sheet" in part 1.

Sadly, there is rampant disinformation circulating about climate models today, despite concerted efforts by climate scientists to correct the public record. The essays contributed to this book provide a foundation for an informed discourse concerning climate models, one based on theory, facts, and evidence.

We have both learned a great deal about climate modeling through editing this collection, and our hope is that anyone dipping into the book will experience the same benefit. Of course, modeling is an ongoing activity, and many of the facets explored in this book will continue to fascinate both modelers, philosophers, and policy analysts for some time to come.

Bloomington, IN, USA — Elisabeth A. Lloyd
Tampa, FL, USA — Eric Winsberg
June 2017

Acknowledgments

As usual for a book of this size, many people were involved in the creation of it, and we are able to thank just a fraction of those, here. We would start by thanking Linda Mearns, Jeffrey Kiehl, and Doug Nychka for making Lisa Lloyd's (EAL's) visits to the National Center for Atmospheric Research (NCAR) possible over the years. They and many climate scientists, including Caspar Amman, Melissa Bukovsky, Jim Hurrell, Brian O'Neill, Claudia Tebaldi, Kevin Trenberth, Tom Wigley, and others too numerous to name, introduced me (EAL) to the fundamentals of climate science and climate modeling and also introduced me to many more scientists who would help Lisa along my journey. Being an Affiliate Scientist at NCAR has also helped me meet many scientists from around the world who contributed enormously to her learning and to this book, such as Reto Knutti, Ricky Rood, Jonathan Rougier, Gabriel Hegerl, and her co-author Vanessa Schweizer, among many others. Her co-organization of a running session at the American Geophysical Union (AGU) allowed the opportunity to meet yet more climate scientists, such as Michael Mann, a key figure in understanding climate. She would also like to thank Ben Santer, to whom a debt is also owed for help, patience, and heroism in the face of adversity.

During my many years of research into the philosophy and foundations of climate modeling, Lisa was supported financially by two sources, my endowed chair and the National Science Foundation (NSF). The Arnold

and Maxine Tanis Chair of History and Philosophy of Science made my annual trips to NCAR possible, as well as the annual trips to the AGU. Lisa has had the privilege of knowing Bud and Maxine Tanis, and they are some of the finest and most lovely people She has met in my entire life. Lisa was also funded through two NSF Scholar Grants, "A case of objectivity in science: Climate change" (2007, #0646253) and "What is 'Value Added' in Regional Climate Modeling?" (2016–2017, #1632202). These grants helped make it possible for me to visit NCAR in Boulder for longer visits and to attend workshops and the AGU during those years. Lisa is indebted to Fred Kronz and the NSF for their support.

Finally, Lisa would also like to thank her research assistants, Chris ChoGlueck, Daniel Lindquist, and, most gratefully, Ryan Ketcham, for their patience and help over the several years that it took to get this book produced. She would also note that she owes much happiness and accomplishment to her beloved husband and partner, Teddy Alfrey. All of these people aided in overcoming the delaying effects of a car accident and spinal surgery on the production of this book. Lisa owes them a great deal indeed.

Eric Winsberg would like to thank the Institute of Advanced Study at Durham University, where he had the opportunity to learn about climate science from many of the practitioners affiliated with the university and to make climate science a focus of his philosophical study, and the Institute of Advanced Study on the Media Cultures of Computer Simulation at Leuphana University, which supported much of the work on this book. He would like to thank Jessica Williams for the support she gives him in all his endeavors.

Contents

Notes on Contributors

Melissa S. Bukovsky is a Project Scientist at the National Center for Atmospheric Research. Her cross-disciplinary work centers on regional climate change in North America and includes both climate modeling and data analysis. Her specialties include extreme weather and storms, climate changes and impacts, mesoscale meteorology, and climate modeling. She works with people outside the atmospheric sciences involved in studying the impacts of weather and climate on society. She has been an integral part of the North American Regional Climate Change Assessment Program (NARCCAP).

Nancy Cartwright is a Professor of Philosophy at the Department of Philosophy, University of Durham and at the University of California, San Diego (UCSD). She is past President of the Philosophy of Science Association and was President of the American Philosophical Association (Pacific Division) in 2008. Her research interests include philosophy and history of science (especially physics and economics), causal inference, causal powers, scientific emergence and objectivity, evidence, especially for evidence-based policy [EBP] and the philosophy of social technology. She has authored a number of books, the most recent being *Improving Child Safety: Deliberation, judgement and empirical research*, with Eileen Munro, Jeremy Hardie, and Eleonora Montuschi.

Michel Crucifix is a Professor at the Université de Louvain and Senior Research Scientist at the Belgian National Fund of Scientific Research. His research group focuses on the dynamics of current and past climates with a range of methods including dynamical systems analysis, numerical simulation, and statistical

inference. He is an Editor of the journal *Earth System Dynamics* and a member of several European societies, including the Royal Academy of Sciences in Belgium.

Melissa Free worked at the Air Resources Laboratory at the National Oceanographic and Atmospheric Administration, Silver Spring, Maryland.

Mathias Frisch is a Professor for Philosophy at the Leibniz Universität Hannover, in Germany. He has held positions at both Northwestern University and the University of Maryland, where he taught until 2016. His research focuses on general philosophy of science, philosophy of physics, and philosophy and climate change. He has written two books: *Inconsistency, Asymmetry, and Non-Locality: A Philosophical Investigation of Classical Electrodynamics* (Oxford 2005) and *Causal Reasoning in Physics* (Cambridge 2014).

Peter J. Gleckler is a Research Scientist at the Program for Climate Model Diagnosis and Intercomparison at Lawrence Livermore National Laboratory, CA. His research involves the analysis, comparison, and evaluation of climate models. Working with National Oceanic and Atmospheric Administration (NOAA), he also applies climate models to study current problems in understanding the dynamics of climate, such as ocean warming and climate change. He has contributed to several Assessment Reports of the Intergovernmental Panel on Climate Change and continues to publish widely on model assessment and evaluation.

Leopold Haimberger is an Associate Professor at the Institute for Meteorology and Geophysics, University of Vienna. His research interests include diagnostics of atmospheric general circulation, analysis of radiosonde data, and numerical weather prediction. He served as a Contributing Author for the UN's Intergovernmental Panel on Climate Change 5th Assessment Report (2013), and he is a regular contributor to the Bulletin of the American Meteorological Society "State of the Climate" supplement.

Philip D. Jones is a Research Professor (and up to 2016 was the Research Director) of the Climatic Research Unit (CRU) and is now and a Professor in the School of Environmental Sciences at the University of East Anglia in Norwich. He is principally known for the time series of hemispheric and global surface temperatures, which he updates on a monthly basis. His other fields include climate change, detection and attribution of climate, proxy climate reconstructions, and climate extremes and impacts. He has produced over 450 research papers over the course of his career and is one of the most widely cited

climate scientists publishing today. He is a Fellow of the Royal Meteorological Society and is an elected member of Academic Europaea, as well as being on the editorial board of *Climatic Change*. He has won numerous awards and prizes and is also a Fellow of the American Meteorological Society and the American Geophysical Union (AGU).

Thomas R. Karl is the former director of the National Oceanic and Atmospheric Administration's National Centers for Environmental Information (NCEI), retiring in August, 2016. He was the Lead Author on several Assessment Reports for the USA, including the 2006 study on tropospheric temperatures (CCSP), and the 2015 study published in the American Association for the Advancement of Science's *Science* journal, concerning the possible existence of a hiatus in global warming, of which he found no evidence, a finding that was later independently confirmed. He was the President and Fellow of the American Meteorological Society, and Chair of the U.S. Global Change Research Program (USGCRP) Subcommittee on Global Change Research. He has served as a Lead Author on several Assessment Reports of the Intergovernmental Panel on Climate Change, and is a fellow of the American Geophysical Union, and a National Associate of the National Research Council.

Stephen A. Klein is a Research Scientist in the Cloud Processes Research Group at the Department of Energy's Lawrence Livermore National Laboratory (DOE LLNL) in Livermore, California. His research interests include clouds, their role in climate change, and the fidelity with which climate models simulate clouds. He is a lead author or co-author on over 100 peer-reviewed publications. Prior to arriving at LLNL in 2004, he was a research scientist at the National Oceanic and Atmospheric Administration's Geophysical Fluid Dynamics Laboratory (NOAA GFDL), a leading climate modeling laboratory in Princeton, New Jersey. While there, he was a leader in the creation of the atmospheric portion of the GFDL CM2 climate model. Most recently, he has been leading a multipronged effort to determine the response of clouds to climate change.

Reto Knutti is a Professor in the Department of Environmental Systems Science at the Institute for Atmospheric and Climate Science at ETH Zurich, Switzerland. He has been a leading researcher into climate models and their foundations, publishing key papers concerning their structure, genealogy, and limits. He has also been a pivotal member of the Intergovernmental Panel on Climate Change for a number of years, leading Fifth Assessment WG1 Model projections. His research topics currently span the following fields: long-term projections, scenarios, climate targets, climate and carbon cycle feedbacks,

uncertainties, model evaluation, model weighting, natural climate variability, detection and attribution, climate sensitivity, ocean heat uptake, extreme events, regional projections, climate services, and more.

John R. Lanzante is a Research Meteorologist at the Climate Impacts and Extremes Group at the Geophysical Fluid Dynamics Laboratory (GFDL)/ NOAA, Princeton University Forrestal Campus, Princeton, NJ. His research involves model-generated and observed data, focusing on large-scale climate diagnostics, on time scales ranging from days to decades. He is an expert in handling and analyzing weather balloon datasets and their preparation and maintenance. Much of his work involves Empirical Statistical Downscaling of climate model output and its critical evaluation. He is Associate Editor of *Journal of Climate* and has served as Contributing Author to several Assessment Reports, including the UN's Intergovernmental Panel on Climate Change Fourth Assessment.

Elisabeth A. Lloyd is a philosopher of climate science and evolutionary biology, as well as a scientist studying women's sexuality. She is Arnold and Maxine Tanis Chair of History and Philosophy of Science; Adjunct Professor in the Department of Biology; Adjunct Professor in the Department of Philosophy; Affiliated Faculty Scholar at The Kinsey Institute for Research in Sex, Gender, and Reproduction; Adjunct Faculty in the Center for the Integrative Study of Animal Behavior; and Faculty in the Cognitive Science Program. She works in the Department of History and Philosophy of Science and Medicine, Indiana University (Bloomington). She was previously Professor of Philosophy at University of California, Berkeley. She has published *The Structure and Confirmation of Evolutionary Theory* (Princeton) and *Science, Politics, and Evolution* (Cambridge). Her fourth book, *The Case of the Female Orgasm* (Harvard), won awards in both philosophy and science.

Victor Magana is a Research Associate at the Institute of Geography at the University of Mexico specializing in the investigation of climate dynamics in the American Tropics, climate change, and the evaluation of the role governmental policy has played in wetland management, drought, and urban development.

Michael Mann is a Distinguished Professor of Meteorology and Director of the Earth System Science Center at The Pennsylvania State University, University Park, PA. One of the most noted climate scientists in the world starting with his creation with R.S. Bradley and M.K. Hughes of the so-called hockey-stick graph of global warming, he is the author of *The Hockey Stick and the Climate Wars:*

Dispatches from the Front Lines, co-author (w/Lee Kump) of *Dire Predictions: Understanding Climate Change* and *The Madhouse Effect*, with Tom Toles. He is a Fellow of the American Meteorological Society and the recipient of numerous prestigious national and international awards.

Alexandre Marcellesi studied philosophy of social science at UC San Diego with Nancy Cartwright, receiving his PhD for a dissertation on causation and evidence-based policy in 2016. He is currently enrolled in the NYU School of Law.

Linda O. Mearns is Director of the Weather and Climate Impacts Assessment Science Program (WCIASP) and Head of the Regional Integrated Sciences Collective (RISC) within the Institute for Mathematics Applied to Geosciences (IMAGe) and Senior Scientist at the National Center for Atmospheric Research, Boulder, Colorado. She has performed research and published mainly in the areas of climate change scenario formation, quantifying uncertainties, and climate change impacts on agro-ecosystems. She has particularly worked extensively with regional climate models and co-Chairs the North American CORDEX program (NA-CORDEX), which is providing multiple high-resolution climate change scenarios for the North American climate science and impacts community. She has been a Contributing or Lead Author in the Intergovernmental Panel on Climate Change (IPCC) 1995, 2001, 2007, and 2013 Assessment Reports.

Carl Mears is Vice President and Senior Research Scientist at Remote Sensing Systems in Santa Rosa, CA. His research is focused on the construction and maintenance of climate-quality data records, including those atmospheric temperatures from satellite datasets, MSU and AMSU, and total column water vapor from SSM/I, AMSRE, SSMIS, and WindSat. He has also studied the use of these datasets for the detection and attribution of human-induced climate change. He was a convening lead author for the US Climate Change Science Program Synthesis and Assessment product 1.1, a lead author for the US Climate Science Special Report, and a contributing author to the Intergovernmental Panel on Climate Change 4th and 5th Assessment Reports.

Doug Nychka is Director of the Institute for Mathematics Applied to Geosciences (IMAGe) at the National Center for Atmospheric Research (NCAR) in Boulder, CO. He has a wide array of interests in data science, including nonparametric regression, detection and properties of nonlinear systems, and spatial statistics applied to large datasets. He has used Bayesian hierarchical

modeling for paleoclimate reconstructions and other applications of statistics to climate observations and model experiments. He is a Fellow of the American Statistical Association and was awarded the Jerry Sacks Award for Multidisciplinary Research.

Jay Odenbaugh is an Associate Professor and Department Chair at the Department of Philosophy of Lewis and Clark College in Portland, Oregon. He specializes in the history and philosophy of science, especially of climate science, evolutionary biology, and ecology, as well as ethics and metaphysics. His work on complex and simple models in ecology has been especially influential.

Naomi Oreskes is a Professor of the History of Science and Affiliated Professor of Earth and Planetary Sciences at Harvard University. She previously served as Professor of History and Science Studies at the University of California, San Diego, and Adjunct Professor of Geosciences at the Scripps Institute of Oceanography. She works on conceptual and historical issues in the Earth and Environmental sciences, as well as science policy, philosophy of science, science and religion, STS, technology and society, and women and gender studies. Her book with Erik M. Conway, *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco to Global Warming*, won the Watson-Davis Prize from the History of Science Society. Her 2004 essay "The Scientific Consensus on Climate Change" (*Science*) has been widely cited, both in the United States and abroad.

Wendy Parker is an Associate Professor in Philosophy and Associate Director of the Center for Humanities Engaging Science and Society at Durham University, UK. Her research concerns the epistemology and methodology of contemporary science, with a special focus on computer simulation models and how they are evaluated and used. She has published numerous papers on issues in climate science and climate modeling.

Sara C. Pryor is a Professor of Earth and Atmospheric Sciences at Cornell University in Ithaca, NY. She previously held the position of Provost's Professor at Indiana University, Bloomington. Her climate science research encompasses both numerical and statistical methods and focuses primarily on mitigation options from, and adaptation for, the energy sector. She is a Fellow of the AAAS and served as Convening Lead Author (Midwest Region) and a member of the Advisory committee for the National Climate Assessment (2011); and as an Editor of *Journal of Geophysical Research-Atmospheres* 2010–14.

Jonathan Rougier is a Professor of Statistical Science at University of Bristol, UK. His research concerns uncertainty and risk assessment in complex systems, particularly for natural hazards. He is co-editor and contributor of *Risk and Uncertainty Assessment for Natural Hazards* (2013). He has published widely in a range of fields including statistics and probability, and economics and finance, as well as applications of statistics to problems and issues in climate science and modeling.

Benjamin D. Santer is an atmospheric scientist at Lawrence Livermore National Laboratory (LLNL), where he works in the Program for Climate Model Diagnosis and Intercomparison. His research focuses on topics such as climate model evaluation, the use of statistical methods in climate science, and identification of natural and anthropogenic "fingerprints" in observed climate records. His awards include a MacArthur Fellowship (1998), the US Department of Energy's E.O. Lawrence Award (2002), a Distinguished Scientist Fellowship from the US Department of Energy (2005), a Fellowship of the American Geophysical Union (2011), and membership in the US National Academy of Sciences.

Gavin A. Schmidt is the Director of the Goddard Institute for Space Studies (GISS) at the National Aeronautics and Space Administration (NASA), and the Principal Investigator for the GISS ModelE Earth System Model. This model was used for the GISS modeling contribution to the CMIP3 and CMIP5 databases, which have been widely used by the IPCC 4th and 5th Assessment Reports (AR4/AR5). He is interested in ways in which model skill can be evaluated over the instrumental period and in paleoclimate records, with a focus on periods that might provide key constraints on the system and how measures of skill in representing past climate changes can be directly used to inform future projections. He recently gave the Steven Schneider Lecture at the American Geophysical Union (AGU) and was awarded the Inaugural AGU Climate Communication Prize.

Steven C. Sherwood is an ARC Laureate Professor in Physical Meteorology and Atmospheric Climate Dynamics at the Climate Change Research Centre of the University of New South Wales, Sydney, Australia, where he was Director from 2012 to 2016. He formerly worked at Yale University and the Goddard Space Flight Center in the US. He leads a research group that applies basic physics and mathematics to complex problems by a combination of simple theoretical ideas and hypotheses and directed analyses of observations. They use advanced statistical techniques and climate models to study various processes in

the climate. He is the recipient of the Clarence Leroy Meisinger award from the American Meteorological Society (2005) and a CAREER award, National Science Foundation. He is currently an Editor at *Environmental Research Letters* and has served as a Contributing and Lead Author for the Fourth and Fifth Assessments of the Intergovernmental Panel on Climate Change.

Susan Solomon is the Ellen Swallow Richards Professor of Atmospheric Chemistry and Climate Science at Massachusetts Institute of Technology (MIT). She formerly worked at the National Oceanic and Atmospheric Administration. She was the first, with her colleagues, to propose the chlorofluorocarbon free radical reaction mechanism that is the cause of the Antarctic ozone hole. She is a member of the US National Academy of Sciences, the European Academy of Sciences, and the French Academy of Sciences and holds numerous honorary doctorates. She is the author of *Aeronomy of the Middle Atmosphere: Chemistry and Physics of the Stratosphere and Mesosphere, 3rd Ed.* (2005). She served as Contributing Author for the UN's Intergovernmental Panel on Climate Change and later was also co-chair of Working Group I of the Fourth Assessment Report.

Karl E. Taylor is a research scientist at the Program for Climate Model Diagnosis and Intercomparison (PCMDI) at Lawrence Livermore Laboratory, California. His focus is theoretical studies of climate and atmospheric circulation including climate modeling, climate change, detection of climate change, paleoclimate, climate sensitivity and processes, unintended consequences of geoengineering, numerical methods, and metrics for gauging model performance. He has contributed to several Assessment Reports for the UN's Intergovernmental Panel on Climate Change as Lead and Contributing Author, as well as Review Editor. He has helped organize and coordinate a number of the international research activities, including the Paleoclimate Modeling Intercomparison Project (as a past co-chair), the CF Metadata conventions (as chair of its Governance Panel), and the World Climate Research Programme-Working Group on Coupled Model (WCRP-WGCM) Infrastructure Panel (as co-chair).

Peter W. Thorne is a Professor of climatology and physical geography in the Department of Geography at the National University of Ireland, Maynooth (NUIM). He is the chair of the International Surface Temperature Initiative, an interdisciplinary effort to create improved land surface air temperature products. He is co-chair of the GCOS Working Group on the Global Climate Observing System Reference Upper Air Network (GRUAN) and is also the project lead on the Horizon 2020 GAIA-CLIM project which aims to use such measurements

to better characterize satellite measurements. He is a Lead Author of the UN's Intergovernmental Panel on Climate Change Fifth Assessment Report.

Frank J. Wentz is the President and CEO of Remote Sensing Systems, a research company specializing in the production of measurement technologies relating to satellite microwave remote sensing of the Earth. His research has focused on radiative transfer models that relate satellite observations to geophysical parameters, with the objective of providing reliable geophysical datasets to the Earth Science Community. Wentz has served on numerous NASA review panels and is a fellow of the American Association for the Advancement of Science, the American Meteorological Society, and the American Geophysical Union. He has received numerous awards, including the Verner E Suomi Award in 2015 for "pioneering, painstaking work to accurately retrieve geophysical parameters from satellite microwave instruments and using these measurements to elucidate climate trends."

Tom M.L. Wigley is an Adjunct Professor in the School of Biological Sciences at the University of Adelaide, Australia, and a former Director of the Climatic Research Unit at the University of East Anglia, UK. He also remains affiliated with the University Corporation for Atmospheric Research; he worked for many years at the National Center for Atmospheric Research in Boulder, CO. He was named a fellow of the American Association for the Advancement of Science (AAAS) for his major contributions to climate and carbon cycle modeling and to climate data analysis and remains one of the world's experts on climate change. He contributed to many of the reports of the UN's Intergovernmental Panel on Climate Change.

Eric Winsberg is a philosopher of science specializing in modeling and simulation, climate science, and the philosophy of physics. He is a Professor at the Department of Philosophy, University of South Florida. He is author of the books *Science in the Age of Computer Simulation* and *Philosophy and Climate Science*. He is also a co-editor of *Time's Arrows and the Probability Structure of the World*.

List of Figures

List of Tables

1

Introduction

Elisabeth A. Lloyd and Eric Winsberg

1.1 A Warming Planet

As we write this in the early summer of 2016, we see news stories reporting that April 2016 was the hottest month of April in the historical record. In fact, the last 12 consecutive months have set global high temperature records. All but one of the ten hottest years going back to 1880 have come in the twenty-first century, with the one exception being 1998. 2015 was the hottest year on record, having broken the previous record (2014) by the largest margin yet, but 2016 looks likely to break both of those records (it will be the hottest year ever, and it will exceed 2015 by an even larger margin than 2015 exceeded 2014). Meteorologists are now predicting that 2016 will surpass the 1.5 °C mark, meaning that

E.A. Lloyd (✉)
Indiana University Bloomington, Bloomington, IN, USA

E. Winsberg
Department of Philosophy, University of South Florida, Tampa, FL, USA

E.A. Lloyd, E. Winsberg (eds.), *Climate Modelling*,
https://doi.org/10.1007/978-3-319-65058-6_1

it will be more than 1.5 °C higher than the pre-industrial average. 2.0 °C has long been considered a dangerous tipping point beyond which we dare not pass. It is now looking more and more unavoidable.

Every year, usually in February or March, the cap of frozen seawater floating over the North Pole in the Arctic Ocean reaches its largest size for the year before its starts to melt back for the summer. The peak in 2016 was reached on 24 March at 5.607 million square miles. That is the smallest size, in the satellite record going back to 1978, to which the Arctic cap has reached; the 13 smallest years have been the last 13 years. This is an especially worrying development, because the melting of ice is an extremely strong feedback effect in the climate system: as the temperature rises, ice melts and the melting ice reduces the amount of sunlight reflected back into space, which makes the temperature rise even more. Other potential tipping points loom on the not-so-distant horizon: the melting of the Arctic permafrost, which would release billions of tons more carbon into the atmosphere; the melting of the Thwaites glacier in Antarctica, which could destabilize enough of the Antarctic ice sheets to drive sea levels up by 16 feet; and the spread of diseases into areas where they have never been before—with dengue fever, for example, now being a significant risk in areas beyond both tropics for the first time in history.

While a fair bit of controversy concerning the cause of these phenomena remains in the body politic (especially in the United States),[1] nothing could be further from the truth when it comes to the scientific community. Multiple studies, appearing in peer-reviewed publications, all show similar findings: that roughly 97–98% of actively publishing climate scientists agree with the claim that it is extremely likely that the past century's warming trend is due to human activities.[2] Eighteen major scientific associations (including the American Association for the Advancement of Science and the American Geophysical Union) have endorsed the claim that "Observations throughout the world make it clear that climate change is occurring, and rigorous scientific research demonstrates that the greenhouse gases emitted by human activities are the primary driver."[3]

Part of this confidence comes from the fact that the scientific basis for the claim of anthropogenesis (caused by human activities) rests on a wide variety of convergent evidence: the recordings of modern instruments concerning the climate going back to around 1880; observations of sea

ice, glaciers, ice sheets, animal migrations, etc.; basic science in the form of energy-balance models; reconstructions of more distant climate history from "proxy data" like ice cores, tree rings, pollen samples, coral reefs, and the like; and of course the detailed study of highly complex and sophisticated computer simulation models of the climate. The same can be said about our confidence in the rather general claim that further increases in greenhouse gas concentrations are going to drive the climate further away from its pre-industrial state. That too is supported by a diverse array of evidence.

But the answers to other important questions about the climate, and its response to increases in the concentrations of greenhouse gases in the atmosphere, remain less certain: what is the correct value of the earth's equilibrium climate sensitivity (the amount that a sustained doubling in the quantity of CO_2 in the atmosphere would raise the equilibrium global surface temperature)? What about the transient features of this response? How long does it take to reach equilibrium? What can we expect from global surface temperature in the meantime?

All these involve hypotheses about the future of a very coarse-grained variable: mean global surface temperature. We would also like to know quite a bit more about how these phenomena will play out regionally. Climate change is likely to make some regions wetter and other regions drier. But which ones, exactly? So far, global warming has been (as the models mostly predicted) concentrated around the poles. Will this continue? At what rate is the Arctic sea ice going to continue to disappear? (So far, it has disappeared faster than most models predicted.) Will the melting of the Arctic ice actually make northern Europe considerably colder? And perhaps most importantly, how likely are, and how close are we to, the kinds of climate *tipping points* we mentioned above: the collapse of ice sheets in Antarctica and Greenland; the cessation of the vitally important thermohaline circulation system [ocean currents driven by surface heat and freshwater flows or fluxes], or the release of massive quantities of heat-trapping gases from frozen storehouses like the Arctic permafrost?

Answers to some of these latter questions are more difficult to come by, in part because they necessarily depend on less diverse sources of evidence than the basic claim of anthropogenesis. For answering most questions

about the expected future pace and tempo of climate change that would come in response to possible emissions scenarios, we are almost wholly dependent on *complex simulation models*.

The core behavior of the atmosphere can be modeled with three simple laws: Newton's laws of motion as they apply to parcels of fluid, the law of conservation of mass, and a simple thermodynamic equation that allows us to calculate the heating effect on each parcel of air via a parameterized value of the radiation from the sun. Unfortunately, what we get out of this is a coupled set of nonlinear partial differential equations for which we have no closed form solution. We can at best hope to get a *numerical approximation* of how a system governed by such equations should behave. Simulation models of the climate do this by transforming the original (continuous) differential equations into discrete difference equations that approximate them, and use a computer to solve the latter step-by-step over discrete intervals of time for discrete points in space. Rather than a function that tells us values for variables like temperature and pressure for arbitrary points in time and space, the computer outputs numerical values for these variables on a space–time grid.

Modern climate models of the most advanced kind do much more than model just the circulation of the atmosphere. The atmosphere, after all, is only one part of the climate system—which consists not only of the atmosphere, but also the hydrosphere (seas), the cryosphere (ice sheets), the land surfaces, and the biosphere, and all the complex interactions between them. Not only does a climate model need to couple the circulation of the atmosphere to the circulation of the oceans, but the atmospheric component must also include representations of physical features like clouds, precipitation, and aerosols; the ocean component must include sea ice dynamics, iceberg transport of fresh water, currents, and wave dynamics; the land component will include precipitation and evaporation, streams, lakes, rivers, etc.; and the ice sheet component will include thickening and thinning and cracks and fissures.[4] A full Earth System Model (ESM) also tracks sources and sinks of carbon into and out of the biosphere and other systems.

All of this makes a good understanding of the conceptual and philosophical foundations of these models vital. It is vital if we are going to be

able to form well-informed judgments not only about what to expect in the future, but also about how we should act—both to mitigate those effects that we possibly can but also to adapt to those that might, at this point in time, be unavoidable.

Unfortunately, despite the fact that computer simulation modeling has played a prominent and ever-growing role in science since the middle of the last century, and despite the fact that it plays a starring role in one of the most socially important sets of scientific questions we have ever faced, it has received, until very recently, only a smattering of interest from philosophers of science. The first goal of this book is to improve on that situation.

The second goal is to explore the philosophical foundations of the other sources of knowledge in climate science. The central component of this goal is to get a better understanding of the relations between models of the climate system and the data that inform them. Data in climate science come from a wide variety of sources and instruments, all of which have strengths and weaknesses. The task of knitting all of those sources together into the most well-informed and responsible representation of the knowledge that is best supported by those sources is highly complex. That, in turn, makes it ripe for philosophical and foundational analysis. In the case of climate science, this kind of analysis by philosophers and foundationally inclined scientists is equally overdue.

In response to these lacunae, we offer this collection of essays by both climate scientists and philosophers writing on a broad array of issues pertaining to climate science and modeling. It is intended for both philosophical and scientific audiences. The essays range from detailed consideration of the evidence for climate models to discussions of models and values, to the robustness of models and its significance, and much more. Each part contains a mixture of pieces by both, philosophers and climate scientists, each offering unique perspectives on the topics at hand, valuable for their insight into climate-related issues and philosophical conundrums involving climate models. The book is not meant to be read from front-to-back, although the pieces in each part do benefit from being read in order. Enjoy!

1.2 Part 1: Confirmation and Evidence

Oreskes, Santer et al., Lloyd, Mann, Mearns et al.

We open Part 1 with an updated reproduction of a classic paper by Naomi Oreskes, "The Scientific Consensus on Climate Change: How Do We Know We're Not Wrong?" Oreskes was one of the first scholars to empirically document the degree of scientific consensus regarding the anthropogenic origin of observed changes in the climate. In this paper, she presents many of her findings, supplements with several others, and then offers a philosophical account of why we should take those findings to provide us with strong reason to believe in the claims. We thought this paper would provide a nice "second introduction" to all that follows.

The piece sets the agenda for the volume by answering two central sets of questions about climate science. First: What is the scientific consensus on climate change? How do we know it exists? What exactly does it assert? And second: What should we conclude from that consensus? Might not the claims, about which the overwhelming majority of climate experts agree, nevertheless be wrong? How strong, after all, is their evidence?

An important element of Oreskes' answer to both sets of questions is a distinction that is central to any discussion of climate science and its epistemology—the distinction between claims about the *existence and anthropogenic origin* of climate change in the recent past, on the one hand, and claims about the pace and mode of future changes, on the other. Oreskes concedes that there is neither consensus, nor overwhelmingly strong evidence for hypotheses about the pace and mode of future changes. What she is concerned with is claims of the first kind—claims about existence and anthropogenesis.

Regarding the consensus in favor of these claims, Oreskes has famously gathered a great deal of bibliographic evidence. She also canvasses the formal positions of a variety of scientific societies. She concludes that when we set aside claims about future pace and mode change, and focus only on existence and origin, the consensus is overwhelming.

Regarding the strength of the evidence, Oreskes provides two avenues of analysis. First, she examines the strength of the evidence for the existence of a warming over the last century and half. She calls this the "inductive" evidence for warming. Here, she cites the large body of evidence for a warming

trend, from measurement records in Europe that go back 150 years, from around the globe that go back 40 years, and from so-called "proxy data"—tree rings, ice cores, and coral reefs, that go back much further.

Regarding the origin, she looks at the evidence from the point of view of a variety of theories of confirmation, and argues that the evidence looks strong from all of those points of view. She examines the evidence from the point of view of four different takes on the nature of scientific evidence: the hypothetico-deductive model (championed by Carl Hempel, among others), Falsificationism (championed by Popper), from the point of view of the consilience of evidence, and of inference to the best explanation. She deftly shows that, looked at from all of these points of view, modest hypotheses regarding past trends and their causes look extremely well supported.

The next set of papers relates to a now famous controversy about the accuracy of satellite data and their role in supporting or undermining the predictions of climate models. In March 2011, a world expert in satellite climate data, John Christy, from University of Alabama, Hunstville, testified to a Congressional Committee that global climate models were contradicted and undermined by those data, and that global warming and the greenhouse effect were not occurring, contrary to what the models said. He placed a published paper into the Congressional Record to support these claims: Douglass et al. (2008). This paper had been thoroughly discredited in the scientific literature by other climate scientists, including those who handled satellite datasets, as well as statisticians and climate modelers (Santer et al. 2008; see below). But this apparently made no difference to Christy, who simply repeated his earlier claims.

In the fourth paper in this part, Elisabeth Lloyd uses this satellite data controversy as a case study of the clash of approaches to thinking about modeling, data, and the role of the scientist in data construction. On one side, we have John Christy, who takes weather balloon data at face value, and uses it to anchor his satellite data measures of temperature. This is despite the fact that the data handlers working with the weather balloons (radiosondes) see their instruments as highly variable, fickle in their readings according to their exposure to the sun, changing from instrument to instrument, having to be recalibrated constantly to keep up, and intended for meteorological, rather than climatological, use.

This latter view is tending toward what Lloyd calls "complex empiricism," while she calls Christy's view "direct empiricism," because he sees the radiosonde measurements as "direct" readings of the true temperature, taken from nature straightforwardly by an instrumental measurement. These measures are relatively simple compared to his satellites, which must undergo constant revision from the transition of one satellite to another. They give "radiances" not "degrees," and must always be recalibrated according to how much the satellite has fallen off course, moved in relation to the others, etc. Lloyd's "complex empiricist" account of the satellite data involves models intertwined with data, the judgments of the scientists in constructing the datasets as representations of the world, the "combined use of observations, theory, and models," and much more (Santer et al. 2005, p. 1555).

What we present here in this part starts with an introductory note by Lloyd (Chap. 3), and a "Fact Sheet" from Ben Santer and his co-authors (Chap. 4) regarding their critique of the Christy satellite data analysis and its comparison to climate models. We also present the Santer paper itself (Chap. 5), which gives the reader more than a taste of the Christy and colleagues' style of reasoning; and finally offer Lloyd's paper (Chap. 6) mapping out the 20+ year controversy over the satellite datasets and their relations with climate models, and arguing that this dispute can be better understood through the contrasts and comparisons of direct with complex empiricism.

1.3 "Satellite Data and Climate Models," by Elisabeth A. Lloyd (Chap. 3; Original for this book)

"Fact Sheet" for "Consistency of Modelled and Observed Temperature Trends in the Tropical Troposphere" by Ben Santer et al. (Chap. 4), & "Consistency of Modelled and Observed Temperature Trends in the Tropical Troposphere" by Ben Santer, Peter Thorne, Leo Haimberger, Karl Taylor, Tom Wigley, John Lanzante, Susan Solomon, Melissa Free, Peter Gleckler, Phil Jones, Tom Karl, Steve Klein, Carl Mears, Doug Nychka, Gavin Schmidt, Steve Sherwood, and Frank Wentz (2008) (Chap. 5)

We start with Elisabeth A. Lloyd's stage setting for the Santer et al. (2008a, b) Fact Sheet and paper, which emphasizes the positive public reception of denialist claims regarding a supposed mismatch between satellite data and climate models. This is followed by Ben Santer's and colleagues' "Fact Sheet" and 2008 paper. These are important because they show exactly why not to believe Christy's testimony to the Congressional Committee. Douglass et al. (2008) claimed to give a highly robust statistical test of the fit between observational datasets and models, and to show that the data undermined the models. Santer et al. (2008) demonstrated the fatal flaw in the Douglass et al. statistical test in a situation in which the answer is known a priori, through "stochastic simulation" methods. Data are generated randomly with known statistical properties, and the test is then made on these data, with certain expected results. What happened was that the Douglass et al. "robust statistical test" failed to give correct results with these stochastic simulation methods. In cases where there was no significant difference between two known data sets, the test frequently yielded the incorrect answer that there *was* a significant difference. Thus, their "robust statistical test" cannot be trusted when given real data, as it will indicate significant differences where there are really none. It is no surprise, then, that the Douglass et al. (2008) paper showed a significant difference between models and datasets, as the test was rigged to show such a result.

Despite this deep flaw in the statistics of the paper, the Douglass et al. paper had for the previous 10 months held very wide interest and was dispersed in the media, and both inside and outside the scientific community. As Ben Santer notes, "the paper received high-level attention within the U.S. Department of Energy and the National Oceanic and Atmospheric Administration."[5] The paper was highlighted by Fox News, and S. Fred Singer, a co-author of the paper, gave a news conference with the paper as its centerpiece at the National Press Club. A press release from that conference claimed that the Douglass et al. findings represented "an inconvenient truth," and proved that "Nature rules the climate: Human-produced greenhouse gases are not responsible for global warming."[6]

Santer sought the expertise of experts in climate modeling, statistical analysis, and the development of observational datasets of several sources. Santer and his colleagues decided, given the fatal statistical flaw in the Douglass et al. (2007 online; 2008 print) paper, and the paper's widespread

influence, as well as the significance of the interpretation of its results as essentially falsifying greenhouse warming effects, that they must perform a thorough examination of the paper's statistical significance testing, and statistical analyses of a wide range of available datasets and the models of their own. The resulting paper was much more than a critique of the Douglass et al. (2008) paper. It contains substantial further research, as a wide range of new datasets were used, several different statistical tests were used to establish significance, and there is a discussion of how these tests performed under controlled conditions. Christy, Douglass, and Singer have so far failed to produce a response to the critical statistical analysis offered in Santer et al. (2008), or to retract or correct their (2008) paper.

1.4 "The Role of 'Complex' Empiricism in the Debates about Satellite Data and Climate Models" by Elisabeth A. Lloyd (Updated for this book)

As noted above, Elisabeth A. Lloyd proposes two distinct approaches to relations among measurement, dataset, model, scientist, and theory, one called "direct" empiricism, which uses a basic Hypothetico-Deductive approach, the other, "complex" empiricism, meant to focus especially on model-evaluation processes, especially important as computational models become more widespread.

As she notes, in the satellite data case, "it now appears that the models were mostly right and the early data were mostly wrong, and therein lies an interesting story about data and their relations to scientists, models, and reality" (2012, p. 391). By sticking by their models and declaring that they did not trust the data, the modelers continued to insist that the data needed to be cleaned up, not so much the models (e.g., Santer et al. 1999, 2005), a move that drove satellite data wrangler John Christy to constant frustration. He was sure that the models were wrong, and believed that his data "proved" that (Christy and Spencer 2006). Christy used radiosondes as independent data against which to compare the satellite data. But Santer and colleagues did not trust the radiosondes the way

that Christy did (1999), so did not share Christy and Spencer's conclusions. Moreover, it seems that Christy et al. had used the radiosondes in building their satellite datasets, so the radiosondes could not be treated as independent datasets.

By focusing on the complex ways that models are evaluated and supported in climate science, the complex empiricists like Ben Santer and colleagues insisted that Christy and his colleague's approach to models and model confirmation was hopelessly shallow and flawed. Complex empiricism includes an approach to model evaluation and confirmation that focuses on the embeddedness of data in models, and on model assumptions and their independent empirical support. In contrast, direct empiricism relies almost completely on predictive success of models, as exemplified in the hypothetico-deductive account of theory testing. Such an emphasis on the predictions of models overshadows crucial information contained in the model assumptions and parameterizations; support for these are key, as aids to model success. Complex empiricism also emphasizes a variety of evidence for model assumptions and the conformation of multiple model results, or model robustness (Lloyd 2015). Now, the consistency of the models and satellite datasets supports the complex empiricists claims (Santer et al. 2008). This case study stands as a good example illustrating the differences between direct and complex empiricism.

1.5 "Reconciling Climate Model/Data Discrepancies: The Case of the 'Trees That Didn't Bark'" by Michael E. Mann (Original for this book)

When estimating the equilibrium climate sensitivity (ECS), a measure of our impact on climate that is defined as the warming we should expect in response to the doubling of the CO_2 concentration relative to pre-industrial levels, the results usually end up around a midrange of 3 °C. Yet when paleoclimate reconstructions of past temperature based on tree-ring proxy measurements are used for these estimates, the mid-range

number comes in a full degree lower, at 2 °C. Since these datasets are often the key to paleoclimate reconstructions, it is especially important to know whether they are biased,[7] so paleoclimatologist Michael Mann scrutinizes this result in his paper.

What is driving the ECS estimate in these comparisons of climate models to paleoclimate reconstructions? Before the industrial age, the primary forcing of climate came from natural changes in radiative forcings, from factors such as solar changes with orbital variations, or small fluctuations in greenhouse gas concentrations; but the greatest pre-industrial forcing came from volcanic eruptions and the cooling effect from aerosols spread into the stratosphere by the eruptions. If the model simulations or the paleo-reconstructions underestimate or overestimate the size of the cooling signal from the volcanic eruptions, the estimates of ECS from these comparisons will be biased.

In the paper, Mann argues that such biases do indeed exist. Thus, he says that the paleo-reconstructions "may selectively underestimate the cooling signal associated with large explosive volcanic eruptions of the past millennium" (p. 178). More specifically, underestimation of volcanic cooling can result from reliance on paleo-reconstructions from tree-ring data; that is, we lose sensitivity to large summer cooling events associated with major explosive volcanic eruptions when we rely on tree-rings. "This loss of sensitivity potentially results in chronological errors in some subset of tree-ring records used to reconstruct past temperatures" (p. 178).

In 2012, Mann and his colleagues published a new hypothesis to account for a discrepancy between the tree-ring reconstructed and climate model predicted magnitude of volcanic cooling in Northern Hemisphere mean temperatures during the pre-industrial era of the past millennium. There is a virtual absence of cooling in tree-ring reconstructions during what ice core and other evidence suggest is the largest explosive volcanic eruption of the past millennium—the 1258 AD eruption. They suspected that the discrepancy ("the trees that didn't bark") had to do with the types of tree-ring information that were being used to reconstruct past temperatures. The data scientists had used trees that had grown at the tree-line, where even a small annual temperature change could mean the difference between annual growth and no growth, that is, that in some extreme years there is no growth in the trees at all. If

no tree-ring is formed in a certain year, an error is introduced into the chronology established by counting rings back in time. The tree-ring data are then flawed and in error. Mann and his colleagues investigated this problem by comparing a tree-growth model driven with climate model simulations of the past millennium with the model-simulated temperatures and tree-ring reconstructions of temperatures. The bias was consistent with these reconstructions.

In other words, Mann argues, these findings provide additional support for the claim that the most likely value of ECS is in the range of 3 °C, and they help to explain why the previous estimates of ECS from the past millennium were low. This is a significant conclusion, since it supports the claim that the climate system is more sensitive to carbon emissions than the previous paleo-climate-based estimates indicated, and puts them more in line with estimates from other sources. It is also an episode with significant philosophical importance because it illustrates the highly complex relationship between models and data in climate science, and the ways in which carefully reached conclusions rely on constantly expanding the range and scope of evidence that needs to be weighed to get at the underlying phenomena.

1.6 "Downscaling of Climate Information" by Linda O. Mearns, Melissa S. Bukovsky, Sara C. Pryor, and Victor Magana (2014)

One of the most significant developments in climate modeling in the past decade is a surge of downscaling global climate models and global information, including the building of regional climate models. Such downscaled models are in some ways very different in approach and form from global climate models, and need to be analyzed and understood on their own terms. Downscaling in general refers to techniques or methods for developing regional or local information from coarser resolution information, usually from global climate models. Linda Mearns and colleagues discuss simple to complex methods for downscaling, from simple statistical methods to complex dynamical modeling, including variable grid global models and regional climate models.

Mearns et al. review these methods and, more importantly, draw out conclusions regarding the value of these techniques for increasing our confidence in regional projections of climate change. This is a subject rarely tackled by other authors, so is particularly valuable. Do the higher resolution models really help increase our confidence about our understanding of future climate? "Added value" is the "additional knowledge about the climate (current and future) gained from applying an RCM or other downscaling method," in comparison to a global climate model (2014, p. 235). Do the downscaled methods really "add value" compared to the global models? (See Lloyd et al. 2017, ms.)

Mearns et al. first apply these questions to "Empirical/statistical Downscaling", or "ESD", a process of making mathematical connections between states of variables representing a large spatial scale and variables representing a much smaller or local spatial scale (Mearns et al. 2014, p. 206). The authors review the uncertainties associated with this type of downscaling, and then review the results of applications of ESD to development of climate projections over North America. They found good results with precipitation and expansion of the growing season with NARCCAP (North American Regional Climate Change Assessment Program, directed by Mearns; Mearns et al. 2009). And most studies indicate a high degree of value added in both the ESD and dynamical downscaling of precipitation variables compared to the parent global model (Mearns et al. 2014; Maraun et al. 2010).

With regard to the most active area of dynamical downscaling, Mearns et al. focus on nested regional climate modeling, or RCMs, with resolutions of 10–50 kms. (In contrast to the global climate models which have 100s of km resolution.) The global source model from which variables such as temperature, moisture, wind, pressure, etc., are used is also called the "parent" or "driver" model. These variable values or "boundary conditions" are used as starting variables from which calculations are started in the RCM. Mearns et al. advise users of RCMs to pay attention to which variables are being used in the models they are applying. For example, some models use mini-models of lakes, but some do not. Some set surface temperatures of their lakes using interpolated values from the nearest ocean points; this may make the lake surface temperature more realistic than using land points, but may

negatively impact the simulated climate near the lakes. That is, more "realism" may lead to less "realism," and use of RCMs needs to take all this into account. Mearns et al. review the estimations of both skill and uncertainties of dynamically downscaled methods. They also review the sweep of results from NARCCAP, an ensemble of 50-km RCM simulations covering most of North America, intended to help climate projections for use in impacts research, such as flood, fire, and drought protections and planning, and to investigate the uncertainties in projections of future climate. Will all of this pay off? What role do these RCMs play in the hierarchy of climate models (Giorgi et al. 2016; Bukovsky et al. 2017)? There are issues of explanation, reduction, complementarity, and compatibility here, as well as other philosophical topics in play.

1.7 Part 2: Robustness and Climate Models

Parker & Odenbaugh

1.8 "The Significance of Robust Model Projections" by Wendy Parker (Updated for this book)

To begin our next segment of the book, Wendy Parker writes on the conditions in which robust predictive modeling results hold special epistemic significance, related to truth, confidence, and security. She considers whether such robust predictive modeling conditions hold in the case of climate modeling today, finding little presence of such robust climate models or their prediction. While Parker acknowledges that there is a "broad scientific consensus—underwritten by a substantial and growing body of evidence—that the earth's climate warmed significantly over the last century, that increased atmospheric concentrations of greenhouse gases due to human activities are a major cause of this warming, and that the earth's climate will be still warmer by the end of the twenty-first century," the quantitative details are less clear "especially regarding future climate change" (2011, p. 579).

How much will the earth's average surface temperature increase by the end of the twenty-first century if greenhouse gas concentrations continue rising as they have in recent decades? Can we operate an ensemble of model runs to get climate predictions? Parker explains how and why multiple models are used to investigate future climate change. She investigates a set of claims regarding inferences from robust modeling results that (i) "an agreed-on predictive hypothesis H is likely to be true"; (ii) significantly increased confidence is H is warranted; and (iii) the security of a climate to have evidence for H is enhanced. "The findings are disappointing," she writes: "When today's climate models agree that an interesting hypothesis about long-term climate change is true, it cannot be inferred—via the arguments considered here anyway—that the hypothesis is likely to be true" (2011, p. 581).

Parker considers a variety of ways that we can increase our confidence using robustness of models, including a Bayesian perspective, Condorcet's Jury Theorem, and a sampling-based perspective, all failing to arrive at a satisfactory result. In the end, she says the prospects for reaching these aims, desired for epistemic significance, "seem slim" in the near future (2011, p. 598).

1.9 "Building Trust, Removing Doubt? Robustness Analysis and Climate Modeling" by Jay Odenbaugh (Original for this book)

The second paper of this part, Jay Odenbaugh's essay, begins by examining climate models' evaluation and independence, and proceeds to discuss how model robustness can make problems with idealizations irrelevant. This topic is in contrast to Wendy Parker's writing, which concentrated on predictive modeling; Odenbaugh is focusing on robust climate modeling of past causal events, rather than predictions of the future. Thus, her conclusions are not relevant, but rather complementary, to the topics covered by Odenbaugh in this paper, which concern robust explanations of past events and their confirmatory evidence. He ends by considering a potentially seri-

ous epistemological problem with model robustness, which he addresses through epistemic contextualism and by drawing a distinction between relative and absolute robustness.

Odenbaugh sets up his discussion of climate model evaluation through the work of Elisabeth Lloyd, using her different components of (i) goodness-of-fit, (ii) independent assumptions of the models, (iii) varieties of evidence (Lloyd 1994, 2009, 2010), and (iv) model robustness (Lloyd 2015). He illustrates the goodness-of-fit through the comparison of a General Circulation Model (GCM) with Mt. Pinatubo measurements (Houghton 2009; Odenbaugh p. x), and uses a variety of other climate models for the various other confirmatory virtues.

He reviews a particular doubt raised by some authors, of issues with idealization: Taking 14 climate models across the twentieth century, which include greenhouse gas forcing, in particular, the average surface mean temperature has increased over the century 58 times over 58 trials, consistently. However, that result is fragile with respect to models that include only natural forcings, without the greenhouse gas forcing, which fail to increase in global mean surface temperature over the century. Thus, he asks, "if one was suspicious of [Temp.] because of an idealization with regard to forcings, atmospheric resolution, atmospheric layers, ocean resolution,… can one remove the doubt regarding those idealizations with robustness analysis?" It may be worrying that we get stuck in an infinite regress by a skeptic.

Odenbaugh offers a "contextualist" response to a regress worry. "Epistemological contextualists often claim that whether one knows or is justified in believing a proposition depends on what standards are at work" (p. X). (Where this is "substantive" is where it concerns whether one knows or is justified in believing a proposition with respect to varying standards.) (p. x). "When conducting robustness analysis, we must distinguish between *relative* versus *absolute* robustness analyses" (p. x). If there is concern about a specific idealized assumption, we can take it out, and replace it with another assumption we are less concerned with, provided that the skeptic is not worried about it. Odenbaugh calls this "relative" robustness analysis.

But suppose that the skeptic is worried about any idealization per se. Then the skeptic's worry would be much more profound; "The only way

to remove this worry is to show that there is some true assumption when conjoined with the substantial core implies the prediction. I will call this the "absolute" robustness analysis." As Odenbaugh cites C.S. Peirce in his piece after some nice discussion, so shall I: 'Let us not pretend to doubt in philosophy what we do not doubt in our hearts' (p. x).

1.10 Part 3: Climate Models as Guides to Policy

Knutti, Rougier & Crucifix, Winsberg, Frisch, and Marcellesi & Cartwright

1.11 "Climate Model Confirmation: From Philosophy to Predicting Climate in the Real World" by Reto Knutti (Original for this book)

Our final part considers the role and suitability of climate models for climate projection, mitigation, and policy making generally. We open the part with "Climate model confirmation: from philosophy to predicting climate in the real world," by Reto Knutti. Knutti is a physicist and climate scientist and a lead author of the IPCC's summary for policymakers. His contribution asks and answers several questions about using models to make projections: Why are models uncertain? Why do we use, and indeed need to use, more than one climate model? How are models evaluated? Are they confirmed? Does robustness and variety of evidence help us to confirm climate models?

Models are uncertain for two principal reasons. First, their structure is imperfect. They differ from reality in that they only describe some of the components and interactions that exist in the real world, and they do even that much imperfectly. Second, because the equations in those models cannot be solved analytically, they must be solved on a grid, and sometimes that grid is too large to capture important processes. These processes, in turn, need to be modeled with parameters, and the best value of these parameters is not

always known. We need more than one model both because different models are useful for different purposes, and because sampling from the predictions of more than one model enables us to get a better grip on what future scenarios are plausible given different plausible ways of modeling the climate.

How are models evaluated? Knutti argues that we need to understand the process of model evaluation not as the process of confirming the model, but as the process of gathering confidence that the model is adequate for the purposes to which we intend to put it. To do this, moreover, we need much more than evidence that the model fits observed data. This is true even if the model has multiple instances of "fit," or if it is supported by a variety of evidence. This is because instances of fit could be the result of compensating biases and instances of misfit could be the result of what philosophers call the "Duhem problem," or from errors in the data. In the end, neither fit nor misfit with observed data tell us definitely whether a model has "skill," or suitability for the purpose to which we intend to apply it—which invariably involves having confidence that the model is skillful outside of the domain of behavior for which we have real data. In the end, Knutti argues that acquiring genuine confidence that our climate models have the kinds of skill we want them to have will require us to have what he calls "process understanding."

To understand what Knutti means by process understanding, we need to understand that though climate models are driven by a mixture of basic physics, and models of other underlying components of the climate system, there are also various features and variables of the climate that are emergent. They arise out of those underlying components rather than being given in them. Process understanding comprises having knowledge of the relevant quantitative relationships and interactions between different emergent components of the system. But it also comprises having well-justified beliefs about how those relationships and interactions will or won't be preserved as we move into time periods or regimes outside of those for which we have data that we can use for evaluation. And it also involves having confidence that we have not neglected any other important relationships and interactions. Only once we have all of this, he argues, can we reliably infer from the fact that a model is adequate for predicting some domain of behavior for which we *have data*, to the claim that the model will be adequate in domains for which we don't.

1.12 "Uncertainty in Climate Science and Climate Policy" by Jonathan Rougier and Michael Crucifix (Original for this book)

Statistician Jonathan Rougier and climate scientist Michael Crucifix argue that what they call "mainstream" climate science—that is to say the kind of climate science that is practiced in universities and major climate centers—is maladapted to meeting the practical needs of policy-makers charged with making decisions about possible interventions one might make in the face of climate change (e.g., do nothing, monetize carbon, attempt geo-engineering, etc.). They identify two sets of reasons why this is so. The first includes the fact that the kinds of simulations climate scientists typically run evolved in a context in which the primary goal of climate science was *explanation*—what Rougier and Crucifix identify as the practice of confirming that the observed patterns in the climate are in fact the emergent features of basic climate physics. Relatedly, climate scientists like to be able to present their funders with highly realistic looking simulations. For both of these reasons, climate science is dominated by a kind of simulation model that has a very high resolution and is expensive to run. The result of this is that we do not have at our disposal the resources for running what they would call well-designed experiments for assessing the degree of uncertainty we ought to take ourselves to have about climate outcomes. Such a well-designed experiment would involve repeated runs of simulations under different configurations of the model parameters and modules, so that we could sample from the entire range of "not-implausible" climate system behaviors. Well-designed experiments would be in stark contrast to what we do have—"ad hoc" collections of simulator runs, like CMIP3 and CMIP5—what some others have called "samples of opportunity." They compare the former to studying a population with a carefully stratified sample of 100 people and the latter to be doing the same by simply selecting the next 100 people to walk by a particular lamppost.

The second reason that Rougier and Crucifix identify for mainstream climate science being maladapted to meeting policy-making needs is that

climate scientists are unwilling to "answer the question, own the judgement, and be coherent."

Climate scientists, they content, are not "answering the questions" that policy-makers are asking them because the models they build are too focused on "consuming CPU cycles" and are not focused on providing climate scientists with the tools they need to assess uncertainties. They are left, instead, relying on flawed intuitions. Climate scientists are not "owning the judgement," the authors argue, because the only notion of probability that makes sense for climate science is subjective probability, or what philosophers sometimes call credences or degrees of belief. They believe, however, that physical scientists are uncomfortable with the notion that the probabilities they report are "subjective"—in part, perhaps, because they confuse the Mertonian norm of disinterestedness with a notion of objectivity that is the antonym of the subjectivity of subjective probabilities. And finally, climate scientists are not always "coherent" in that they do not always strictly adhere to the rules of the probability calculus.

1.13 "Communicating Uncertainty to Policy Makers: The Ineliminable Role of Values" by Eric Winsberg (Original for this book)

Eric Winsberg, a philosopher, is concerned with the uncertainty quantification (UQ) associated with the forecasts of global and regional climate models. The advantages of UQ are clear. UQ can be an extremely effective tool for protecting the objectivity of science by dividing our intellectual labor into the epistemic and the normative. If scientists can manage to objectively assign probabilities to various outcomes given certain choices of action, then they can effectively leave decisions about the relative social value of these outcomes out of the work they do as experts. Climate scientists, for example, might tell us the probability of the arctic ice disappearing if we double carbon emissions. If they do, then the consumers of this scientific knowledge—the people or their elected leaders—can decide for themselves what value they place on the various outcomes

associated with the possible policy choices they might make, and act accordingly based on those probabilities and the usual principles of decision theory. In this way it is commonly thought that scientists can keep ethical questions—like questions about the relative value of environmental stability vs. the availability of fossil fuels for economic development—separate from the purely scientific questions about the workings of the climate system. Accepting or rejecting hypotheses about climate, on the other hand, would require climate scientists to make value judgments about the relative dangers of being wrong in each case.

Such an approach, and the attendant objectivity which comes from the division of labor that it affords between those who discover the facts and those who decide what we should value, has obvious advantages. And it is in line with a famous defense of scientific objectivity, mounted by Richard Jeffrey against the arguments of Richard Rudner in the 1950s: scientists qua scientists can avoid making value judgments by assigning probabilities to hypotheses rather than by accepting or rejecting them. These are the very considerations, or so Winsberg argues, that offer the strongest reasons for attaching precise UQs to the predictions of climate models.

This defense of the value-free ideal of science has drawn criticism from Heather Douglas, among others, who has pointed out that scientists often have to make value judgments at the lab bench; they often have to make discreet choices of methodology which themselves reflect values, and they have to do it prior to the stage where they are ready to turn over judgments to policy-makers. Winsberg points out that this is not necessarily fatal to the Jeffreyan strategy, provided that the scientists can still, as Jonathan Rougier and Michael Crucifix call for, "own the judgment" of their probability estimates, in light of the methodological choices they have made.

All of this, however, is predicated on the assumption that a conceptually coherent methodology is available for judging uncertainties based on the forecasts of complex climate models. Against this, Winsberg argues that there are features of climate science—in particular, its dependence on computer simulation models that are massively complex, constructed by experts from a wide domain of fields

of expertise, and analytically impenetrable—that make it extremely difficult for the Jeffrey strategy to succeed. In this respect, Winsberg is making a kindred point to the one made by Rougier and Crucifix, in that they both identify the complexity of climate models as an obstacle to eliciting well-considered expert judgment about climate uncertainties. Winsberg further argues that this feature of climate science, and the difficulties it creates for experts in making judgments of uncertainty, makes it nearly impossible to avoid the intrusion of normative assumptions. Worse still, it leads to those assumptions being buried in the "nooks and crannies," where they can no longer be recovered or made explicit. Consequently, some of the usual strategies proposed in the "science and values" literature for dealing with difficulties in attaining the value-free ideal of science might not work in climate science or in other disciplines that rely on similarly complex models.

1.14 "Modeling Climate Policies: A Critical Look at Integrated Assessment Models" by Mathias Frisch (Original for this book)

Philosopher Mathias Frisch's contribution is concerned with the role of values in a different kind of climate model: the so-called "economy-climate integrated assessment models" (IAMs). IAMs, according to their developers and users, are modeling tools for weighing the cost and benefit of potential climate mitigation measures by calculating the future benefits of such measures and weighing them against their present costs. In other words, IAMs are purported to be able to tell us: if we spend X amount today on a climate mitigation measure, will we reap more, less, or the same value from it in the future?

Frisch points out that many IAMs produce outcomes that lead their users to be rather sanguine about the prospective damages from climate change and that many of them suggest that only very modest reductions in greenhouse gas emissions are needed in order to maximize eco-

nomic utility. But these results, and the models that deliver them, are dangerous, according to Frisch, because while they purport to offer us precise numbers to use for policy guidance, that precision is illusory and fraught with dangerous assumption and value judgments. Specifically, he argues that IAMs involve simplifying assumptions that are hard to defend. First, they vastly understate the uncertainty that we have regarding climate sensitivity (the amount the climate will change in response to a unit change in greenhouse gases). Second, they are extremely sensitive to assumptions about the economic impact of potential climate changes, built into the so-called "damage function," about which we have no good reasons for adopting the ones that particular modelers have chosen. Third, rather than modeling the preferences of each of the billions of people on earth, which would of course be impossible, they assume an "ideal consumer" who takes the various goods in the economy to be inter-substitutable. In particular, they only value goods in so far as they affect Gross Domestic Product. This makes them highly vexed with respect to how they treat the value of the enjoyment of an undamaged environment. Finally, they are highly sensitive to how much they discount the present value of future goods. All of which is to say: they necessarily involve deep normative commitments about the relative value of various possible goods, for which there are no possible empirical arguments. And unlike the kind of value-ladenness of climate modeling that Eric Winsberg argues for, where the values are in the nooks and crannies, Frisch argues that the values implicated in IAMs systematically affect their predictions in ways that directly track the values of their makers.

Finally, Frisch does conclude with some suggestions of how IAMs might nevertheless be employed fruitfully. He argues that the best possible use of IAMs is as very simple "toy models" that might allow us to explore possible scenarios and examine a range of possible risks we might face from various possible policy choices. This would enable us to use IAMs in a way that would avoid pretending that our models are precise when they are not. At the opposite extreme, he is somewhat more sanguine about the usefulness of much more complex and sophisticated models than the ones that appear in mainstream economic reports—

models that avoid some of Frisch's main criticisms by abstaining from trying to optimize the costs of abatement strategies with respect to future costs of climate change. As such, these more sophisticated models can avoid problematic assumptions like a choice of a damage function or a discount rate.

1.15 "Modeling Mitigation and Adaptation Policies to Predict their Effectiveness: The Limits of Randomized Controlled Trials" by Alexandre Marcellesi and Nancy Cartwright (Original for this book)

Our final contribution comes from philosophers Alexandre Marcellesi and Nancy Cartwright. Now we move beyond the interface between climate forecasting and policy response forecasting that Mathias Frisch explored and purely into the domain of policy response forecasting—that is, the domain of predicting the outcomes of various possible mitigation strategies that policy makers might adopt in the face of a potentially changing climate. Do not be surprised, therefore, that Marcellesi and Cartwright are interested in the strengths and limitation of randomized controlled trials (RCTs). This is not because they write about the use of RCTs to study the impact of increased greenhouse gases on the climate! (We hope that no one is proposing this idea.) Rather, they are interested in the use of RCTs and other causal inference methods for studying the impact potential of mitigation strategies. Suppose, for example, policymakers were to employ the various so-called "Payment for Environmental Services" (PES) programs wherein landowners are paid to change the way they use land so as to, say, reduce their carbon emissions. How effective would such a policy intervention be? How can we tell?

Marcellesi and Cartwright argue that there is a misplaced confidence in RCTs in the policy community. RCTs, they argue, are not the silver bullets that they are often thought to be because, though they almost assure high confidence in their internal validity, the question of their external validity is another matter entirely. In the end, the authors come

neither to praise nor bury RCT, but merely to point out that, like any other tool, they have their strengths and their limitations.

Notes

1. While a recent Gallup poll found that roughly 70% of Americans believe the claim that 2015 was the warmest year on record, we Americans remain split roughly 50/50 regarding the claim that the change in temperatures are caused by human activity (Gallup n.d.; http://www.gallup.com/poll/190319/americans-believe-2015-record-warm-split-why.aspx?g_source=CATEGORY_CLIMATE_CHANGE&g_medium=topic&g_campaign=tiles).
2. J. Cook et al. (2016), "Consensus on consensus: a synthesis of consensus estimates on human-caused global warming," Environmental Research Letters Vol. 11 No. 4, (13 April 2016); DOI:https://doi.org/10.1088/1748%E2%80%939326/11/4/048002 Quotation from page 6: "The number of papers rejecting AGW [Anthropogenic, or human-caused, Global Warming] is a miniscule proportion of the published research, with the percentage slightly decreasing over time. Among papers expressing a position on AGW, an overwhelming percentage (97.2% based on self-ratings, 97.1% based on abstract ratings) endorses the scientific consensus on AGW."
3. http://www.aaas.org/sites/default/files/migrate/uploads/1021climate_letter1.pdf
4. See http://www.gfdl.noaa.gov/earth-system-model for a description of one of the "flagship" American models.
5. Ben Santer, Personal Communication, 2011.
6. Press release from conference held at US National Press Club, January 2008.
7. In climate science, this generally means that the results tend to lean in one direction without a good reason or apparent cause.

References

Bukovsky, Melissa S., Rachel R. McCrary, Anji Seth, and Linda O. Mearns. (2017). A Mechanistically Credible, Poleward Shift in Warm-Season Precipitation Projected for the U.S. Southern Great Plains? *Journal of Climate*.

Christy, J.R., and R.W. Spencer. 2006. Satellite Temperature Data. In *Washington Roundtable on Science & Public Policy*, 1–37. Washington, DC: George Marshall Institute.

Cook, John, Naomi Oreskes, Peter T. Doran, William R.L. Anderegg, Bart Verheggen, Ed W. Maibach, J. Stuart Carlton, et al. 2016. Consensus on Consensus: A Synthesis of Consensus Estimates on Human-Caused Global Warming. *Environmental Research Letters* 11 (4): 048002.

Douglass, David H., John R. Christy, Benjamin D. Pearson, and S. Fred Singer. 2008. A Comparison of Tropical Temperature Trends with Model Predictions. *International Journal of Climatology* 28 (13): 1693–1701.

Houghton, J., Y. Ding, D. Griggs, M. Noguer, P. van der Linden, X. Dai, K. Maskell, and C. Johnson. 2001. *Climate Change 2001: The Scientific Basis*. Cambridge, UK: Cambridge University Press.

Gallup. Americans Believe 2015 Was Record-Warm, but Split on Why. *Gallup.Com*. http://www.gallup.com/poll/190319/americans-believe-2015-record-warm-split-why.aspx. Accessed 6 Aug 2017.

Giorgi, Filippo, Csaba Torma, Erika Coppola, Nikolina Ban, Christoph Schar, and Samual Samot. 2016. Enhanced Summer Convective Rainfall at Alpine High Elevations in Response to Climate Warming. *Nature Geoscience Letters* 9: 584–590. https://doi.org/10.1038/NGE02761.

Lloyd, Elisabeth Anne. 1994. *The Structure and Confirmation of Evolutionary Theory*. Princeton: Princeton University Press.

Lloyd, Elisabeth A. 2010. Confirmation and Robustness of Climate Models. *Philosophy of Science* 77 (5): 971–984.

Lloyd, Elisabeth A. 2015. *Model Robustness* as a Confirmatory Virtue: The Case of Climate Science. *Studies in History and Philosophy of Science* 49: 58–68. https://doi.org/10.1016/j.shpsa.2014.12.002.

Lloyd, Elisabeth A., Melissa Bukovsky, and Linda Mearns. 2017 Ms. An Analysis of Disagreement About Added Value by Regional Climate Models (under review).

Maraun, D., F. Wetterhall, A.M. Ireson, R.E. Chandler, E.J. Kendon, M. Widmann, S. Brienen, H.W. Rust, T. Sauter, M. Themessl, V.K.C. Venema, K.P. Chun, C.M. Goodess, R.G. Jones, C. Onof, M. Vrac, and I. Thiele-Eich. 2010. Precipitation Downscaling Under Climate Change: Recent Developments to Bridge the Gap Between Dynamica Models and the End User. *Reviews of Geophysics* 48. https://doi.org/10.1029/2009rg000314.

Mearns, Linda O., W.J. Gutowski, R. Jones, L.-Y. Leung, S. McGinnis, A.M.B. Nunes, et al. 2009. A Regional Climate Change Assessment Program for North America. *EOS* 90: 311–312.

Mearns, Linda O., Melissa S. Bukovsky, Sara C. Pryor, and Victor Magana. 2014. Downscaling of Climate Information. In *Climate Change in North America*, Regional Climate Studies, ed. G. Ohring, 201–250. Switzerland: Springer International Publishing.

Parker, Wendy S. 2011. When Climate Models Agree: The Significance of Robust Model Predictions. *Philosophy of Science* 78: 579–600.

Santer, B.D., et al. 1999. Uncertainties in Observationally Based Estimates of Temperature Change in the Free Atmosphere. *Journal of Geophysical Research 104*: 6305–6333.

Santer, B.D., T.M.L. Wigley, C. Mears, F.J. Wentz, S.A. Klein, D.J. Seidel, K.E. Taylor, et al. 2005. Amplification of Surface Temperature Trends and Variability in the Tropical Atmosphere. *Science* 309 (5740): 1551–1556.

Santer, B.D., P.W. Thorne, L. Haimberger, K.E. Taylor, T.M.L. Wigley, J.R. Lanzante, S. Solomon, M. Free, P.J. Gleckler, P.D. Jones, T.R. Karl, S.A. Klein, C. Mears, D. Nychka, G.A. Schmidt, S.C. Sherwood, and F.J. Wentz. 2008. Consistency of Modelled and Observed Temperature Trends in the Tropical Troposphere. *International Journal of Climatology* 28 (13): 1703–1722.

Part I

Confirmation and Evidence

2

The Scientific Consensus on Climate Change: How Do We Know We're Not Wrong?

Naomi Oreskes

2.1 Introduction

In December 2004, *Discover* magazine ran an article on the top science stories of the year. One of these was climate change, and the story was the emergence of a scientific consensus over the reality of global warming. *National Geographic* similarly declared 2004 as the year that global warming "got respect" (Roach 2004).

Many scientists felt that respect was overdue. As early as 1995, the Intergovernmental Panel on Climate Change (IPCC) had concluded that "the balance of evidence" supported the conclusion that humans were having an impact on the global climate (Houghton et al. 1995). By 2007, the IPCC's Fourth Assessment Report found a stronger voice, declaring that warming was "unequivocal," and noting that it is "extremely unlikely that the global climate changes of the past fifty years can be explained without invoking

N. Oreskes (✉)
Harvard University, Cambridge, MA, USA

E.A. Lloyd, E. Winsberg (eds.), *Climate Modelling*,
https://doi.org/10.1007/978-3-319-65058-6_2

human activities" (Alley et al. 2007). Prominent scientists and major scientific organizations have all ratified the IPCC conclusion (Oreskes 2004). Today, all but a tiny handful of climate scientists are convinced that earth's climate is heating up and that human activities are a primary driving cause (Doran and Zimmerman 2009; Anderegg et al. 2010; Cook et al., 2016).

Yet many Americans continued to wonder. A 2006 poll reported in *Time* magazine found that only just over half (56 percent) of Americans thought that average global temperatures had risen—despite the fact that virtually all climate scientists think that they have.[1] Since 2006, public opinion has wavered—influenced by short-term fluctuations in weather, as well as by political and cultural considerations whose relationship to climate change is indirect at best (Leiserowitz et al. 2011, and refs cit.). But one thing that has remained consistent is a gap between the virtually unanimous opinion of scientists that man-made climate change is underway and the continued doubts of a significant proportion of the American people (Leiserowitz et al. 2011; see also Borick et al. 2010). Moreover, as Jon Krosnick and his colleagues have stressed, while the scientific community has for some time believed that the evidence of climate change "justifies substantial public concern," the public has not broadly shared that view (Krosnick et al. 2006, see also Lorenzoni and Pidgeon 2006).

This book addresses the scientific study of climate change and its impacts. By definition, predictions are uncertain, and people may wonder why we should spend time, effort, and money addressing a problem that may not affect us for years or decades to come. Some people have gone further, suggesting that it would be foolish to spend time and money addressing a problem that might not actually even be a problem. After all, how do we really know?

This chapter addresses the question: how *do* we know? Put another way, even if there is a scientific consensus, how do we know it's not wrong? If the history of science teaches anything, it is humility. There are numerous historical examples where expert opinion turned out to be wrong. At the start of the twentieth century, Max Planck was advised not to go into physics because all the important questions had been answered, medical doctors prescribed arsenic for stomach ailments, and geophysicists were confident that continents did not drift. In any scientific community there are individuals who depart from generally accepted views, and occasion-

ally they turn out to be right. At present, there is a scientific consensus on global warming, and that consensus has been stable for at least a decade. But how do we know it's not wrong?

2.2 The Scientific Consensus on Climate Change

Let's start with a simple question: What is the scientific consensus on climate change, and how do we know it exists? Scientists do not vote on contested issues, and most scientific questions are far too complex to be answered by a simple yes or no response. So how does anyone know what scientists think about global warming?

Scientists glean their colleagues' conclusions by reading their results in published scientific literature, listening to presentations at scientific conferences, and discussing data and ideas in the hallways of conference centers, university departments, research institutes, and government agencies. For outsiders, this information is difficult to access: scientific papers and conferences are by experts for experts and are difficult for outsiders to understand.

Climate science is a little different. Because of the political importance of the topic, scientists have been motivated and asked to explain their research results in accessible ways, and explicit statements of the state of scientific knowledge are easy to find.

An obvious place to start is the Intergovernmental Panel on Climate Change (IPCC). Created in 1988 by the World Meteorological Organization and the United Nations Environment Programme, the IPCC evaluates the state of climate science as a basis for informed policy action, primarily on the basis of peer-reviewed and published scientific literature (IPCC 2005). The IPCC has issued five assessments, with a sixth due in 2014. Already in 2001, the IPCC had stated the consensus scientific opinion that Earth's climate is being affected by human activities. This view is expressed throughout the report, but perhaps the clearest statement is this: "Human activities … are modifying the concentration of atmospheric constituents …that absorb or scatter radiant energy… [M]ost of the observed warming over the last 50 years is likely to have

been due to the increase in greenhouse gas concentrations" (McCarthy et al. 2001, 21). The 2007 IPCC reports updates this to "very likely" (Alley et al. 2007).

From a historical perspective, the IPCC is a somewhat unusual scientific organization: it was created not to discover new knowledge but to compile and assess existing knowledge on a politically sensitive and economically significant issue. Its conclusions might be skewed by these extra-scientific concerns. But the IPCC is by no means alone in its conclusions; its results have been repeatedly ratified by other scientific organizations.

All of the major scientific bodies in the United States whose membership's expertise bears directly on the matter have issued reports or statements that confirm the IPCC conclusion. One is the National Academy of Sciences report, *Climate Change Science: An Analysis of Some Key Questions* (2001), which originated from a White House request. Here is how it opens: "Greenhouse gases are accumulating in Earth's atmosphere as a result of human activities, causing surface air temperatures and subsurface ocean temperatures to rise" (National Academy of Sciences 2001, 1). The report explicitly addresses whether the IPCC assessment is a fair summary of professional scientific thinking and answers yes: "The IPCC's conclusion that most of the observed warming of the last 50 years is likely to have been due to the increase in greenhouse gas concentrations accurately reflects the current thinking of the scientific community on this issue" (National Academy of Sciences 2001, 3).

Other US scientific groups have agreed. In February 2003, the American Meteorological Society adopted the following statement on climate change: "There is now clear evidence that the mean annual temperature at the Earth's surface, averaged over the entire globe, has been increasing in the past 200 years. There is also clear evidence that the abundance of greenhouse gases has increased over the same period… Because human activities are contributing to climate change, we have a collective responsibility to develop and undertake carefully considered response actions" (American Meteorological Society 2003). So too says the American Geophysical Union: "Scientific evidence strongly indicates that natural influences cannot explain the rapid increase in global near-surface temperatures observed during the second half of the 20th century" (American Geophysical Union Council 2003). Likewise the

American Association for the Advancement of Science: "The world is warming up. Average temperatures are half a degree centigrade higher than a century ago. The nine warmest years this century have all occurred since 1980, and the 1990s were probably the warmest decade of the second millennium. Pollution from 'greenhouse gases' such as carbon dioxide (CO_2) and methane is at least partly to blame" (Harrison and Pearce 2000). In short, these groups have all affirmed that global warming is real and substantially attributable to human activities. (And today, the observed increase in mean global temperature is nearly a full degree, centrigrade.)

If we extend our purview beyond the United States, we find this conclusion further reinforced. In 2005, the Royal Society of the UK, one of the world's oldest and most respected scientific societies, issued a "Guide to Facts and Fictions about Climate Change," debunking various myths asserting that climate change is not occurring, that it is not caused by human activities, that observed changes are within the range of natural variability, that CO_2 is too trivial to matter, that climate models are unreliable, and that the IPCC is biased and does not fairly represent the scientific uncertainties.

On the latter point, the report takes pains to underscore the scientific authority of the IPCC, noting that "the IPCC is the world's leading authority on climate change and its impacts," and that its work is backed by the worldwide scientific community.[2] This point was underscored in 2007, when the National Academies of 13 countries (G8+ 5) issued a joint statement calling attention to the problem of anthropogenic climate change, and urging a rapid transition to a low carbon society.[3]

One website dedicated to evaluating the scientific consensus on climate change counts 27 scientific societies that have formally endorsed the conclusion that "most of the global warming in recent decades can be attributed to human activities"—just in North America, Europe, and Australia—as well as 13 National Academies in Africa.[1] If we were to do a comprehensive count of scientific societies in Asia, Africa, and South America, the figure would no doubt be still higher.

Consensus reports and statements are drafted through a careful process involving many opportunities for comment, criticism, and revision, so it is unlikely that they would diverge greatly from the opinions of the

societies' memberships. Nevertheless, it could be the case that they downplay dissenting opinions.[3]

One way to test that hypothesis is by analyzing the contents of published scientific papers, which contain the views that are considered sufficiently supported by evidence that they merit publication in expert journals. After all, any one can *say* anything, but not anyone can get research results published in a refereed journal.[4] Papers published in scientific journals must pass the scrutiny of critical, expert colleagues. They must be supported by sufficient evidence to convince others who know the subject well. So one must turn to the scientific literature to be certain of what scientists really think.

Before the twentieth century, this would have been a trivial task. The number of scientists directly involved in any given debate was usually small. A handful, a dozen, perhaps a hundred, at most, participated—in part because the total number of scientists in the world was small (Price 1986). Moreover, because professional science was a limited activity, many scientists used language that was accessible to scientists in other disciplines as well as to serious amateurs. It was relatively easy for an educated person in the nineteenth or early twentieth century to read a scientific book or paper and understand what the scientist was trying to say. One did not have to be a scientist to read *The Principles of Geology* or *The Origin of Species.*

Our contemporary world is different. Today, hundreds of thousands of scientists publish over a million scientific papers each year.[5] The American Geophysical Union has over 60,000 members in 135 countries, and the American Meteorological Society has nearly 14,000. The IPCC reports involved the participation of many hundreds of scientists from scores of countries (Houghton et al. 1990; Alley et al. 2007), still more if reviewers are included in the head count. No individual could possibly read all the scientific papers on a subject without making a full-time career of it.

Fortunately, the growth of science has been accompanied by the growth of tools to manage scientific information. One of the most important of these is the database of the Institute for Scientific Information (ISI). In its Web of Science, the ISI indexes all papers published in refereed scientific journals every year—over 8500 journals. Using a key word or phrase, one

can sample the scientific literature on any subject and get an unbiased view of the state of knowledge.

Figure 2.1 shows the results of an analysis of 928 abstracts, published in refereed journals during the period 1993–2003, that I completed in 2004, to evaluate the state of scientific debate at that time, using the Web of Science data base.[6]

After a first reading to determine appropriate categories of analysis, the papers were divided as follows: (1) those explicitly endorsing the consensus position, (2) those explicitly refuting the consensus position, (3) those discussing methods and techniques for measuring, monitoring, or predicting climate change, (4) those discussing potential or documenting actual impacts of climate change, (5) those dealing with paleoclimate change, and (6) those proposing mitigation strategies. How many fell into category 2—that is, how many of these papers present evidence that refutes the statement: "Global climate change is occurring, and human activities are at least part of the reason why"? The answer is remarkable: none.

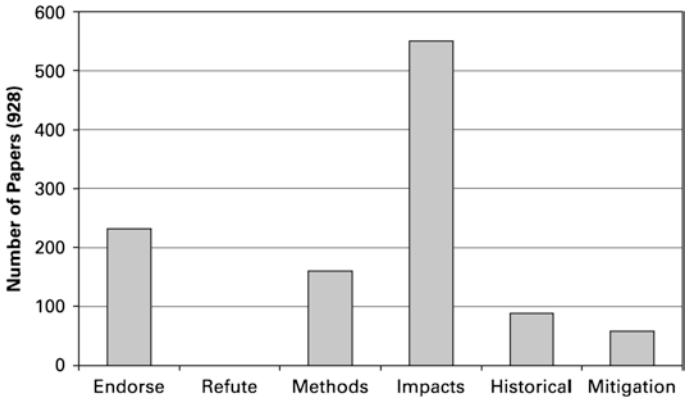

Fig. 2.1 A Web of Science analysis of 928 abstracts using the keywords "global climate change." No papers in the sample provided scientific data or theoretical arguments to refute the consensus position on the reality of global climate change (It should be acknowledged that in any area of human endeavor, leadership may diverge from the views of the led. For example, many Catholic priests endorse the idea that priests should be permitted to marry (Watkin 2004))

A few comments are in order. First, often it is challenging to determine exactly what the authors of a paper do think about global climate change. This is a consequence of experts writing for experts: many elements are implicit. If a conclusion is widely accepted, then it is not necessary to reiterate it within the context of expert discussion. Scientists generally focus their discussions on questions that are still disputed or unanswered rather than on matters about which everyone agrees.

This is clearly the case with the largest portion of the papers examined (approximately half of the total)—those dealing with impacts of climate change. The authors evidently accept the premise that climate change is real and want to track, evaluate, and understand its impacts. Nevertheless, such impacts could, at least in principle, be the results of natural variability rather than human activities. Strikingly, none of the papers used that possibility to argue against the consensus position.

Roughly 15 percent of the papers dealt with methods, and slightly less than 10 percent dealt with paleoclimate change. The most notable trend in the data is the recent increase in such papers; concerns about global climate change have given a boost to research in paleoclimatology and to the development of methods for measuring and evaluating global temperature and climate. Such papers are essentially neutral with respect to the reality of current anthropogenic change: developing better methods and understanding historic climate change are important tools for evaluating current effects, but they do not commit their authors to any particular opinion about those effects. Perhaps some of these authors are in fact skeptical of the current consensus, and this could be a motivation to work on a better understanding of the natural climate variability of the past. But again, none of the papers used that motivation to argue openly against the consensus, and it would be illogical if they did because a skeptical motivation does not constitute scientific evidence. Finally, approximately 20 percent of the papers explicitly endorsed the consensus position, and an additional five percent proposed mitigation strategies. In short, by 2003, the basic reality of anthropogenic global climate change was no longer a subject of scientific debate.[7]

Some readers were surprised by this result and questioned the reliability of a study that failed to find arguments against the consensus position when such arguments clearly existed. After all, anyone who watched Fox

news or MSNBC or trolled the Internet knew that there was an enormous debate about climate change, right? Well, no.

First, let's make clear what the scientific consensus is. It is over the reality of human-induced climate change. Scientists predicted a long time ago that increasing greenhouse gas emissions could change the climate, and now there is overwhelming evidence that it *is* changing the climate. These changes are *in addition* to natural variability. Therefore, when contrarians try to shift the focus of attention to natural climate variability, they are misrepresenting the situation. No one denies the fact of natural variability, but natural variability alone does not explain what we are now experiencing. Scientists have also documented that many of the changes that are now occurring are deleterious to both human and nonhuman communities (Arctic Council 2004, IPCC AR4). Because of global warming, sea level is rising, humans are losing their homes and hunting grounds, plants and animals are losing their habitats, and extreme weather events (particularly droughts and heat waves) are becoming more common and in some cases more extreme (Kolbert 2006; Flannery 2006, IPCC AR4, IPCC 2012).

Second, to say that man-made global warming is underway is not the same as agreeing about what will happen in the future. Much of the recent and continuing debate in the scientific community involves the likely rate of future change. A good analogy is evolution. In the early twentieth century, paleontologist George Gaylord Simpson introduced the concept of "tempo and mode" to describe questions about the manner of evolution—how fast and in what manner evolution proceeded. Biologists by the mid-twentieth century agreed about the reality of evolution, but there were extensive debates about its tempo and mode. So it is now with climate change. Virtually all professional climate scientists agree that human-induced climate change is underway, but debate continues on tempo and mode.

Third, there is the question of what kind of dissent still exists. My analysis of the published literature was done by sampling published papers, using a keyword phrase that was intended to be fair, accurate, and neutral: "global climate change" (as opposed to, e.g., "global warming," which might be viewed as biased). The total number of scientific papers published over that 10-year period having anything at all to do with climate

change was over 10,000; it is likely that some of the authors of the unsampled papers expressed skeptical or dissenting views. But given that the sample turned up no dissenting papers at all, professional dissention must have been very limited.

Recent work has supported this conclusion, showing that 97–98 percent of professional climate scientists affirm the reality of anthropogenic climate change as outlined by the IPCC (Anderegg et al. 2012; see also Cook et al. 2013, 2016). This also affirms the conclusions of Max and Jules Boykoff (2004, see also Freudenburg and Muselli 2010; Boykoff 2011) that the mass media have given air and print space to a handful of dissenters to a degree that is greatly disproportionate with their representation in the scientific community. Many articles on climate change, for example, will quote two mainstream scientists and one dissenter, where an accurate reflection of the state of the science would be to quote 30 or 40 mainstream scientists for every dissenter. (On television and radio the situation is even worse, where a debate is set up between one mainstream scientist and one dissenter, as if the actual distribution of views in the scientific community were fifty-fifty.) There are climate scientists who actively do research in the field but disagree with the consensus position, but their number is very, very small. This is not to say that there are not a significant number of *contrarians*, but to point out that the vast majority of them are not climate scientists.

In fact, most contrarians are not even scientists at all. Some, like the physicist Frederick Seitz (who for many years challenged the scientific evidence of the harms of tobacco along with the threat of climate change), were once scientific researchers but not in the field of climate science. (Seitz was a solid-state physicist.) Others, like Michael Crichton, who for many years was a prominent speaker on the contrarian lecture circuit, are novelists, actors, or others with access to the media, but no scientific credentials. What Seitz and Crichton have in common, along with most other contrarians, is that they do no new scientific research. They are not producing new evidence or new arguments. They are simply attacking the work of others, and doing so in the court of public opinion and in the mass media rather than in the halls of science.

This latter point is crucial and merits underscoring: the vast majority of books, articles, and websites denying the reality of global warming do not pass the most basic test for what it takes to be counted as scientific—

namely, being published in a peer-reviewed journal. Contrarian views have been published in books and pamphlets issued by politically motivated think tanks and widely spread across the Internet, but so have views promoting the reality of UFOs or the claim that Lee Harvey Oswald was an agent of the Soviet Union.

Moreover, some contrarian arguments are frankly disingenuous, giving the impression of refuting the scientific consensus when their own data do no such thing. One example will illustrate the point. In 2001, Willie Soon, a physicist at the Harvard-Smithsonian Center for Astrophysics, with several colleagues published a paper entitled "Modeling Climatic Effects of Anthropogenic Carbon Dioxide Emissions: Unknowns and Uncertainties" (Soon et al. 2001). This paper has been widely cited by contrarians as an important example of a legitimate dissenting scientific view published in a peer-reviewed journal.[8] But the issue under discussion is how well models can predict the future—in other words, tempo and mode. The paper does not refute the consensus position, and the authors acknowledge so: "The purpose of [our] review of the deficiencies of climate model physics and the use of GCMs is to illuminate areas for improvement. Our review does not disprove a significant anthropogenic influence on global climate" (Soon et al. 2001, 259; see also 2002).

The authors needed to make this disclaimer because many contrarians *do* try to create the impression that arguments about tempo and mode undermine the whole picture of global climate change. But they don't. Indeed, one could reject all climate models and still accept the consensus position because models are only one part of the argument—one line of evidence among many.

Is there disagreement over the details of climate change? Yes. Are all aspects of climate past and present well understood? No, but who has ever claimed that they were? Does climate science tell us what policy to pursue? Definitely not. But it does identify the problem, explain why it matters, and give society insights that can help to frame an efficacious policy response (e.g., Smith 2002; Oreskes et al. 2010).

So why does the public have the impression of disagreement among scientists? If the scientific community has forged a consensus, then why do so many Americans have the impression that there is serious scientific uncertainty about climate change?[9]

There are several reasons. First, it is important to distinguish between scientific and political uncertainties. There are reasonable differences of opinion about how best to respond to climate change and even about how serious global warming is relative to other environmental and social issues. Some people have confused—or deliberately conflated—these two issues. Scientists are in agreement about the reality of global climate change, but this does not tell us what to do about it.

Second, climate science involves prediction of future effects, which by definition are uncertain. It is important to distinguish among what is known to be happening now, what is likely to happen based on current scientific understanding, and what might happen in a worst-case scenario. This is not always easy to do, and scientists have not always been effective in making these distinctions. Uncertainties about the future are easily conflated with uncertainties about the current state of scientific knowledge.

Third, scientists have evidently not managed well enough to explain their arguments and evidence beyond their own expert communities. The scientific societies have tried to communicate to the public through their statements and reports on climate change, but what average citizen knows that the American Meteorological Society even exists or visits its home page to look for its climate-change statement?

There is also a deeper problem. Scientists are finely honed specialists trained to create new knowledge, but they have little training in how to communicate to broad audiences and even less in how to defend scientific work against determined and well-financed contrarians (Moser and Dilling 2004, idem 2007; Hassol 2008; Somerville and Hassol 2011). Moreover, until recently, most scientists have not been particularly anxious to take the time to communicate their message broadly. Most scientists consider their "real" work to be the production of knowledge, not its dissemination, and often view these two activities as mutually exclusive, or at least competitive. Some sneer at colleagues who communicate to broader audiences, dismissing them as "popularizers."

If scientists do jump into the fray on a politically contested issue, they may be accused of "politicizing" the science and compromising their objectivity.[10] This places scientists in a double bind: the demands of objectivity seem to suggest that they should keep aloof from contested issues, but if they don't get involved, no one will know what an objective view of

the matter looks like. Scientists' reluctance to present their results to broad audiences has left scientific knowledge open to misrepresentation, and recent events show that there are plenty of people ready and willing to misrepresent it.

It's no secret that politically motivated think tanks such as the American Enterprise Institute and the George Marshall Institute have been active for some time in trying to communicate a message that is at odds with the consensus scientific view (Gelbspan 1997, 2005; Mooney 2006; Oreskes and Conway 2012). These organizations have successfully garnered a great deal of media attention for the tiny number of scientists who disagree with the mainstream view and for nonscientists, like Crichton, who pronounce loudly on scientific issues.

This message of scientific uncertainty has been reinforced by the public relations campaigns of certain corporations with a large stake in the issue.[11] The most well-known example is ExxonMobil, which in the late 1990s and throughout the 2000s, ran a highly visible advertising campaign on the op-ed page of the *New York Times*. Its carefully worded advertisements—written and formatted to look like newspaper columns and called op-ed pieces by ExxonMobil—suggested that climate science was far too uncertain to warrant action on it (Supran and Oreskes, 2017).[12] One advertisement concluded that the uncertainties and complexities of climate and weather mean that "there is an ongoing need to support scientific research to inform decisions and guide policies" (Environmental Defense 2005; see also van den Hove et al., 2002). Not many would argue with this unobjectionable claim, unless it is taken to imply that decisions and policies taken now would be premature. Our scientists have long ago concluded that existing research warrants that decisions and policies be made *today*.[13]

In any scientific debate, past or present, one can always find intellectual outliers who diverge from the consensus view. Even after plate tectonics was resoundingly accepted by earth scientists in the late 1960s, a handful of persistent resisters clung to the older views, and some idiosyncratics held to alternative theoretical positions, such as earth expansion. Some of these men were otherwise respected scientists, including Sir Harold Jeffreys, one of Britain's leading geophysicists, and Gordon J. F. MacDonald, a one-time science adviser to Presidents Lyndon Johnson and Richard Nixon. Both these men rejected plate tectonics until their

dying day, which for MacDonald was in 2002. Does that mean that scientists should reject plate tectonics, that disaster-preparedness campaigns should not use plate tectonics theory to estimate regional earthquake risk, or that schoolteachers should give equal time in science classrooms to the theory of earth expansion? Of course not. That would be silly and a waste of time. In the case of earthquake preparedness, it would be dangerous as well.

No scientific conclusion can ever be proven, and new evidence may lead scientists to change their views, but it is no more a "belief" to say that earth is heating up than to say that continents move, that germs cause disease, that DNA carries hereditary information, that HIV causes AIDS, and that some synthetic organic chemicals can disrupt endocrine function. You can always find someone, somewhere, to disagree, but these conclusions represent our best current understandings and therefore our best basis for reasoned action (Oreskes 2004).

2.3 How Do We Know We're Not Wrong?

Might the consensus on climate change be wrong? Yes, it might be, and if scientific research continues, it is almost certain that some aspects of the current understanding will be modified, perhaps in significant ways. This possibility can't be denied. The relevant question for us as citizens is not whether this scientific consensus *might* be mistaken but rather whether there is any substantive reason to think that it *is* mistaken.

How can outsiders evaluate the robustness of any particular body of scientific knowledge? Many people expect a simple answer to this question. Perhaps they were taught in school that scientists follow "the scientific method" to get correct answers, and they have heard some climate-change deniers suggesting that climate scientists do not follow the scientific method (because they rely on models, rather than laboratory experiments) so their results are suspect. These views are wrong.

Contrary to popular opinion, there is no scientific method (singular). Despite heroic efforts by historians, philosophers, and sociologists to identify "the" scientific method, they have failed. There is no generally agreed-upon answer as to what the methods and standards of science are

(or even what they should be). There is no methodological litmus test for scientific reliability and no single method that guarantees valid conclusions that will stand up to all future scrutiny.

A positive way of saying this is that scientists have used a variety of methods and standards to good effect and that philosophers have proposed various helpful criteria for evaluating the methods used by scientists. None is a magic bullet, but each can be useful for thinking about what makes scientific information a reliable basis for action.[14] So we can pose the question: how does current scientific knowledge about climate stand up to these diverse models of scientific reliability?

The Inductive and Deductive Models of Science

The most widely cited models for understanding scientific reasoning are induction and deduction. *Induction* is the process of generalizing from specific examples. If I see 100 swans and they are all white, I might conclude that all swans are white. If I saw 1000 white swans or 10,000, I would surely think that *all* swans were white, yet a black one might still be lurking somewhere. As David Hume famously put it, even though the sun has risen thousands of times before, we cannot *prove* that it will rise again tomorrow.

Nevertheless, common sense tells us that the sun will rise again tomorrow, even if we can't logically prove that it's so. Common sense similarly tells us that if we had seen 10,000 white swans, then our conclusion that all swans were white would be more robust than if we had seen only 10. Other things being equal, the more we know about a subject, and the longer we have studied it, the more likely our conclusions about it are to be true.

How does climate science stand up to the inductive model? Does climate science rest on a strong inductive base? Yes. Humans have been making temperature records consistently for over 150 years, and nearly all scientists who have looked carefully at these records see an overall temperature increase since the industrial revolution. (Houghton et al. 1990; Bruce et al. 1996; Watson et al. 1996; McCarthy et al. 2001; Houghton et al. 2001; Metz et al. 2001; Watson 2001; Weart 2008). According to the IPCC's AR4, the temperature rise over the 100-year

period from 1906 to 2005 was 0.74 °C [0.56–0.92 °C] with a confidence interval of 90 percent (Alley et al. 2007). The empirical signal is clear, even if all the details are not.

How reliable are the early records? And how do you average data to be representative of the globe as a whole, when most of the early data comes from only a few places, generally in Europe? Scientists have spent quite a bit of time addressing these questions; most have satisfied themselves that the empirical signal is clear (Edwards 2010). Even if scientists doubted the older records, the more recent data show a strong increase in temperatures over the past 30 to 40 years, just when the amount of CO_2 and other greenhouses gases in the atmosphere was growing dramatically (McCarthy et al. 2001; Houghton 2001; Metz et al. 2001; Watson 2001). Recently, an independent assessment by the Berkeley Earth Surface Temperature group found that over the past 50 years the land surface warmed by 0.91 °C, a result that confirms the prior work by NASA, NOAA, and the U.K. Hadley Centre (Muller et al. 2013). The Berkeley group has also reviewed the question of the "heat island effect"—the possible exaggeration of the warming effect due to the location of weather stations in urban areas, which are warmer than rural ones because of buildings, concrete, automobiles, etc.—a potential source of error much emphasized by some contrarians (Wickham et al. 2013)—and finds that the observed warming cannot be explained away this way.

The Berkeley study received a good deal of media attention—arguably out of proportion to its scientific significance—because its spokesman, physicist Richard Muller, was previously a self-proclaimed skeptic, and because some of his funding came from the Koch Industries, a Fortune 500 company heavily involved in petroleum refining, oil and gas pipelines, and petrochemicals. (Both Koch brothers are political libertarians, opposed to environmental regulation: David Koch ran in 1980 for Vice President on the Libertarian party ticket, and Charles Koch is one of the founders of the Cato Institute, which has played a large role in US climate change denial; see Oreskes and Conway 2012.) But despite a flurry of media attention, Richard Muller's late-stage conversion had little political impact, and even less scientific, because the conclusions from the instrumental records that he first questioned but then affirmed have been amply corroborated by other, independent evidence from tree rings, ice cores,

and coral reefs (IPCC, Alley et al. 2007). A paper in 2003 by a team lead by Jan Esper at the Swiss Federal Research Center, for example, had already demonstrated that tree rings can provide a reliable, long-term record of temperature variability, one which largely agrees with the instrumental records over the past 150 years (Esper et al. 2002).

Muller's reanalysis of existing temperature records raises the fundamental problem facing all inductive science: how many data are enough? If you have counted 10,000 white swans—or 100,000, or even 1,000,000—how do you know that a black swan isn't lurking around the corner? How do you know that the generalization you made from your observations is correct? After all, other generalizations could also be consistent with your observations.

The logical limitations of the inductive view of science have led some to argue that the core of scientific method is testing theories through logical deductions. *Deduction* is drawing logical inferences from a set of premises—the stock-in-trade of Sherlock Holmes. In science, deduction is generally presumed to work as part of what has come to be known as the *hypothetico-deductive model*—the model you will find in most textbooks that claim to teach the scientific method (sometimes also called the *deductive-nomological* model, referring to the idea that ultimately science seeks to develop not just hypotheses, but laws, from which conclusions may be deduced).

In this view, scientists develop hypotheses and then test them. Every hypothesis has logical consequences—deductions—and one can try to determine, primarily through experiment and observation, whether the deductions are correct. If they are, they support the hypothesis. If they are not, then the hypothesis must be revised or rejected. It's often considered especially good if the prediction is something that would otherwise be quite unexpected, because that would suggest that it didn't just happen by chance.

The most famous example of successful deduction in the history of science is the case of Ignaz Semmelweis, who in the 1840s deduced the importance of handwashing to prevent the spread of infection (Gillispie 1975; Hempel 1965). Semmelweis had noticed that many women were dying of fever after giving birth at his Viennese hospital. Surprisingly, women who had their infants on the way to the hospital—seemingly

under more adverse conditions—rarely died of fever. Nor did women who gave birth at another hospital clinic where they were attended by midwives. Not surprisingly, Semmelweis was troubled by this pattern, which seemed to suggest that it was more dangerous to give birth when attended by a doctor than by a midwife, and more dangerous to give birth in a hospital than in a horse-drawn carriage.

In 1847, a friend of Semmelweis, Jakob Kolletschka, cut his finger while doing an autopsy and soon died. Autopsy revealed a pathology very similar to the women who had died after childbirth; something in the cadaver had apparently caused his death. Semmelweis knew that many of the doctors at his clinic routinely went directly from conducting autopsies to attending births, but midwives did not perform autopsies. So he hypothesized that the doctors were carrying cadaveric material on their hands, which was infecting the women (and killed his friend). He deduced that if physicians washed their hands before attending the women, then the infection rate would decline. They did so, and the infection rate did decline, demonstrating the power of the hypothetico-deductive method.

How does climate science stand up to this standard? Have climate scientists made predictions that have come true? Absolutely. The most obvious is the fact of global warming itself. Scientific concern over the effects of increased atmospheric CO_2 is based on physics—the fact that CO_2 is a greenhouse gas, something that has been known since the mid-nineteenth century. In the early twentieth century, Swedish chemist Svante Arrhenius predicted that increasing CO_2 from the burning of fossil fuels would lead to global warming, and by midcentury, a number of other scientists, including G. S. Callendar, Roger Revelle, and Hans Suess, concluded that the effect might soon be quite noticeable, leading to sea level rise and other global changes (Fleming 1998; Weart 2008). In 1965, Revelle and his colleagues wrote, "By the year 2000, the increase in atmospheric CO_2 … may be sufficient to produce measurable and perhaps marked change in climate, and will almost certainly cause significant changes in the temperature and other properties of the stratosphere" (Revelle 1965, 9). This prediction has come true (McCarthy et al. 2001; Houghton et al. 2001; Metz et al. 2001; Watson 2001).

Another prediction fits the category of something unusual that you might not even think of without the relevant theory. In 1980, climatologist Suki Manabe predicted that the effects of global warming would be strongest first in the polar regions. *Polar amplification* was not an induction from observations but a deduction from theoretical principles: the concept of ice-albedo feedback. The reflectivity of a material is called its *albedo*. Ice has a high albedo, reflecting sunlight into space much more effectively than grass, dirt, or water. One reason polar regions are as cold as they are is that snow and ice are very effective in reflecting solar radiation back into space. But if the snow starts to melt and bare ground (or water) is exposed, this reflective effect diminishes. Less ice means less reflection, which means more solar heat is absorbed, leading to yet more melting in a feedback loop. So once warming begins, its effects accelerate; Manabe and his colleagues thus predicted that warming would be more pronounced in polar regions than in temperate ones. The Arctic Climate Impact Assessment concluded in 2004 that this prediction had come true (Manabe and Stouffer 1980, 1994; Holland and Bitz 2003; Arctic Council 2004).

Falsification

Ignaz Semmelweis is among the famous figures in the history of science because his work in the 1840s foreshadows the germ theory of disease and the saving of millions of human lives. His story is a great one, told and retold many times. But the story has a twist because Semmelweis was right for the wrong reason. Cadaveric matter was *not* the cause of the infections: germs were. In later years, this would be demonstrated by James Lister, Robert Koch, and Louis Pasteur, who realized that handwashing was effective not because it removed the cadaveric material, but because it removed the germs associated with that material.

The story illustrates a fundamental flaw with the hypothetico-deductive model—the fallacy of affirming the consequent. If I make a prediction and it comes true, I may assume that my theory is correct. But this would be a mistake, for the accuracy of my deduction does not prove that my hypothesis was correct; my prediction may have come true for other

reasons, as indeed Semmelweis' did. The other reasons may be related to the hypothesis—germs *were* associated with cadaveric matter—but in other cases the connection may be entirely coincidental. I can convince myself that I have proved my theory right, but this would be self-deception.

This realization led the twentieth-century philosopher Karl Popper to suggest that you can never prove a theory true. Any affirmation of a hypothesis through deduction runs to the risk of the fallacy of affirming the consequent. However, if the prediction does not come true, then you do know that there is something wrong with your hypothesis. Thus Popper emphasized that while we cannot prove a theory true, we can prove it false. Thus, scientific theories must be "falsifiable"—able to be shown, through experiment or observation—that they are false, and the scientific method is not to prove theories, but to show them to be false, a view known as *falsificationism* (Popper 1959).

How does climate science hold up to this modification? Can climate models be refuted? Falsification is a bit of a problem for models—not just climate models—because many models are built to forecast the future and the results will not be known for some time. By the time we find out whether the long-term predictions of a model are right or wrong, that knowledge won't be of much use. So while model predictions might be falsifiable in principle, many are not actually falsifiable in practice.

For this reason, many models are tested by seeing if they can accurately reproduce past events—what is sometimes called *retrodiction*. In principle, retrodiction should be a rigorous test: a climate model that fails to reproduce past temperature records is obviously faulty, and could be considered falsified. In reality, it doesn't work quite that way.

Climate models are complex, and they involve many variables—some that are well measured and others that are not. If a model does not reproduce past data very well, most modelers assume that one or more of the model parameters are not quite right, and they make adjustments in an attempt to obtain a better fit. This is generally referred to as *model calibration,* and many modelers consider it an essential part of the process of building a good model. But calibration can make models refutation-proof: the model doesn't get rejected; it gets revised. Given the complex-

ity of climate models, there are myriad ways a model can be revised to ensure that it successfully retrodicts past climate change. Thus, in practice, the idea of falsification is not of great use in judging climate models.

Recently, however, one modeler has put his model to the test by making a genuine prediction of the future. When the Philippine volcano Mt. Pinatubo erupted in 1991, millions of tons of sulfur dioxide, ash, and dust were thrown into the atmosphere. NASA climate modeler James Hansen realized that these materials were likely to cause a global cooling effect, and that it was possible to use the NASA-GISS climate model to predict what that cooling would be. The model had been built to simulate long-term global warming, not short-term global cooling, but still, if the physics of the model were correct, he reasoned, it ought to be able to make this prediction. Hansen and his team ran the model, and forecast a short-term cooling effect of about 0.5 degree, that would briefly overwhelm the general warming trend from greenhouse gases (Hansen et al. 1992). That prediction came true (Kerr 1993).

This is still only one test, however, and if model results were the only basis for current scientific understanding, there would be grounds for some healthy skepticism. Models are therefore best viewed as heuristic devices: a means to explore what-if scenarios. This is, indeed, how most modelers use them: to answer questions like "If we double the amount of CO_2 in the atmosphere, what is the most likely outcome?"

One way in which modelers address the fact that a model can't be proved right or wrong is to make lots of different models that explore diverse possible outcomes—what modelers call *ensembles.* An example of this is <climateprediction.net>, a Web-based mass-participation experiment that enlists members of the public to run climate models on their home computers to explore the range of likely and possible climate outcomes under a variety of plausible conditions.

Over 90,000 participants from over 140 countries have produced tens of thousands of runs of a general circulation model produced by the Hadley Centre for Climate Prediction and Research. Figure 2.2 presents some initial results, published in the journal *Nature* in 2005, for a steady-state model in which atmospheric CO_2 is doubled relative to preindustrial levels and the model earth is allowed to adjust.

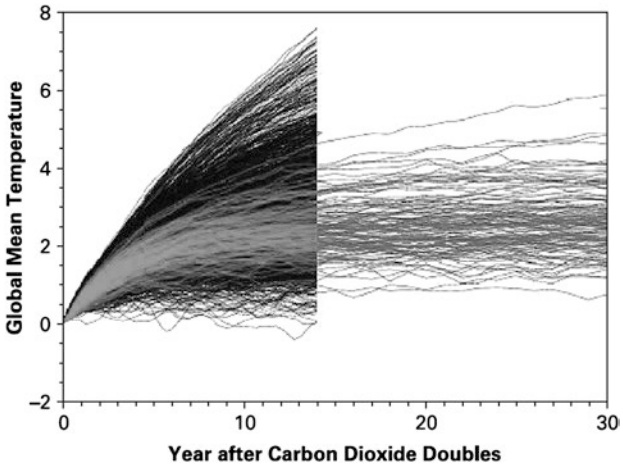

Fig. 2.2 Changes in global mean surface temperature after carbon dioxide values in the atmosphere are doubled. The *black lines* show the results of 2579 fifteen-year simulations by members of the general public using their own personal computers. The *gray lines* show comparable results from 127 thirty-year simulations completed by Hadley Centre scientists on the Met Office's supercomputer (<www.metoffive.gov.uk>). Figure prepared by Ben Sanderson with help from the <climateprediction.net> project team (Source: Reproduced by permission from http://www.climateprediction.net/science/results_cop10.phpi)

The results in black are the climateprediction.net's mass-participation runs; the results in gray come from runs made by professional climate scientists at the Hadley Centre on a supercomputer (Stainforth et al. 2005).

What does an ensemble like this show? For one thing, no matter how many times you run the model, you almost always get the same qualitative result: the earth will warm. The unanswered question is how much and how fast—in other words, tempo and mode.

The models vary quite a bit in their tempo and mode, but nearly all fall within a temperature range of 1–7 °C (2–14 °F) within 15 years after the earth's atmosphere reaches a doubling of atmospheric CO_2. Moreover, most of the runs are still warming at that point. The model runs were stopped at year 15 for practicality, but most of them had not yet reached equilibrium: model temperatures were still rising. Look again at Fig. 2.2. If the general-public model runs had been allowed to continue out to 30

years, as the Hadley Centre scientists' model runs do, many of them would apparently have reached still higher temperatures, perhaps as high as 12 °C!

How soon will our atmosphere reach a CO_2 level of twice the preindustrial level? The answer depends largely on how much CO_2 we humans put into the atmosphere—a parameter that cannot be predicted by a climate model. Note also that in these models CO_2 does not continue to rise: it is fixed at twice preindustrial levels. Nearly all experts now believe that even if major steps are taken soon to reduce the global production of greenhouse gases, atmospheric CO_2 levels will go well above that level. If CO_2 triples or quadruples, then the expected temperature increase will also increase. No one can say precisely when earth's temperature will increase by any specific value, but the models indicate that it almost surely *will* increase. With scant exceptions, the models show the earth warming, and some of them show the earth warming very quickly and very much.

Is it possible that *all* these model runs are wrong? Yes, because they are variations on a theme. If the basic model conceptualization were wrong in some way, then all the models runs could be wrong, too. Perhaps there is a negative feedback loop that we have not yet recognized. Perhaps the oceans can absorb more CO_2 than we think, or we have missed some other carbon sink (Smith 2002). This is one reason that continued scientific investigation is warranted. But note that Svante Arrhenius and Guy Callendar predicted global warming before anyone ever built a global circulation model (or even had a digital computer). You do not need to have a computer model to predict global warming, and you do not need to have a computer model to know that Earth is, currently, warming.

If climate science stands with or without climate models, then is there any information that would show that climate science is wrong? Yes. Scientists might discover a mistake in their basic physical understanding that showed they had misconceptualized the whole issue. They could discover that they had overestimated the significance of CO_2 and underestimated the significance of some other parameter. But if such mistakes are found, there is no guarantee that correcting them will lead to a more optimistic scenario. It could well be the case that scientists discover neglected factors that show that the problem is worse than we'd

supposed. (Indeed, some scientists now think this is the case: that we have underestimated the cooling or "masking" effect of sulfate aerosols, and therefore the impact of greenhouse gases will be worse if and when China, for example, cleans up its air pollution problems.)

Moreover, there is another way to think about this issue. Contrarians have put inordinate amounts of effort into trying to find something that is wrong with climate science, and despite all this effort, they have come up empty-handed. Year after year, the evidence that global warming is real and serious has only strengthened.[15] Perhaps that is the strongest argument of all. Contrarians have repeatedly tried to falsify the consensus position, and they have repeatedly failed.

Consilience of Evidence

Most philosophers and historians of science agree that there is no iron-clad means to prove a scientific theory. But if science does not provide proof, then what is the purpose of induction, hypothesis testing, and falsification? Most would answer that, in various ways, these activities provide warrant for our views. Do they?

An older view, which has come back into fashion of late, is that scientists look for consilience of evidence. *Consilience* means "coming together," and the term is generally credited to the English philosopher William Whewell, who defined it as the process by which sets of data—independently derived—coincided and came to be understood as explicable by the same theoretical account (Gillispie 1981; Wilson 2000). The idea is not so different from what happens in a legal case. To prove a defendant guilty beyond a reasonable doubt, a prosecutor must present a variety of evidence that holds together in a consistent story. The defense, in contrast, might need to show only that some element of the story is at odds with another to sow reasonable doubt in the minds of the jurors. In other words, scientists are more like lawyers than they might like to admit. They look for independent lines of evidence that hold together.

Do climate scientists have a consilience of evidence? Again the answer is yes. Instrumental records, tree rings, ice cores, borehole data, and coral reefs all point to the same conclusion: things are getting warmer overall. Keith Briffa and Timothy Osborn of the Climate Research Unit of the

University of East Anglia compared Esper's tree-ring analysis with six other reconstructions of global temperature between the years 1000 and 2000 (Briffa and Osborn 2002). All seven analyses agree: temperatures increased dramatically in the late twentieth century relative to the entire record of the previous millennium. Temperatures vary naturally, of course, but the absolute magnitude of global temperatures in the late twentieth century was higher than *any* known temperatures in the previous 1000 years, and many different lines of evidence point in this direction.

Inference to the Best Explanation

The various problems in trying to develop an account of how and why scientific knowledge is reliable have led some philosophers to conclude that the purpose of science is not proof, but explanation. Not just any explanation will do, however; the best explanation is the one that is consistent with the evidence (e.g., Lipton 1991). Certainly, it is possible that a malicious or mischievous deity placed fossils throughout the geological record to trick us into believing organic evolution—perhaps to test our faith?—but to a scientist this is not the best explanation because it invokes supernatural effects, and the supernatural is beyond the scope of scientific explanation. (It might not be the best explanation to a theologian, either, if that theologian was committed to heavenly benevolence.) Similarly, I might try to explain the drift of the continents through the theory of the expanding earth—as some scientists did in the 1950s—but this would not be the best explanation because it fails to explain why the earth has conspicuous zones of compression as well as tension. The philosopher of science Peter Lipton has put it this way: every set of facts has a diversity of possible explanations, but "we cannot infer something simply because it is a possible explanation. It must somehow be the best of competing explanations" (Lipton 2004, 56). (Isaac Newton, in the *Principia Mathematica*, argued that our explanations must invoke causes that we know actually exist—so-called *vera causa*. We might hypothesize that Martians hunted dinosaurs to extinction, thereby explaining their demise, but this would not be an inference to the best explanation, because we have no evidence that Martians exist, but invoking a meteorite can be, because large meteorites do.)

Best is a term of judgment, so it doesn't entirely solve our problem, but it gets us thinking about what it means for a scientific explanation to be the best available—or even just a good one. It also invites us to ask the question, "Best for what purpose?" For philosophers, *best* generally means that an explanation is consistent with all the available evidence (not just selected portions of it), that the explanation is consistent with other known laws of nature and other bodies of accepted evidence (and not in conflict with them), and that the explanation does not invoke supernatural events or causes that by definition cannot be refuted. In other words, *best* can be judged in terms of the various criterion invoked by *all* the models of science discussed above: Is there an inductive basis? Does the theory pass deductive tests? Do the various elements of the theory fit with each other and with other established scientific information? And is the explanation potentially refutable and not invoking unknown, inexplicable, or supernatural causes?

Contrarians have tried to suggest that the climate effects we are experiencing are simply natural variability. Climate does vary, so this is a *possible* explanation. No one denies that. But is it the *best* explanation for what is happening now? Most climate scientists would say that it's not the best explanation. In fact, it's not even a good explanation—because it is inconsistent with much of what we know.

Should we believe that the global increase in atmospheric CO_2 has had a negligible effect even though basic physics tells us it should be otherwise? Should we believe that the correlation between increased CO_2 and increased temperature is just a peculiar coincidence? If there were no theoretical reason to relate them, and if Arrhenius, Callendar, Suess, and Revelle had not predicted that all this would all happen, then one might well conclude that rising CO_2 and rising temperature were merely coincidental. But we have many reasons to believe that there is a causal connection and no good reason to believe that it is a coincidence. Indeed, the only reason we might think otherwise is to avoid committing to action: if this is just a natural cycle in which humans have played no role, then global warming might go away on its own in due course, and we would not have to do spend money or be otherwise inconvenienced by trying to remedy the problem.

2.4 Conclusion

To deny that global warming is real is to deny that humans have become geological agents, changing the most basic physical processes of the earth, and therefore to deny that we bear responsibility for adverse changes that are taking place around us. For centuries, scientists thought that earth processes were so large and powerful that nothing we could do would change them. This was a basic tenet of geological science: that human chronologies were insignificant compared with the vastness of geological time; that human activities were insignificant compared with the force of geological processes. And once they were. But no more. There are now so many of us cutting down so many trees and burning so many billions of tons of fossil fuels that we have become geological agents. We have changed the chemistry of our atmosphere, causing sea level to rise, ice to melt, and climate to change. There is no reason to think otherwise. And, in my view, there is, at this point in history, no excuse for not taking action to prevent the very significant losses that are likely to ensue—indeed, losses that are already becoming evident—if we sit around denying the reality that science has made clear.

Notes

1. Contrast this with the results of the Intergovernmental Panel on Climate Change's *Third* and *Fourth Assessment Reports*, which state unequivocally that average global temperatures have risen (Houghton et al. 2001; Alley et al. 2007).
2. http://royalsociety.org/uploadedFiles/Royal_Society_Content/News_and_Issues/Science_Issues/Climate_change/climate_facts_and_fictions.pdf
3. http://www.science.org.au/policy/climatechange-g8+5.pdf
4. In recent years, climate-change deniers have increasingly turned to non-scientific literature as a way to promulgate views that are rejected by most scientists (see, for example, Deming 2005). http://www.skepticalscience.com/global-warming-scientific-consensus-intermediate.htm

5. An e-mail inquiry to the Thomson Scientific Customer Technical Help Desk produced this reply: "We index the following number of papers in Science Citation Index—2004, 1,057,061 papers; 2003, 1,111,398 papers."
6. The analysis begins in 1993 because that is the first year for which the database consistently published abstracts. Some abstracts initially compiled were deleted from our analysis because the authors of those papers had put "global climate change" in their key words, but their papers were not actually on the subject.
7. This is consistent with the analysis of historian Spencer Weart, who concluded that scientists achieved consensus in 1995 (see Weart 2008).
8. In e-mails that I received after publishing my essay in *Science* (Oreskes 2004), this paper was frequently invoked. It did appear in the sample.
9. According to *Time* magazine, in 2006 a Gallup poll reported that "64 percent of Americans think scientists disagree with one another about global warming" (Americans see a climate problem 2006).
10. Objectivity certainly can be compromised when scientists address charged issues. This is not an abstract concern. It has been demonstrated that scientists who accept research funds from the tobacco industry are much more likely to publish research results that deny or downplay the hazards of smoking than those who get their funds from the National Institutes of Health, the American Cancer Society, or other nonprofit agencies (Bero 2003). On the other hand, there is a large difference between accepting funds from a patron with a clearly vested interest in a particular epistemic outcome and simply trying one's best to communicate the results of one's research clearly and in plain English.
11. Some petroleum companies, such as BP and Shell, have largely refrained from participating in misinformation campaigns (see Browne 1997). Browne began his 1997 lecture by focusing on what he accepted as "two stark facts. The concentration of carbon dioxide in the atmosphere is rising, and the temperature of the Earth's surface is increasing." On the other hand, after an initial flurry of attention caused by Lord Browne's public statements, BP continued to develop its petroleum resources and only to put modest efforts into developing renewables and carbon sequestration technologies. For an analysis of diverse corporate responses, see Van den Hove et al. (2002).
12. For an analysis of one ad, "Weather and Climate," see Environmental Defense (2005). An interesting development in 2003 was that Institutional

Shareholders Services advised ExxonMobil shareholders to ask the company to explain its stance on climate-change issues and to divulge financial risks that could be associated with it. For further information, see https://www.nytimes.com/2017/05/31/business/energy-environment/exxon-shareholders-climate-change.html?mcubz=1.

13. These efforts to generate an aura of uncertainty and disagreement have had an effect. This issue has been studied in detail by academic researchers (see, for example, Boykoff and Boykoff 2004).
14. *Reliable* is a term of judgment. By *reliable basis for action,* I mean that it will not lead us far astray in pursuing our goals, or if it does lead us astray, at least we will be able to look back and say honestly that we did the best we could given what we knew at the time.
15. This is evident when the three IPCC assessments—1990, 1995, 2001—are compared (Houghton et al. 1990, 2001; Bruce et al. 1996; Watson et al. 1996; Metz et al. 2001; Watson 2001; see also Weart 2008).

References

Alley, Richard, Terje Bernsten, Nathaniel Bindoff, Zhenlin Chen, Amnat Chidthaisong, Pierre Fredlingstein, Johnathan Gregory, et al. 2007. AR4 *Climate Change 2007: The Physical Science Basis: Summary for Policy Makers*. Geneva: IPPC Secretariat: Intergovernmental Panel on Climate Change. Accessed 31 Mar 2007.

American Geophysical Union Council. 2003. *Human Impacts of Climate*. Washington, DC: American Geophysical Union.

American Meteorological Society. 2003. Climate Change Research: Issues for the Atmospheric and Related Sciences. *Bulletin of the American Meteorological Society* 84: 508–515. https://www.ametsoc.org/ams/index.cfm/about-ams/ams-statements/archive-statements-of-the-ams/climatechange-research-issues-for-the-atmospheric-and-related-sciences/.

Americans See a Climate Problem. 2006. Time.com, March 26.

Anderegg, William R.L., James W. Prall, Jacob Harold, and Stephen H. Schneider. 2010. Expert Credibility in Climate Change. *Proceedings of the National Academy of Sciences* 107 (27): 12107–12109. https://doi.org/10.1073/pnas.1003187107.

Arctic Council. 2004. *Arctic Climate Impact Assessment.* Oslo: Arctic Council. Accessed 14 Mar 2005.

Bero, Lisa. 2003. Implications of the Tobacco Industry Documents for Public Health and Policy. *Annual Review of Public Health* 24: 267–288. https://doi.org/10.1146/annurev.publhealth.24.100901.140813.

Borick, Christopher, Erick Iachapelle, and Barry Rabe. 2010. Climate Compared: Public Opinion on Climate Change in the United States and Canada. *People*.

Boykoff, M.T. 2011. *Who Speaks for the Climate?: Making Sense of Media Reporting on Climate Change*. Cambridge: Cambridge University Press.

Boykoff, Maxwell T., and Jules M. Boykoff. 2004. Balance as Bias: Global Warming and the US Prestige Press. *Global Environmental Change* 2 (14): 125–136. https://doi.org/10.1016/j.gloenvcha.2003.10.001.

Briffa, Keith R., and Timothy J. Osborn. 2002. Blowing Hot and Cold. (Perspectives: Paleoclimate). *Science* 295 (5563): 2227–2229. https://doi.org/10.1126/science.1069486.

Browne, John. 1997. *Climate Change: The New Agenda*. Stanford University, May 19.

Bruce, James P., Hoesung Lee, and Erik F. Haites. 1996. *Climate Change 1995: Economic and Social Dimensions of Climate Change*. Cambridge: Cambridge University Press.

Cook, John, D. Nuccitelli, S.A. Green, M. Richardson, B. Winkler, R. Painting, R. Way, P. Jacobs, and A. Skuce. 2013. Quantifying the Consensus on Anthropogenic Global Warming in the Scientific Literature. *Environmental Research Letters* 8 (2). https://doi.org/10.1088/1748-9326/8/2/024024.

Cook, J., N. Oreskes, P.T. Doran, W.R. Anderegg, B. Verheggen, E.W. Maibach, J.S. Carlton, S. Lewandowsky, A.G. Skuce, S.A. Green, and D. Nuccitelli. 2016. Consensus on Consensus: A Synthesis of Consensus Estimates on Human-Caused Global Warming. *Environmental Research Letters 11* (4): 048002.

Deming, David. 2005. How 'Consensus' on Global Warming Is Used to Justify Draconian Reform. *Investor's Business Daily*, March 18, A16.

Doran, Peter T., and Maggie Kendall Zimmerman. 2009. Examining the Scientific Consensus on Climate Change. *EOS Transactions* 90: 22–23. https://doi.org/10.1029/2009EO030002.

Edwards, Paul N. 2010. *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. Cambridge: MIT Press.

Environmental Defense. 2005. *Too Slick: Stop ExxonMobil's Global Warming Misinformation Campaign*. https://www.edf.org/sites/default/files/3820_Solutions_07_04.pdf.

Esper, Jan, Edward R. Cook, and Fritz H. Schweingruber. 2002. Low-Frequency Signals in Long Tree-Ring Chronologies for Reconstructing Past Temperature Variability. *Science (New York, N.Y.)* 295 (5563): 2250–2253. https://doi.org/10.1126/science.1066208.

Flannery, Tim Fridtjof. 2006. *The Weather Makers: How Man Is Changing the Climate and What It Means for Life on Earth*. New York: Grove Press.

Fleming, K., P. Johnston, D. Zwartz, Y. Yokoyama, K. Lambeck, and J. Chappell. 1998. Refining the Eustatic Sea-Level Curve Since the Last Glacial Maximum Using Far-and Intermediate-Field Sites. *Earth and Planetary Science Letters 163* (1): 327–342.

Freudenburg, William R., and Violetta Muselli. 2010. Global Warming Estimates, Media Expectations, and the Asymmetry of Scientific Challenge. *Global Environmental Change, Governance, Complexity and Resilience* 20 (3): 483–491. https://doi.org/10.1016/j.gloenvcha.2010.04.003.

Gelbspan, R. 1997. *The Heat Is On: The High Stakes Battle Over Earth's Threatened Climate*. Reading: Addison Wesley Publishing Company.

Gelbspan, Ross. 2005. *Boiling Point: How Politicians, Big Oil and Coal, Journalists, and Activists Are Fueling the Climate Crisis-and What We Can Do to Avert Disaster*. New York: Basic Books.

Gillispie, Charles. 1975. *Dictionary of Scientific Biography. Vol._12: Ibn Rushd—Jean-Servais Stas*. Vol. 12. New York: Scribner.

Gillispie, Charles Coulston. 1981. *Dictionary of Scientific Biography*. New York: Scribner.

Hansen, James, Andrew Lacis, Reto Ruedy, and Makiko Sato. 1992. Potential Climate Impact of Mount Pinatubo Eruption. *Geophysical Research Letters* 19: 215–218. https://doi.org/10.1029/91GL02788.

Harrison, Paul, and Fred Pearce. 2000. *AAAS Atlas of Population and Environment*. Berkeley: University of California Press.

Hassol, Susan Joy. 2008. Improving How Scientists Communicate About Climate Change. *EOS Transactions* 89: 106–107. https://doi.org/10.1029/2008EO110002.

Hempel, Carl G. 1965. *Aspects of Scientific Explanation, and Other Essays in the Philosophy of Science*. New York: Free Press.

Holland, M.M., and C.M. Bitz. 2003. Polar Amplification of Climate Change in Coupled Models. *Climate Dynamics* 21 (3/4): 221–232. https://doi.org/10.1007/s00382-003-0332-6.

Houghton, John, G.J. Jenkins, and J.J. Ephraums, eds. 1990. *Scientific Assessment of Climate Change: Report of Working Group 1. Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press.

Houghton, John, Meria Filho, B.A. Callander, N. Harris, A. Katteberg, and K. Maskell, eds. 1995. *Climate Change 1995: The Science of Climate Change: Report of Working Group 1. Intergovernmental Panel on Climate Change*. Cambridge/New York: Cambridge University Press.

Houghton, John, Y. Ding, D.J. Griggs, M. Nouger, P.J. van der Linden, X. Dai, K. Maskell, and C.A. Johnson. 2001. *Climate Change 2001: The Scientific Basis: Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press.

Intergovernmental Panel on Climate Change (IPCC). 2005. *About IPCC*, February 7.

Kerr, Richard A. 1993. Pinatubo Global Cooling on Target. *Science* 259: 594. https://doi.org/10.1126/science.259.5095.594.

Kolbert, Elizabeth. 2006. *Field Notes from a Catastrophe: Man, Nature, and Climate Change*. New York: Bloomsbury.

Krosnick, J.A., A.L. Holbrook, L. Lowe, and P.S. Visser. 2006. The Origins and Consequences of Democratic Citizens' Policy Agendas: A Study of Popular Concern About Global Warming. *Climatic Change 77* (1): 7–43.

Leiserowitz, Anthony, Edward Maibach, and Connie Roser-Renouf. 2011. *Global Warming's Six Americas in March 2012 and November 2011*. Yale Program on Climate Change Communication. Accessed 7 June 2017.

Lipton, Peter. 1991. *Inference to the Best Explanation*. London: Routledge.

———. 2004. *Inference to the Best Explanation*. London: Routledge.

Lorenzoni, Irene, and Nick F. Pidgeon. 2006. Public Views on Climate Change: European and USA Perspectives. *Climatic Change* 77 (1–2): 73–95. https://doi.org/10.1007/s10584-006-9072-z.

Manabe, Syukuro, and R.J. Stouffer. 1980. Sensitivity of a Global Climate Model to an Increase of CO_2 Concentration in the Atmosphere. *Journal of Geophysical Research* 85 (C10): 5529–5554.

———. 1994. Multiple-Century Response of a Coupled Ocean-Atmosphere Model to an Increase of Atmospheric Carbon Dioxide. *Journal of Climate; (United States)* 7: 5.

McCarthy, James J., Osvaldo F. Canziani, Neil A. Leary, David Dokken, and Kasey S. White. 2001. *Climate Change 2001: Impacts, Adaptation, and Vulnerability*. Cambridge: Cambridge University Press.

Metz, Bert, Ogunlade Davidson, Rob Swart, and Jiahua Pan. 2001. *Climate Change 2001: Mitigation*. Cambridge University Press.

Mooney, Chris, and Basic Books. 2006. *The Republican War on Science*. New York: Basic Books.

Moser, Susanne C., and Lisa Dilling. 2004. Making Climate Hot: Communicating the Urgency and Challenge of Global Climate Change. *Environment: Science and Policy for Sustainable Development* 46 (10): 32–46.

Muller, Richard A., Robert Rohde, Robert Jacobsen, Elizabeth Muller, Saul Perlmutter, Arthur Rosenfeld, Jonathan Wurtele, Donald Groom, and Charlotte Wickham. 2013. A New Estimate of the Average Earth Surface Land Temperature Spanning 1753 to 2011. *Geoinformatics & Geostatistics: An Overview* (January 2, 2014). https://doi.org/10.4172/2327-4581.1000101.

National Academy of Sciences, Committee on the Science of Climate Change. 2001. *Climate Change Science; An Analysis of Some Key Questions*. Washington, DC: National Academy Press.

Oreskes, Naomi. 2004. Beyond the Ivory Tower. The Scientific Consensus on Climate Change. *Science (New York, N.Y.)* 306 (5702).

Oreskes, Naomi, and Erik M. Conway. 2012. *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming*. London: Bloomsbury.

Oreskes, Naomi, David A. Stainforth, and Leonard A. Smith. 2010. Adaptation to Global Warming: Do Climate Models Tell Us What We Need to Know? *Philosophy of Science* 77 (5): 1012–1028. https://doi.org/10.1086/657428.

Popper, Karl Raimund. 1959. *The Logic of Scientific Discovery*. New York: Basic Books.

Price, Derek J. De Solla. 1986. *Little Science, Big Science ... and Beyond*. New York: Columbia University Press.

Revelle, Roger. 1965. Atmospheric Carbon Dioxide. In *Restoring the Quality of Our Environment: A Report of the Environmental Pollution Panel*, 111–133. Washington, DC: President's Science Advisory Committee.

Roach, John. 2004. The Year Global Warming Got Respect. *National Geographic*.

Smith, Leonard A. 2002. What Might We Learn from Climate Forecasts? *Proceedings of the National Academy of Sciences of the United States of America* 99 (Suppl. 1): 2487–2492.

Somerville, Richard C.J., and Susan Joy Hassol. 2011. Communicating the Science of Climate Change. *Physics Today* 64 (10): 48–53.

Soon, Willie, Sallie Baliunas, Sherwood Idso, Kirill Kondratyev, and Eric Posmentier. 2001. Modeling Climatic Effects of Anthropogenic Carbon Dioxide Emissions: Unknowns and Uncertainties. *Climate Research* 18 (3): 259. https://doi.org/10.3354/cr018259.

Soon, Willie, Sallie Baliunas, Sherwood B. Idso, Kirill Ya. Kondratyev, and Eric S. Posmentier. 2002. Modeling Climatic Effects of Anthropogenic Carbon Dioxide Emissions: Unknowns and Uncertainties. Reply to Risbey (2002). *Climate Research* 22 (2): 187.

Stainforth, David, Tolu Aina, C. Christensen, M. Collins, N. Faull, D.J. Frame, J.A. Kettleborough, et al. 2005. Uncertainty in Predictions of the Climate Response to Rising Levels of Greenhouse Gases. *Nature* 433 (7024): 403–406. https://doi.org/10.1038/nature03301.

Supran, G., and N. Oreskes. 2017. Assessing ExxonMobil's Climate Change Communications (1977–2014). *Environmental Research Letters 12* (8): 084019.

van den Hove, Sybille, Marc Le Menestrel, and Henri-Claude de Bettignies. 2002. The Oil Industry and Climate Change: Strategies and Ethical Dilemmas. *Climate Policy* 1: 3. https://doi.org/10.3763/cpol.2002.0202.

Watkin, Daniel. 2004. Roman Catholic Priests' Group Calls for Allowing Married Clergy Members. *New York Times*, April 28.

Watson, Robert T. 1996. Groupe d'experts intergouvernamental sur l'évolution du clima. In *Climate Change 1995: Impacts, Adaptations and Mitigation of Climate Change: Scientific-Technical Analyses*, ed. Organisation météorologique mondiale and Programme des Nations Unies pour l'environnement. Cambridge/New York: Cambridge University Press.

Watson, R.T., M.C. Zinyowera, and R.H. Moss. 1996. *Climate Change 1995. Impacts, Adaptations and Mitigation of Climate Change: Scientific-Technical Analyses*. Cambridge/New York: Cambridge University Press.

Watson, R.T., and D.L. Albritton, eds. 2001. *Climate Change 2001: Synthesis Report: Third Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press.

Weart, Spencer R. 2008. *The Discovery of Global Warming*. Cambridge: Harvard University Press.

Wickham, Charlotte, Robert Rohde, Richard A. Muller, Jonathan Wurtele, Judith A. Curry, Donald Groom, Robert Jacobsen, S. Perimutter, Arthur Rosenfeld, and S. Mosher. 2013. Influence of Urban Heating on the Global Temperature Land Average Using Rural Sites Identified from MODIS Classifications. *Geoinformatics & Geostatistics: An Overview* 1 (2). https://doi.org/10.4172/2327-4581.1000104.

Wilson, Edward O. 2000. *Consilience: The Unity of Knowledge*. New York: Alfred A. Knopf.

3

Satellite Data and Climate Models

Elisabeth A. Lloyd

3.1 Context

A paper by John Christy, David Douglass, Benjamin D. Pearson, and S. Fred Singer was published online in December 2007 in the *International Journal of Climatology* (print version 2008). John Christy was the "primary developer of a satellite-based temperature record which suggests that there has been minimal warming of Earth's lower atmosphere since 1979," according to the blog Real Climate,[1] while S. Fred Singer was a longtime science skeptic and tobacco/cancer denialist.[2] Thus, this paper was coauthored by some climate "skeptics" (or "denialists").

The Douglass, Christy, et al. paper claimed to establish that climate models are inconsistent with data from satellites and weather balloons. Here is the beginning of the conclusion of their paper:

E.A. Lloyd (✉)
History and Philosophy of Science and Medicine Department, Indiana University, Bloomington, IN, USA

E.A. Lloyd, E. Winsberg (eds.), *Climate Modelling*,
https://doi.org/10.1007/978-3-319-65058-6_3

> Models are very consistent, as this article demonstrates, in showing a significant difference between surface and tropospheric trends, with tropospheric temperature trends warming faster than the surface. (Douglass et al. 2008, p. 1700)

And this is the essence of what is at stake: climate models predict that there is extra warming of the tropospheric level—the level of atmosphere above the surface—in the tropical region of the earth. The question then becomes: is this what is found in observational data? And for many years, the satellite data, supervised and processed by John Christy and Roy Spencer, did *not* find such tropospheric warming (Spencer and Christy 1990; Christy and Spencer 2003; Christy et al. 2000; Spencer et al. 2006). It thus seemed that the models were getting something wrong, and Christy and coauthors wanted to conclude that the models could not be trusted…. Or was it possibly the satellite data getting something wrong? The satellite data had been challenged repeatedly (Hurrell and Trenberth 1997; Santer et al. 2003a, b; see discussion in Lloyd Chap. 6). The weather balloons seemed to agree with the satellite data, but the weather balloons were never designed to collect climate data, as opposed to short-term weather data, nor were they considered trustworthy by many climate modelers (for the history and discussion, see Thorne et al. 2011; Lanzante et al. 2003; see discussion in Lloyd Chap. 6).

Thus, the Douglass et al. paper focused on the temperatures in the tropics, particularly the lower troposphere, where the temperatures measured by the satellites seemed to contradict those predicted by many climate models (2008). Christy and the coauthors on this paper produced a "robust statistical test" to compare climate model results with Christy and Spencer's interpretations of the satellite data. When this "robust statistical test" was applied, it showed that the models were "significantly different from observations" (2008).

3.2 Fact Sheet

This new "robust statistical test" that they set up in this paper is discussed extensively in the Santer et al. (2008a) paper (Chap. 5) and explained by Santer et al. in their "Fact Sheet" (Chap. 4). One question that Santer

et al. (2008a, b) asked about the "robust statistical test" set up by Douglass et al., was how well it performed under controlled conditions, using random data with known statistical properties, a standard test. In brief, it failed such a challenge spectacularly, and should never have been used at all in a scientific paper, as Santer et al. explain in the "Fact Sheet." But let us proceed to what Douglass et al. would like to conclude from their use of these tests:

> These [model results] are compared with several equally robust updated estimates of trends from observations which disagree with trends from the models. The last 25 years constitute a period of more complete and accurate observations and more realistic modelling efforts. Yet the models are seen to disagree with the observations. We suggest, therefore, that projections of future climate based on these models be viewed with much caution. (2008, p. 1700)

As we can see here, Douglass et al. compare "complete and accurate" observations with climate models, finding disagreement. As Santer et al. (2008a) will show, there is, in fact, little disagreement between updated data sets and contemporary models, when analyzed fairly. Santer et al., in their "Fact Sheet" (Chap. 4), detail some of the errors used to achieve the Douglas et al. conclusion, and all of the claims there are backed up by the Santer et al. paper published alongside it in this collection (Chap. 5).

For the big picture, it is good to know a bit more about the context of climate models and data sets at the time that Douglass et al. and Santer et al. were writing. At the time, three alternate versions of the satellite temperature record produced by alternate teams of researchers using the same raw satellite data were available, two producing substantially more warming of the lower troposphere than the Christy and Spencer interpretation of that same satellite data (Mears et al. 2003; Vinnikov and Grody 2003; Vinnikov et al. 2006; see Karl et al. 2006 for analysis; see Lloyd Chap. 6 for discussion). In other words, the interpretation of the satellite data that Douglass, Christy, et al. used in their paper had already been challenged and an agreement come to, it was thought (see Karl et al. 2006), on which Christy signed off (see discussion in Lloyd Chap. 6).

3.3 Larger Social Context and Santer et al.'s Accomplishments

It is also crucial to understand that the Douglass et al. paper immediately attracted a great deal of media and political attention. It was claimed to represent an "inconvenient truth" and to prove that "Nature, not humans, rules the climate." In Santer's words on a Real Climate blogpost: "These statements were absurd. No single study can overturn the very large body of scientific evidence supporting 'discernible human influence' findings. [This was a reference to the IPCC findings at the time.] Nor does any individual study provide the sole underpinning for the conclusion that human activities are influencing global climate."[3]

Santer and a host of other climate scientists, including leaders in satellite data, weather balloon data, modeling, and statistical analysis, felt they needed to respond to this new paper. While the errors in the Douglass et al. paper were obvious, "it required a substantial amount of new and original work to repeat the statistical analysis properly," according to Santer (2010; http://www.realclimate.org/index.php/archives/2010/02/close-encounters-of-the-absurd-kind/#sthash.2OEvv0sp.dpuf Accessed on June 6, 2017).

The Santer et al. paper went far beyond what Douglass et al. (2008) had done: "We looked at the sensitivity of model-versus-data comparisons to the choice of statistical test, to the test assumptions, to the number of years of record used in the tests, and to errors in the computer model estimates of year-to-year temperature variability." Again, the Douglass et al. paper showed no evidence that they had considered any of these important issues before making their highly publicized claims (http://www.realclimate.org/index.php/archives/2010/02/close-encounters-of-the-absurd-kind/#sthash.2OEvv0sp.dpuf).

As is clear from Chaps. 3, 4, and 5, the Santer et al. analysis showed that the models and the observational data were clearly consistent, exactly contrary to the claims of Douglass et al. (2008). The models were refined, and the previous data corrected, to yield a confluence of the two, models and datasets, that would not undermine the predictive power of the models on the bases claimed by Christy, Singer, Douglas, and Pearson. The Santer et al. research was published on October 10,

2008, online. On November 15, 2008, the Douglass et al. and the Santer et al. papers both appeared in the same print version of the *International Journal of Climatology*. Climate "skeptics" and "deniers," unhappy with the Santer et al. paper, later accused Santer and coauthors of manipulating the journal to publish the papers at the same time. But this is false, totally unfounded in fact.

As Santer et al. make clear in their "Fact Sheet" (Chap. 4), created to frame and introduce the Santer et al. article (Chap. 5), the Douglass et al. (2008) conclusions were unfounded, because of the serious problems with both the observations and the statistical test they developed. We now turn to Chap. 4, "Fact Sheet for 'Consistency of Modelled and Observed Temperature Trends in the Tropical Troposphere,' by B.D. Santer et al.," written by Ben Santer et al.

Notes

1. "Close Encounters of the Absurd Kind," Ben Santer, February 24, 2010, Real Climate: Climate science from climate scientists Blog. http://www.realclimate.org/index.php/archives/2010/02/close-encounters-of-the-absurd-kind/ Accessed on August 7, 2015
2. See Naomi Oreskes and Erik M. Conway's (2011) discussion, in *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming*, of S. Fred Singer's roles in skepticism about science and regulation since the 1960s.
3. See more at: http://www.realclimate.org/index.php/archives/2010/02/close-encounters-of-the-absurd-kind/#sthash.2OEvv0sp.dpuf. Accessed on June 6, 2017.

References

Christy, J.R., and R.W. Spencer. 2003. *Global Temperature Report: 1978–2003*. Huntsville: Earth System Science Center, University of Alabama in Huntsville.

Christy, J.R., R.W. Spencer, and W.D. Braswell. 2000. MSU Tropospheric Temperatures: Dataset Construction and Radiosonde Comparisons. *Journal of Atmospheric and Oceanic Technology* 17: 1153–1170.

Douglass, D.H., J.R. Christy, B.D. Pearson, and S.F. Singer. 2008. A Comparison of Tropical Temperature Trends with Model Predictions. *International Journal of Climatology* 28: 1693–1701.

Hurrell, J., and K. Trenberth. 1997. Response to "How Accurate are Satellite 'Thermometers'?". *Nature* 389: 342–343.

Karl, T.R., S.J. Hassol, C.D. Miller, and W.L. Murray. 2006. *Temperature Trends in the Lower Atmosphere: Steps for Understanding and Reconciling Differences. A Report by the U.S. Climate Change Science Program and the Subcommittee on Global Change Research*. Asheville: National Oceanic and Atmospheric Administration, National Climatic Data Center.

Lanzante, J.R., S.A. Klein, and D.J. Seidel. 2003. Temporal Homogenization of Monthly Radiosonde Temperature Data. Part I: Methodology. *Journal of Climate* 16: 224–240.

Mears, C., M.C. Schabel, and F.J. Wentz. 2003. A Reanalysis of the MSU Channel 2 Tropospheric Temperature Record. *Journal of Climate* 16: 3650–3664.

Oreskes, Naomi, and Erik M. Conway. 2011. *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming*. London: Bloomsbury Press.

Santer, B.D. 2010. Close Encounters of the Absurd Kind, February 24. Real Climate: Climate Science from Climate Scientists Blog. http://www.realclimate.org/index.php/archives/2010/02/close-encounters-of-the-absurd-kind/. Accessed 7 Aug 2015.

Santer, B.D., T.M.L. Wigley, G.A. Meehl, M.F. Wehner, C. Mears, M. Schabel, et al. 2003a. Influence of Satellite Data Uncertainties on the Detection of Externally Forced Climate Change. *Science* 300: 1280–1284.

———. 2003b. Response to Christy and Spencer 2003. *Science* 301: 1047–1049.

Santer, B.D., P.W. Thorne, L. Haimberger, K.E. Taylor, T.M.L. Wigley, J.R. Lanzante, et al. 2008a. Consistency of Modeled and Observed Temperature Trends in the Tropical Troposphere. *International Journal of Climatology* 28: 1703–1722.

———. 2008b. Fact Sheet for Consistency of Modelled and Observed Temperature Trends in the Tropical Troposphere, by B.D. Santer et al. https://www.llnl.gov/file/27178/download?token=DL2W9BH2

Spencer, R.W., and J.R. Christy. 1990. Precise Monitoring of Global Temperature Trends from Satellites. *Science* 247: 1558–1562.

Spencer, R.W., J.R. Christy, W.D. Braswell, and W.B. Norris. 2006. Estimation of Tropospheric Temperature Trends from MSU Channels 2 and 4. *Journal of Atmospheric and Oceanic Technology* 23: 417–423.

Thorne, P.W., J.R. Lanzante, T.L. Peterson, D.J. Seidel, and K.P. Shine. 2011. Tropospheric Temperature Trends: History of an Ongoing Controversy. *Wiley Interdisciplinary Reviews: Climate Change* 2: 66–88.

Vinnikov, K.Y., and N.C. Grody. 2003. Global Warming Trend of Mean Tropospheric Temperature Observed by Satellites. *Science* 302: 269–272.

Vinnikov, K.Y., N.L. Grody, A. Robock, R.J. Stouffer, P.D. Jones, and M.D. Goldberg. 2006. Temperature Trends at the Surface and in the Troposphere. *Journal of Geophysical Research* 111: D03106.

4

Fact Sheet for "Consistency of Modeled and Observed Temperature Trends in the Tropical Troposphere"

Ben Santer, Peter Thorne, Leo Haimberger, Karl Taylor, Tom Wigley, John Lanzante, Susan Solomon, Melissa Free, Peter Gleckler, Phil Jones, Tom Karl, Steve Klein, Carl Mears, Doug Nychka, Gavin Schmidt, Steve Sherwood, and Frank Wentz

4.1 QUESTION 1: What is the scientific context for the research published in the Santer et al. *International Journal of Climatology* paper?

Our paper compares modeled and observed atmospheric temperature changes in the tropical troposphere.[1] We were interested in this region because of an apparent inconsistency between computer model results

This paper was published online in the *International Journal of Climatology* during the week of Oct. 6–10, 2008.

B. Santer (✉) • P. Thorne • L. Haimberger • K. Taylor • T. Wigley • J. Lanzante • S. Solomon • M. Free • P. Gleckler • P. Jones • T. Karl • S. Klein • C. Mears • D. Nychka • G. Schmidt • S. Sherwood • F. Wentz
Lawrence Livermore National Laboratory, University of California, Livermore, CA, USA

E.A. Lloyd, E. Winsberg (eds.), *Climate Modelling*,
https://doi.org/10.1007/978-3-319-65058-6_4

and observations. Since the late 1960s, scientists have performed experiments in which computer models of the climate system are run with human-caused increases in atmospheric concentrations of greenhouse gases (GHGs).[2] These experiments consistently showed that increases in atmospheric concentrations of GHGs should lead to pronounced warming, both at the Earth's surface and in the troposphere. The models also predicted that in the tropics, the warming of the troposphere should be larger than the warming of the surface.[3]

Observed estimates of surface temperature changes are in good agreement with computer model results, confirming the predicted surface warming.[4] Until several years ago, however, most available estimates of tropospheric temperature changes obtained from satellites and weather balloons (radiosondes) implied that the tropical troposphere had actually cooled slightly over the last 20–30 years (in sharp contrast to the computer model predictions, which show tropospheric warming).

For nearly a decade, this apparent disconnect between models and reality has been used by some scientists and politicians to argue that:

- The surface thermometer record is wrong
- The Earth has not experienced any surface or tropospheric warming since the beginning of satellite measurements of atmospheric temperature in 1979
- Human-caused changes in greenhouse gases have no effect on climate
- Computer models have no skill in simulating the observed temperature changes in the tropics, and therefore cannot be used to predict the climatic "shape of things to come" in response to further increases in greenhouse gases

Our paper attempts to determine whether there is indeed a real and statistically significant discrepancy between modeled and observed temperature changes in the tropics, as was claimed in a paper published online in December 2007 in the *International Journal of Climatology*. As discussed in QUESTION 9, we find that this claim is incorrect.

4.2 QUESTION 2: What arguments were made to support this claim?

David Douglass, John Christy, Benjamin Pearson, and S. Fred Singer[5] devised a statistical test to determine whether modeled and observed atmospheric temperature trends in the tropical troposphere were significantly different. They applied this test in several different ways. First, they considered temperature trends in two different layers of the troposphere (the lower troposphere and the mid- to upper troposphere). In each of these layers, their test suggested that the modeled warming trends were larger than and significantly different from the warming trends estimated from satellite data. Second, they compared trends in the temperature differences between the surface and the lower troposphere—a measure of the "differential warming" of the surface and lower atmosphere. Once again, their test pointed toward the existence of statistically significant differences in modeled and observed trends.

The bottom-line conclusion of Douglass et al. was that "models and observations disagree to a statistically significant extent." As discussed in QUESTIONS 6–8, we show that this statistical test is flawed and that the conclusions reached by Douglass et al. are incorrect.

4.3 QUESTION 3: But hadn't the scientific community already resolved this issue?

The community had already achieved a partial resolution of this issue in a 2006 Report issued by the U.S. Climate Change Science Program (CCSP).[6] The CCSP Report concluded that, when one examined temperature changes at the global scale, newer satellite and weather balloon datasets showed "no significant discrepancy" between surface and tropospheric warming trends, and were therefore consistent with computer model results. But the same CCSP Report noted that it was not possible (in 2006) to reconcile modeled and observed temperature changes in the tropics, where "most observational datasets show more

warming at the surface than in the troposphere, while most model runs have larger warming aloft than at the surface."

The CCSP Report relied almost exclusively on published literature. At the time of its publication in 2006, there were no peer-reviewed studies on the formal statistical significance of differences between modeled and observed tropical temperature trends. The Douglass et al. paper attempted to assess the statistical significance of the model-versus-observed tropical trend differences noted in the CCSP Report.

4.4 QUESTION 4: What was the thrust of your new research?

Our primary goal was to determine whether the findings of Douglass et al. were sound. As noted above, Douglass et al. reported that "models and observations disagree to a statistically significant extent." They interpreted their results as evidence that computer models are seriously flawed and that the projections of future climate change made with such models are untrustworthy. If Douglass et al. were right, this would imply that there was some fundamental flaw—not only in all state-of-the-art climate models, but also in our basic theoretical understanding of how the climate system should respond to increases in GHGs. We wanted to know whether such a fundamental flaw really existed.

4.5 QUESTION 5: What specific issues did you focus on?

We focused on two issues. First, Douglass et al. claimed that they had applied a "robust statistical test" to identify statistically significant differences between modeled and observed temperature trends. We sought to understand whether their test was indeed "robust" and appropriate. Second, Douglass et al. claimed to be using the "best available updated observations" for their study. We did not believe that this claim was accurate.

We decided to check their analysis by applying a variety of different statistical tests to modeled and observed temperature trends, and by employing temperature data from more recent observational datasets—datasets that were either unavailable to Douglass et al. at the time of their study, or which were available, but had not been used by them.

4.6 QUESTION 6: What did you learn about the appropriateness of the Douglass et al. test?

We found that there was a serious flaw in the "robust statistical test" that Douglass et al. had used to compare models and observations. Their test ignored the effects of natural climate "noise" on observed temperature trends, and the resulting statistical uncertainty in estimating the "signal component" of these trends (see QUESTION 7 for a definition of the "signal component").

4.7 QUESTION 7: Why was this a problem?

We know that in the real world, changes in temperatures are due to a combination of human effects and natural factors. The "natural factors" can be things like volcanic eruptions or changes in the Sun's energy output. Another type of "natural factor" is referred to as "internal variability", which is unrelated to changes in the Sun or volcanic dust, and involves phenomena like El Niños, La Niñas, and other natural climate oscillations. In the tropics in particular, El Niños and La Niñas have a substantial effect on surface and atmospheric temperature. They introduce climate "noise", which complicates the separation of human and natural effects on temperature.

Douglass et al. effectively assumed that the observed surface and tropospheric temperature trends were perfectly known and that these trends were purely due to human-caused changes in greenhouse gases.[7] The inappropriateness of this assumption is immediately obvious by looking

at any observed temperature time series, such as the surface and tropospheric temperature time series shown in Fig. 4.1.

Figure 4.1 illustrates that both tropical surface and tropospheric temperatures have gradually warmed since 1979. Superimposed on this overall warming is climate "noise", which in this case arises primarily from El Niños and La Niñas. When temperatures are averaged over the tropics (and indeed, over the globe), El Niños tend to warm the surface and

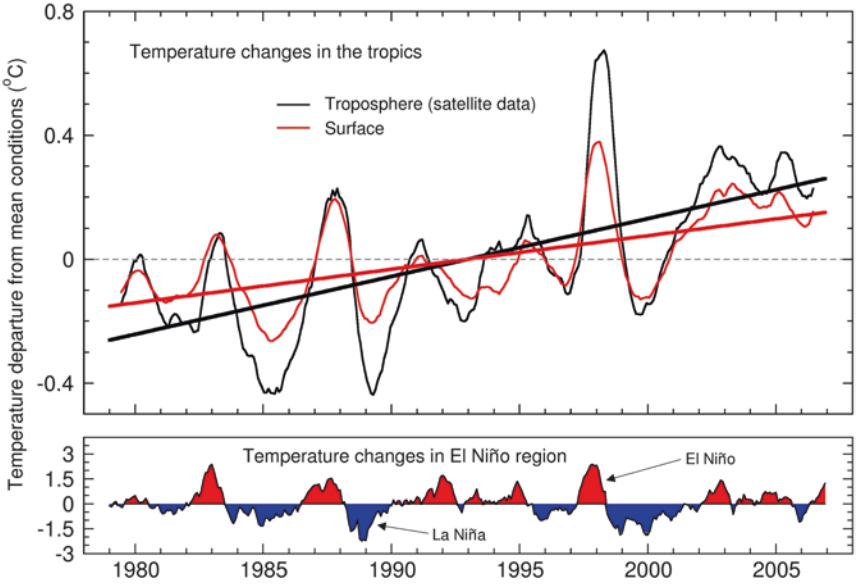

Fig. 4.1 Estimates of observed temperature changes in the tropics (30 °N–30 °S). Changes are expressed as departures from average conditions over 1979–2006. The *top panel* shows results for the surface[13] and lower troposphere.[13] The *thin red* and *black lines* in the *top panel* are 12-month running averages of the temperature changes for individual months. The *thick straight lines* are trends that have been fitted to the time series of surface and tropospheric temperature changes. The warming trend is larger in the tropospheric temperature data than in the surface temperature record, in accord with computer model results. The *bottom panel* shows a commonly used index of El Niño and La Niña activity, consisting of sea surface temperature changes averaged over the so-called Niño 3.4 region of the tropical Pacific. The *bottom panel* shows that much of the year-to-year variability in surface and lower tropospheric temperatures is related to changes in El Niños and La Niñas

lower atmosphere, and La Niñas tend to cool these regions.[8] As is visually obvious, El Niños and La Niñas introduce considerable year-to-year variability in surface and tropospheric temperature.

Because of the climate noise introduced by El Niños and La Niñas, there is uncertainty in estimating any underlying temperature trend, such as that arising from slow, human-caused increases in GHGs. In the real world and in many model simulations of twentieth-century climate change, this underlying trend in temperature is not caused by GHG increases alone—it results from the combined changes in GHGs and other external forcing factors, and is partly masked by climate noise.

The underlying "signal trend" is what we really want to compare in climate models and observations. Any meaningful statistical test of the differences between modeled and observed temperature trends must therefore account for the statistical uncertainty in estimating this "signal trend" from noisy observational data. The Douglass et al. test did not account for this uncertainty.

4.8 QUESTION 8: What were the consequences of the flaw in the Douglass et al. test?

The primary consequence was that Douglass et al. reached incorrect conclusions about the true statistical significance of differences between modeled and observed temperature trends in the tropics. When we applied modified versions of their test—versions that properly accounted for uncertainties in estimating the "signal component" of observed temperature trends—we obtained results that were strikingly different from theirs. Like Douglass et al., we applied our tests to modeled and observed temperature trends:

- In individual layers of the troposphere
- In the trend difference between surface and tropospheric warming rates

Unlike Douglass et al., however, we found that most of our tests involving temperature trends in individual layers of the troposphere did not show statistically significant differences between models and observations. This result was relatively insensitive to which model or satellite dataset we chose for the trend comparison.

The situation was a little more complex for tests involving the trend difference between surface and tropospheric warming rates. In this case, the statistical significance of the differences between models and observations was sensitive to our choice of observational datasets. When we used a satellite-based tropospheric temperature dataset developed at Remote Sensing Systems (RSS) in Santa Rosa, California, we found that the warming in the tropical troposphere was always larger than the warming at the surface.[9] This behavior is consistent with the behavior of the climate models and with our understanding of the physical processes that govern tropospheric temperature profiles. It is contrary to the findings of Douglass et al.

However, when we used a satellite-based tropospheric temperature dataset developed at the University of Alabama at Huntsville (UAH),[10] the tropospheric warming was less than the surface warming. But even when we employed UAH data, our statistical test showed that the observed difference between surface and tropospheric warming trends was not always significantly different from the trend difference in model simulations. Whether or not trend differences were statistically significant was dependent on the choice of model and the choice of observed surface dataset used in the test.[11]

4.9 QUESTION 9: So what is the bottom line of your study?

The bottom line is that we obtained results strikingly different from those of Douglass et al. The "robust statistical test" that they used to compare models and observations had at least one serious flaw—its failure to account for any uncertainty in the "signal component" of observed temperature trends (see QUESTION 7). This flaw led them to reach incor-

rect conclusions. We showed this by applying their test to randomly generated data with the same statistical properties as the observed temperature data, but without any underlying "signal trend." In this "synthetic data" case, we knew that significant differences in temperature trends could occur by chance only, and thus would happen infrequently. When we applied the Douglass et al. test, however, we found that even randomly generated data showed statistically significant trend differences much more frequently than we would expect on the basis of chance alone. A test that fails to behave properly when used with random data—when one knows in advance what results to expect—cannot be expected to perform reliably when applied to real observational and model data.

4.10 QUESTION 10: Final question: have you reconciled modeled and observed temperature trends in the tropics?

We've gone a long way toward such a reconciliation. There are at least two reasons for this.[12] The first reason is that we have now applied appropriate statistical tests for comparing modeled and observed temperature trends in the tropics. Unlike the Douglass et al. test, our test properly accounts for uncertainty in estimating the "signal component" of observed temperature trends. Results from these more appropriate tests do not support the claim that there are fundamental, pervasive, and statistically significant differences between modeled and observed tropical temperature trends. This claim is not tenable for temperature trends in individual layers of the troposphere. Nor is it tenable for the differences in the warming rates of the surface and troposphere.

Second, we now have many more estimates of recent temperature changes. These have been produced by a number of different research groups, often using completely independent methods.

Research groups involved in the development of newer sea surface temperature datasets have reported improvements in the treatment of information from buoys and satellites. This has led to slightly reduced estimates of the warming of the tropical ocean surface (relative to the

warming in the earlier surface temperature datasets used by Douglass et al. and in the CCSP Report). Additionally, newly developed satellite and radiosonde datasets now show larger warming of the tropical troposphere than was apparent in the datasets used by Douglass et al. The enhanced tropospheric warming is due to improvements in our ability to identify and adjust for biases introduced by changes over time in the instruments used to measure temperature.[13]

Access to such a rich variety of independently produced datasets has provided us with a valuable perspective on the inherent uncertainty in observed estimates of recent climate change. Based on our current best estimates of these observational uncertainties, there is no fundamental discrepancy between modeled and observed tropical temperature trends. In fact, many of the recently developed observational datasets now show tropical temperature changes that are larger aloft than at the surface—behavior that is entirely consistent with climate model results.

One of the lessons from this work is that even with improved datasets, there are still important uncertainties in observational estimates of recent tropospheric temperature trends. These uncertainties may never be fully resolved, and are partly a consequence of historical observing strategies, which were geared toward weather forecasting rather than climate monitoring. We should apply what we learned in this study toward improving existing climate monitoring systems, so that future model evaluation studies are less sensitive to observational ambiguity.

Notes

1. The troposphere is the lowest layer of the atmosphere, where most weather phenomena take place. In the tropics, the troposphere extends from the surface to a height of about 10 miles (16 km) above the Earth's surface.
2. Both climate models and the experiments performed with them have become more realistic over time. Since the mid-1990s, many climate model experiments have incorporated not only human-caused changes in GHGs, but also changes in other "forcing agents" that have effects on global or regional climate. Examples include human-caused changes in

various aerosol particles (such as sulfate and soot aerosols), and natural changes in the Sun's energy output and the amount of volcanic dust in the atmosphere.

3. This prediction of larger warming aloft than at the surface holds for all factors that tend to warm the surface of the Earth—it is not unique to human-caused changes in GHGs.
4. This agreement between models and observations was also found for complex geographical patterns of surface temperature changes—not simply for trends in temperature changes averaged over very large areas (such as the tropics).
5. Douglass DH, Christy JR, Pearson BD, Singer SF. 2007. A comparison of tropical temperature trends with model predictions. *International Journal of Climatology* **27**: https://doi.org/10.1002/joc.1651.
6. Karl TR, Hassol SJ, Miller CD, Murray WL (*eds*). 2006. *Temperature Trends in the Lower Atmosphere: Steps for Understanding and Reconciling Differences*. A Report by the U.S. Climate Change Science Program and the Subcommittee on Global Change Research. National Oceanic and Atmospheric Administration, National Climatic Data Center, Asheville, NC, 164 pp.
7. In their paper, Douglass et al. claim to be testing "the proposition that greenhouse model simulations and observations can be reconciled." The model simulations of twentieth-century climate change that they used to test this proposition, however, include a variety of different human and natural forcing factors, such as changes in sulfate and soot aerosols, volcanic dust, the Sun's energy output, and land surface properties. These so-called "20CEN" experiments are not just driven by human-caused increases in GHGs. Douglass et al.'s proposition that they are only testing the response of climate models to GHG increases is simply incorrect.
8. For example, 1998 was unusually warm because of the effects of a very large El Niño.
9. Irrespective of which one of four different observational datasets was used to characterize changes in tropical surface temperatures.
10. Developed by John Christy (one of the coauthors of the Douglass et al. paper), Roy Spencer, and colleagues.
11. See Table V in our paper.
12. A third reason is that several studies published within the last 12 months provide independent evidence for substantial warming of the tropical

troposphere. These studies have documented pronounced increases in surface specific humidity and atmospheric water vapor that are in accord with tropospheric warming.

13. Several of the newer radiosonde and satellite datasets that exhibit pronounced tropospheric warming are based on novel approaches to the construction of homogeneous datasets. These approaches often involve bringing in data from new sources (such as hitherto unused satellite data, or data on the physical relationship between temperature and wind) in order to better constrain uncertainties in estimated tropospheric temperature changes.

5

Consistency of Modeled and Observed Temperature Trends in the Tropical Troposphere

B.D. Santer, P.W. Thorne, L. Haimberger, K.E. Taylor, T.M.L. Wigley, J.R. Lanzante, S. Solomon, M. Free, P.J. Gleckler, P.D. Jones, T.R. Karl, S.A. Klein, C. Mears, D. Nychka, G.A. Schmidt, S.C. Sherwood, and F.J. Wentz

5.1 Introduction

There is now compelling scientific evidence that human activities have influenced global climate over the past century (e.g., IPCC 1996, 2001, 2007; Karl et al. 2006). A key line of evidence involves "fingerprint" studies, which attempt to identify the causes of historical climate change

B.D. Santer (✉) • K.E. Taylor • P.J. Gleckler • S.A. Klein
Program for Climate Model Diagnosis and Intercomparison (PCMDI), Lawrence Livermore National Laboratory, Livermore, CA, USA

P.W. Thorne
U.K. Meteorological Office Hadley Centre, Exeter, UK

L. Haimberger
Department of Meteorology and Geophysics, University of Vienna, Vienna, Austria

E.A. Lloyd, E. Winsberg (eds.), *Climate Modelling*,
https://doi.org/10.1007/978-3-319-65058-6_5

through rigorous statistical comparison of models and observations (e.g., Santer et al. 1996; Mitchell et al. 2001; Hegerl et al. 2007). Fingerprint research consistently finds that natural causes alone cannot explain the recent changes in many different aspects of the climate system—the simplest, most internally consistent explanation of the observations invariably involves a pronounced human effect.

One recurring criticism of such findings is that the climate models employed in fingerprint studies are in fundamental disagreement with

---

T.M.L. Wigley • D. Nychka
National Center for Atmospheric Research, Boulder, CO, USA

J.R. Lanzante
National Oceanic and Atmospheric Administration/Geophysical Fluid Dynamics Laboratory, Princeton, NJ, USA

S. Solomon
National Oceanic and Atmospheric Administration Earth System Research Laboratory, Chemical Sciences Division, Boulder, CO, USA

M. Free
National Oceanic and Atmospheric Administration/Air Resources Laboratory, Silver Spring, MD, USA

P.D. Jones
Climatic Research Unit, School of Environmental Sciences, University of East Anglia, Norwich, UK

T.R. Karl
National Oceanic and Atmospheric Administration/National Climatic Data Center, Asheville, NC, USA

C. Mears • F.J. Wentz
Remote Sensing Systems, Santa Rosa, CA, USA

G.A. Schmidt
NASA/Goddard Institute for Space Studies, New York, NY, USA

S.C. Sherwood
Yale University, New Haven, CT, USA

observations of tropospheric temperature change (Douglass et al. 2004, 2007). In climate model simulations, increases in well-mixed greenhouse gases cause the tropical troposphere to warm relative to the surface (Manabe and Stouffer 1980). In contrast, some satellite and radiosonde datasets show little or no warming of the tropical troposphere since 1979, and imply that temperature changes aloft are smaller than at the surface.

The "differential warming" of the surface and troposphere has been the subject of intense scrutiny (NRC 2000; Santer et al. 2005; Karl et al. 2006; Trenberth et al. 2007). It has raised questions about both model performance and the reliability of observed estimates of surface warming (Singer 2001). In addressing the latter concern, the first report of the U.S. Climate Change Science Program (CCSP) noted that progress had been made in identifying and correcting for errors in satellite and radiosonde data. At the global scale, newer upper air datasets showed "no significant discrepancy" between surface and tropospheric warming, consistent with model results (Karl et al. 2006, page iii). The Fourth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC) reached similar findings, concluding that "*New analyses of balloon-borne and satellite measurements of lower- and mid-tropospheric temperature show warming rates that are similar to those of the surface temperature record*" (IPCC 2007, page 5).

The CCSP report used several of these newer observational datasets in extensive comparisons of simulated and observed temperature changes. For global-mean changes, model estimates of differential warming were consistent with observations. In the tropics, however, it was noted that "*most observational datasets show more warming at the surface than in the troposphere, while most model runs have larger warming aloft than at the surface*" (Karl et al. 2006, page 90). Although the CCSP report did not make a definitive determination of the cause or causes of these tropical discrepancies, it found that "structural uncertainties" in observations were large enough to encompass the model estimates of temperature change. Residual errors in the satellite and radiosonde data were therefore judged to be the most likely explanation for the remaining discrepancies (Karl et al. 2006, page 3).

Structural uncertainties arise because different groups make different processing choices in the complex procedure of adjusting raw measurements for inhomogeneities (Thorne et al. 2005a). In radiosonde temperature records, inhomogeneous behavior can be caused by changes in site location, measurement time, instrumentation, and the effectiveness of thermal shielding of the temperature sensor (Lanzante et al. 2003; Seidel et al. 2004; Sherwood et al. 2005; Randel and Wu 2006; Mears et al. 2006). Nonphysical temperature changes in satellite records can occur through orbital drift or decay, inter-satellite instrumental biases, and drifts in instrumental calibration (Wentz and Schabel 1998; Christy et al. 2000, 2003; Mears et al. 2003, 2006; Mears and Wentz 2005; Trenberth et al. 2007). Because of these large uncertainties, neither satellite- nor radiosonde-based atmospheric temperature measurements constitute an unimpeachable gold standard for evaluating model performance (Thorne et al. 2007).

A recent study by Douglass, Christy, Pearson, and Singer (Douglass et al. 2007; hereinafter DCPS07) revisits earlier comparisons of simulated and observed tropospheric temperature changes performed by Santer et al. (2005, 2006), and concludes that "*models and observations disagree to a statistically significant extent*." This contradicts the findings of both Santer et al. (2005) and the previously mentioned CCSP and IPCC reports (Karl et al. 2006; IPCC 2007). As DCPS07 note, their conclusions were reached "*based on essentially the same data*" used in earlier work.

DCPS07 interpret their results as evidence that models are seriously flawed and that model-based projections of future climate change are unreliable. Singer (2008) makes an additional and even stronger assertion: that the information presented in DCPS07 "*clearly falsifies the hypothesis of anthropogenic greenhouse warming*."

If such claims were correct, they would have significant scientific implications. It is therefore of interest to examine (as we do here) the "robust statistical test" that DCPS07 rely on in order to reach the conclusion that models are inconsistent with observations. We also evaluate other formal statistical tests of the significance of modeled and observed temperature trend differences. We use a variety of different observational datasets, which enables us to explore the sensitivity of our results to

current "structural uncertainties" in observed estimates of surface and tropospheric temperature change.

The structure of our paper is as follows. In Sect. 5.2, we introduce the observational and model tropospheric temperature datasets analyzed here. Section 5.3 covers basic statistical issues that arise in comparisons of modeled and observed trends. Section 5.4 describes various tests (among them the DCPS07 test) of the formal statistical significance of trend differences. Results obtained after applying these tests to model and observational data are discussed in Sect. 5.5. Test behavior with synthetic data is considered in Sect. 5.6. This is followed in Sect. 5.7 by a comparison of vertical profiles of temperature change in climate models and radiosonde data. A summary and conclusions are given in Sect. 5.8. An Appendix summarizes the statistical notation used in the paper.

5.2 Observational and Model Temperature Data

Observational Data

Satellite Data

Since late 1978, atmospheric temperatures have been monitored routinely from space by the Microwave Sounding Units (MSU) and Advanced Microwave Sounding Units (AMSU) flown on NOAA polar-orbiting satellites. Both instruments measure the microwave emissions of oxygen molecules, which are roughly proportional to atmospheric temperature (Spencer and Christy 1990). By measuring emissions at different frequencies, it is possible to retrieve the temperatures of different atmospheric layers. Most scientific attention has focused on MSU-derived temperatures for the lower stratosphere (T_4), the mid-troposphere to lower stratosphere (T_2), and the lower to mid-troposphere (T_{2LT}). The bulk (90%) of the emissions contributing to these temperatures occur between roughly 14 and 29 km for T_4, the surface to 18 km for T_2, and the surface to 8 km for T_{2LT} (Karl et al. 2006).

To date, four different groups have been actively involved in the development of multi-decadal temperature records from MSU data. These groups are based at the *University of Alabama at Huntsville* (UAH; Spencer and Christy 1990; Christy et al. 2007), *Remote Sensing Systems* in Santa Rosa, California (RSS; Mears et al. 2003; Mears and Wentz 2005), the *University of Maryland* (UMd; Vinnikov and Grody 2003; Vinnikov et al. 2006), and the NOAA *National Environmental Satellite, Data, and Information Service* (NOAA/NESDIS; Zho et al. 2006). All four groups have made different choices in the complex process of adjusting raw MSU and AMSU data for inhomogeneities. This leads to structural uncertainties in tropical tropospheric temperature trends that are at least as large as 0.14 °C/decade for T_2 and 0.10 °C/decade for T_{2LT} (Lanzante et al. 2006).[1]

Our interest here is primarily in the T_2 and T_{2LT} data produced by UAH and RSS.[2] Data from both groups are employed in the DCPS07 consistency test between modeled and observed trends. We use results from version 3.0 of the RSS data and versions 5.1 and 5.2 (respectively) of the UAH T_2 and T_{2LT} data.[3] Data were available in the form of gridded, monthly-mean products for the period January 1979 through December 2007.

Radiosonde Data

DCPS07 compared model-simulated profiles of atmospheric temperature change with vertical profiles estimated from radiosondes. We perform a similar comparison in Sect. 5.7. Like DCSP07, we rely on radiosonde datasets produced by the U.K. Meteorological Office Hadley Centre (HadAT2; Thorne et al. 2005b; McCarthy et al. 2008), NOAA (RATPAC-A; Free et al. 2005), and the University of Vienna (RAOBCORE version 1.2; Haimberger 2007).[4] For the latter dataset, information from the ERA-40 reanalysis (Uppala et al. 2005) was used to identify and adjust for inhomogeneities in the radiosonde data assimilated by the reanalysis model. HadAT2 and RATPAC-A do not utilize reanalysis information in adjusting for inhomogeneities.

We also analyze four newly developed radiosonde datasets that were not considered by DCPS07. The first two (RAOBCORE v1.3 and v1.4; Haimberger et al. 2008) are more recent versions of the RAOBCORE dataset

used by DCPS07. The third (RICH; *Radiosonde Innovation Composite Homogenization*) uses a new automatic data homogenization method involving information from both reanalysis and composites of neighboring radiosonde stations (Haimberger et al. 2008). The fourth (IUK; *Iterative Universal Kriging*) employs an iterative approach to fit the raw radiosonde data to a statistical model of natural climate variability plus step changes associated with instrumental biases (Sherwood 2007; Sherwood et al. 2008). As will be shown later, all four newer radiosonde datasets exhibit larger warming of the tropical lower troposphere than the datasets selected by DCPS07.

Surface Data

Comparisons of surface and tropospheric warming trends provide a simple measure of changes in temperature lapse rates (Gaffen et al. 2000). Here, we use four different surface temperature datasets to estimate changes in lower tropospheric lapse rates in the deep tropics. The first three datasets contain information on sea surface temperatures only (T_{SST}), while the fourth dataset is a blend of 2 m temperatures over *Land plus Ocean* SSTs (T_{L+O}). The three SST datasets are more appropriate to analyze in order to determine whether observed lower tropospheric temperature changes follow a moist adiabatic lapse rate (Wentz and Schabel 2000).

The three SST datasets are spatially complete, and rely on statistical procedures to "infill" SST information in data-sparse regions. The first dataset, HadISST1, was developed at the U.K. Meteorological Office Hadley Centre (Rayner et al. 2003). SSTs were reconstructed from in situ observations using an optimal interpolation procedure, with subsequent "*superposition of quality-improved gridded observations onto the reconstructions to restore local detail*" (see http://www.hadobs.org/). The other two SST products are versions 2 and 3 of the NOAA ERSST ("Extended Reconstructed SST") dataset developed at the National Climatic Data Center (NCDC; Smith and Reynolds 2005; Smith et al. 2008). Differences between ERSST-v2 and ERSST-v3 are primarily related to differences in treatment of low-frequency variability and to the inclusion of bias-adjusted satellite infrared data in ERSST-v3. The newer dataset is

regarded as "*an improved extended reconstruction over version 2*" (see http://www.ncdc.noaa.gov/oa/climate/research/sst/ersstv3.php).

The fourth dataset, HadCRUT3v, consists of a blend of land 2 m temperatures from the Climatic Research Unit's CRUTEM3 dataset (Brohan et al. 2006) and SSTs from the Hadley Centre HadSST2 product (Rayner et al. 2006). Unlike the SST datasets described above, HadCRUT3v is not spatially complete. Calculation of lapse rate changes with HadCRUT3v facilitates comparison with previous work by Santer et al. (2005, 2006) and DCPS07, which also relied on surface datasets comprised of combined SSTs and land 2 m temperatures.

Model Data

A number of different climate model experiments were performed in support of the IPCC Fourth Assessment Report (IPCC 2007). In the experiment of most interest here, nearly two dozen different climate models were forced with estimates of historical changes in both anthropogenic and natural external factors.[5]

These so-called "twentieth century" ("20CEN") simulations are the most appropriate runs for direct comparison with satellite and radiosonde data, and provide valuable information on current structural and statistical uncertainties in model-based estimates of historical climate change. Inter-model differences in 20CEN results reflect differences in model physics, dynamics, parameterizations of sub-grid scale processes, horizontal and vertical resolution, and the applied forcings (Santer et al. 2005, 2006).

Santer et al. (2005) examined a set of 49 simulations of twentieth-century climate performed with 19 different models. The same suite of runs is analyzed here.[6] Santer et al. (2005) were primarily concerned with comparisons of modeled and observed amplification of surface warming in the tropical troposphere,[7] while the focus of the present work is on testing the significance of trend differences.

To facilitate the comparison of simulated and observed tropospheric temperature trends, we calculate synthetic MSU T_2 and T_{2LT} temperatures from gridded, monthly-mean model data using a static

global-mean weighting function. For temperature changes averaged over large areas, this procedure yields results similar to those estimated with a full radiative transfer code (Santer et al. 1999). Since most of the 20CEN experiments end in 1999, our trend comparisons primarily cover the 252-month period from January 1979 to December 1999—the period of maximum overlap between the observed MSU data and the model simulations.

5.3 Basic Statistical Issues

We assume a simulated tropospheric temperature time series $y_m(t)$ of the form:

$$y_m(t) = \phi_m(t) + \eta_m(t) \tag{5.1}$$

where $\phi_m(t)$ is the underlying signal in response to external forcing, $\eta_m(t)$ is a specific realization of natural internal climate variability superimposed on $\phi_m(t)$, t is a nominal index of time in months, and the subscript m denotes *model* data. The corresponding *observed* time series $y_o(t)$ is given by

$$y_o(t) = \phi_o(t) + \eta_o(t) \tag{5.2}$$

The slopes of the least-squares linear trends in these time series (b_m and b_o) provide one measure of overall change in temperature. Estimates of b_m and b_o are sensitive to the behavior of both signal and noise components in the time series.

In the tropics, the El Niño/Southern Oscillation (ENSO) phenomenon explains most of the year-to-year variability in observed tropospheric temperatures. The real world provides only one sample of how ENSO and other modes of internal climate variability influence atmospheric temperature. This makes it difficult to achieve an unambiguous separation of signal from noise in observational data. Models, however, can be run many times to generate many different realizations of historical climate change,[8] thus

facilitating the separation of $\phi_m(t)$ from $\eta_m(t)$. Since $\eta_m(t)$ is uncorrelated from one realization to the next, averaging over many realizations reduces noise levels and improves estimates of any overall trend in $\phi_m(t)$.

This is clearly illustrated in Figs. 5.1a–e, which show tropical T_{2LT} changes over 1979–1999 in five 20CEN realizations performed with the Japanese Meteorological Research Institute (MRI) model. The character of $\eta_m(t)$ is different in each realization, resulting in a large range of trends in $y_m(t)$ (from 0.042 °C to 0.371 °C/decade).

The small overall trend in realization 1 is partly due to the chance occurrence of El Niños near the beginning and middle of the time series, and the presence of a La Niña at the end. Averaging over these five realizations reduces the amplitude of $\eta_m(t)$, and improves the estimate of the true forced change in $y_m(t)$ (Fig. 5.1f). The key point to note is that the same MRI model, with exactly the same physics and forcings, produces a range of self-consistent estimates of tropical T_{2LT} trends over a particular time interval, not a single discrete value. Many other models with ensembles of 20CEN runs also show substantial inter-realization trend differences (see Sect. 5.5.1.1).

A number of factors may contribute to differences between modeled and observed temperature trends. These include:

1. Missing or inaccurately specified values of the external forcings applied in the model 20CEN run.

Fig. 5.1 Anomaly time series of monthly-mean T_{2LT}, the spatial average of lower tropospheric temperature over tropical (20°N–20°S) land and ocean areas. Results are for five different realizations of twentieth-century climate change performed with a coupled A/OGCM (the MRI-CGCM2.3.2). Each of the five realizations (*panels A–E*) was generated with the same model and the same external forcings, but with initialization from a different state of the coupled atmosphere-ocean system. This yields five different realizations of internally generated variability, $\eta_m(t)$, which are superimposed on the true response to the applied external forcings. The ensemble-mean T_{2LT} change is shown in panel F. Least-squares linear trends were fitted to all time series; values of the trend and lag-1 autocorrelation of the regression residuals (r_1) are given in each panel. Anomalies are defined relative to climatological monthly means over January 1979 to December 1999, and synthetic T_{2LT} temperatures were calculated as described in Santer et al. (1999)

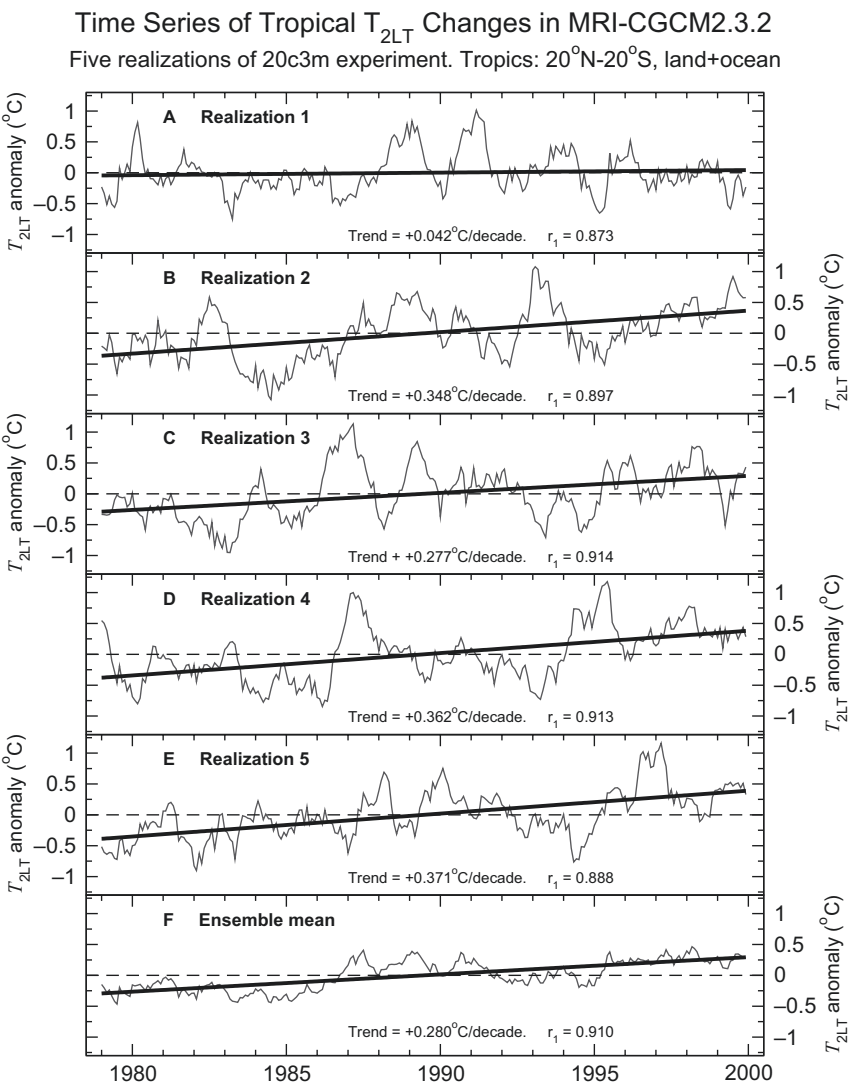

Fig. 5.1 (continued)

2. Errors in $\phi_m(t)$, the model's response to the imposed forcing changes.
3. Errors in the variability and other statistical properties of $\eta_m(t)$.
4. The irreproducibility of the specific, essentially random sequence of observed noise, even by a model which correctly simulates the statistical properties of $\eta_o(t)$.
5. The number of 20CEN realizations for any given model, which influences how well we can estimate $\phi_m(t)$. In the limit of many realizations of $y_m(t)$, the model's ensemble-mean trend would provide an accurate estimate of the forced component of change in $y_m(t)$.
6. Residual inhomogeneities in $y_o(t)$.

Even in a model with no errors in forcing, response, or internally generated variability, we could by chance have realizations of noise that differed markedly from that in the real world, leading to a large difference between modeled and observed trends that was completely unrelated to model error. Any procedure for testing the significance of differences between simulated and observed trends must therefore account for the (potentially different) effects of internally generated variability on b_m and b_o.

5.4 Significance Tests

Our significance testing strategy addresses two different questions. The first is whether models can simulate individual temperature trends that are consistent with the single observed trend. The second question is whether our current best estimate of the model response to external forcing is consistent with our estimate of the externally forced temperature trend in observations.

Each question involves testing a different hypothesis. In the first question, we are testing hypothesis H_1 that the trend in any given realization of $y_m(t)$ is consistent with the trend in $y_o(t)$. As noted previously, interannual climate noise makes it difficult to obtain reliable estimates of the forced components of temperature change [$\phi_o(t)$ and $\phi_m(t)$] from the single $y_o(t)$ time series and from any individual realization of $y_m(t)$. Under hypothesis H_1, therefore, we are comparing trends arising from a combination of forced and unforced temperature changes.

The hypothesis H_2 tested in the second question involves the multimodel ensemble-mean trend. Averaging over realizations and models reduces noise and provides a better estimate of the true model signal in response to external forcing. Under H_2, we seek to determine whether the model average signal is consistent with the trend in $\phi_o(t)$ (the signal contained in the observations).

Tests with Individual Model Realizations

To examine H_1, we apply a "paired trends" test (Santer et al. 2000b; Lanzante 2005), in which b_o is tested against each of the 49 individual b_m trends considered here. The test statistic is of the form

$$d = \left(b_m - b_o\right)\sqrt{s\left\{b_m\right\}^2 + s\left\{b_o\right\}^2} \tag{5.3}$$

where d is the normalized difference between the trends in any two modeled and observed time series, and $s\{b_m\}$and $s\{b_o\}$are (respectively) the standard errors of b_m and b_o. The standard errors are measures of the inherent statistical uncertainty in fitting a linear trend to noisy data. For the model data, $s\{b_m\}$is defined as

$$s\left\{b_m\right\} = \left[s_e^2 \Big/ \sum_{n_t}^{t=1}\left(t - \bar{t}\right)^2\right]^{\frac{1}{2}} \tag{5.4}$$

where t is the time index, $\bar{t}$ is the average time index, n_t is the total number of time samples (252 here), and s_e^2 is the variance of the regression residuals, given by

$$s_e^2 = \frac{1}{n_t - 2}\sum_{n_t}^{t=1} e\left(t\right)^2 \tag{5.5}$$

(see Wilks 1995). Note that the observed standard error, $s\{b_o\}$, is calculated similarly, but using observational rather than model data.

Assuming that d has a Normal distribution, we can compute its associated p-value and test whether the trend in $y_m(t)$ is consistent with the trend in $y_o(t)$. This test is two-tailed, since we have no expectation a priori regarding the direction of the trend difference.

In the case of most atmospheric temperature series, the regression residuals $e(t)$ are not statistically independent. For RSS tropical T_{2LT} data, for example (Fig. 5.2a), values of $e(t)$ have pronounced month-to-month and year-to-year persistence, with a lag-1 temporal autocorrelation coefficient of $r_1 = 0.884$ (Table 5.1). This persistence reduces the number of statistically independent time samples. Following Santer et al. (2000a), we account for the nonindependence of $e(t)$ values by calculating an effective sample size n_e:

$$n_e = n_t \frac{1 - r_1}{1 + r_1} \tag{5.6}$$

By substituting $n_e - 2$ for $n_t - 2$ in Eq. (5.5), the standard error is adjusted for the effects of temporal autocorrelation (see Supporting Online Material). In the RSS example in Fig. 5.2a, $n_e \approx 16$, and the adjusted standard error is over four times larger than the unadjusted

Fig. 5.2 Calculation of unadjusted and adjusted standard errors for least-squares linear trends. The standard error $s\{b_o\}$ of the least-squares linear trend b_o (see Sect. 5.4.1) is a measure of the uncertainty inherent in fitting a linear trend to noisy data. Two examples are given here. Panel A shows observed tropical T_{2LT} anomalies from the RSS group (Mears and Wentz 2005). The regression residuals (shaded blue) are highly autocorrelated ($r_1 = 0.884$). Accounting for this temporal autocorrelation reduces the number of effectively independent time samples from 252 to 16, and inflates $s\{b_o\}$ by a factor of four (see "Results from A" in panel C). The anomalies in panel B were generated by adding Gaussian noise to the RSS tropical T_{2LT} trend, yielding a trend and temporal standard deviation that are very similar to those of the actual RSS data. For this synthetic data series, the regression residuals (shaded red) are uncorrelated and r_1 is close to zero, so that the actual number of time samples is similar to the effective sample size, and the unadjusted and adjusted standard errors are small and virtually identical (see "Results from B" in panel C). All results in panel C are 2σ confidence intervals (C.I.). The analysis period is from January 1979 to December 1999

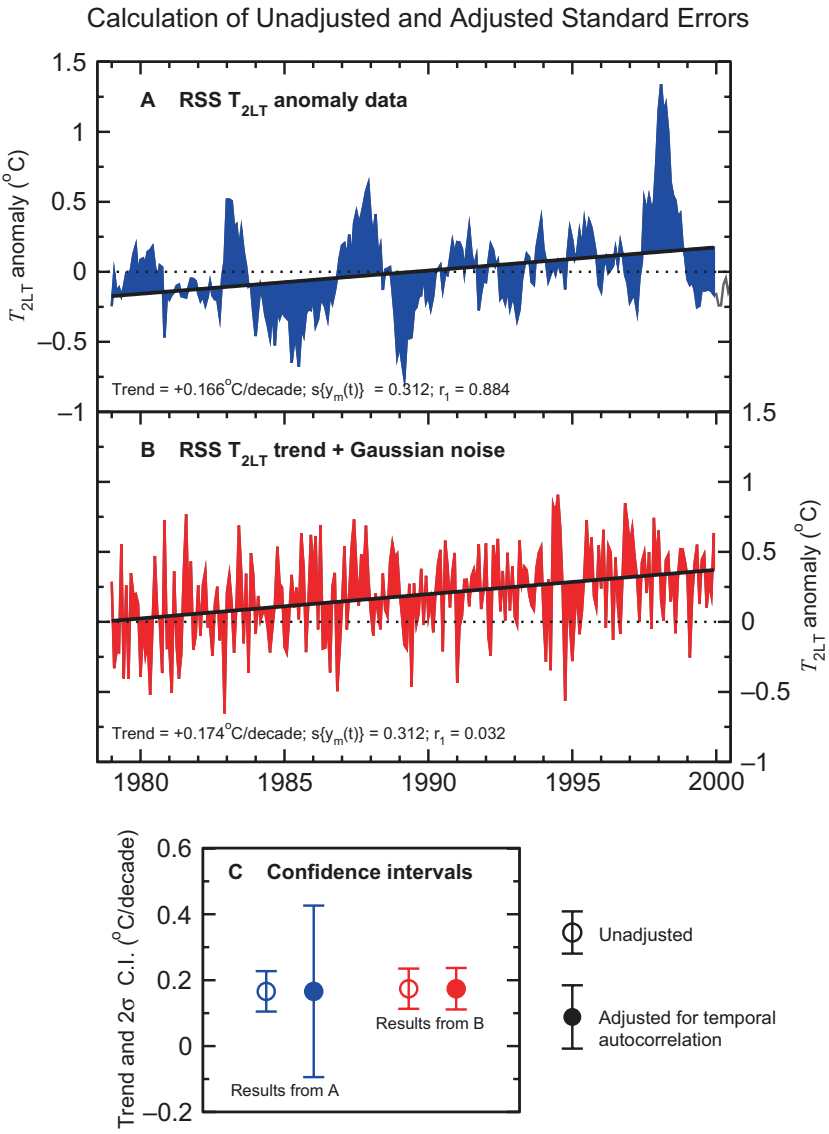

Fig. 5.2 (continued)

standard error (Fig. 5.2c). The unadjusted standard error should only be used if the regression residuals are uncorrelated. In the case of the synthetic data in Fig. 5.2b, for example, r_1 is close to zero, n_e and n_t are of similar size (236 and 252), and the adjusted and unadjusted standard errors are small and virtually identical (Fig. 5.2c). Our subsequent discussion of the paired trends test (Sect. 5.5) deals exclusively with results computed correctly with adjusted standard errors rather than with unadjusted standard errors.

Table 5.1 Statistics for observed and simulated time series of land and ocean surface temperatures, SST, and tropospheric temperatures

Dataset	Trend	1σ S.E.	Std. Dev.	r_1	n_e
HadCRUT3v T_{L+O}	0.119	0.117	0.197	0.934	8.6
Multi-model mean T_{L+O}	0.146	0.214	0.274	0.915	11.7
Inter-model S.D. T_{L+O}	0.066	0.163	0.093	0.087	13.9
HadISST1 T_{SST}	0.108	0.133	0.197	0.944	7.3
ERSST-v2 T_{SST}	0.100	0.131	0.186	0.947	6.9
ERSST-v3 T_{SST}	0.077	0.121	0.190	0.936	8.3
Multi-model mean T_{SST}	0.130	0.333	0.243	0.959	5.3
Inter-model S.D. T_{SST}	0.062	0.336	0.084	0.024	3.2
UAH T_{2LT}	0.060	0.138	0.299	0.891	14.5
RSS T_{2LT}	0.166	0.132	0.312	0.884	15.6
Multi-model mean T_{2LT}	0.215	0.198	0.376	0.876	17.2
Inter-model S.D. T_{2LT}	0.092	0.133	0.127	0.080	12.2
UAH T_2	0.043	0.129	0.306	0.873	17.1
RSS T_2	0.142	0.129	0.319	0.871	17.3
Multi-model mean T_2	0.199	0.181	0.370	0.855	20.3
Inter-model S.D. T_2	0.098	0.133	0.132	0.085	13.0

Basic statistical properties of observed and simulated time series of tropical temperatures. Results are for time series of monthly-mean anomalies in land and ocean surface temperature (T_{L+O}), sea surface temperature (T_{SST}), and tropospheric temperature (T_{2LT}, T_2). Analyses are over the 252-month period from January 1979 through December 1999 (the period of maximum overlap between the observations and most model 20CEN experiments). Gridded anomaly data were spatially averaged over 20°N–20°S. The time series statistics are the least-squares linear trend (b_o, b_m; °C/decade); the standard error of the linear trend, adjusted for temporal autocorrelation effects ($s\{b_o\}$, $s\{b_m\}$; °C/decade); the temporal standard deviation of the anomaly data ($s\{y_o(t)\}$, $s\{y_m(t)\}$; °C); the lag-1 autocorrelation of the regression residuals (r_1); and the effective number of independent time samples (n_e). The multi-model mean and inter-model standard deviation were calculated using the ensemble-mean values of the time series statistics for the 19 models [see Eqs. (5.7, 5.8, and 5.9)]. Anomalies were defined relative to climatological monthly means computed over the analysis period. For sources of model and observed data, see Sect. 5.2

The underlying assumption in our method of adjusting standard errors is that the temporal persistence of $e(t)$ can be well represented by a lag-1 autoregressive (AR) statistical model. This assumption is not uncommon in meteorological applications (e.g., Wilks 1995; Lanzante et al. 2006). If the autocorrelation structure is more complex and exhibits long-range dependence, it may be more appropriate to use higher-order AR models for estimating n_e (Thiébaux and Zwiers 1984). However, it is difficult to reliably estimate the parameters of such statistical models given the relatively short length (20–30 years) and high temporal autocorrelation of the temperature data available here.

Experiments with synthetic data reveal that the use of an AR-1 model for calculating n_e tends to overestimate the true effective sample size (Zwiers and von Storch 1995). This means that our d test is too liberal, and is more likely to indicate that there are significant differences between modeled and observed trends, even when significant differences do not actually exist.[9] It should therefore be easier for us to confirm DCPS07's finding that modeled and observed trends are inconsistent. As described in Sect. 5.5, however, our results do *not* confirm DCPS07's findings. DCPS07's conclusions are erroneous, and are primarily due to the neglect of observed trend uncertainties in their statistical test (see Sect. 5.4.2).

Tests with Multi-Model Ensemble-Mean Trend

Here we examine two different tests of the hypothesis H_2 (see Sect. 5.4). Both rely on the multi-model ensemble-mean trend,[10] $<< b_m >>$:

$$<< b_m >> = \frac{1}{n_m} \sum_{n_m}^{i=1} \langle b_m(i) \rangle \tag{5.7}$$

where $< b_m(i) >$ is the ensemble-mean trend in the i^{th} model:

$$\langle b_m(i) \rangle = \frac{1}{n_r(i)} \sum_{n_r(i)}^{j=1} b_m(i,j); \; i = 1, \ldots, n_m \tag{5.8}$$

The indices i and j are over model number and realization number (respectively). The total number of models is n_m (19 here), and n_r (i) is the total number of 20CEN realizations for the i^{th} model (which varies from 1 to 5). The standard deviation of ensemble-mean trends, $s\{<b_m>\}$, is given by

$$s\left\{\langle b_m\rangle\right\} = \left[\frac{1}{n_m - 1}\sum_{n_m}^{i=1}\left(\langle b_m(i)\rangle - << b_m >>\right)^2\right]^{1/2} \tag{5.9}$$

In the DCPS07 "consistency test", the difference between $<< b_m>>$ and b_o is compared with σ_{SE}, "*an estimate of the uncertainty of the* (multi-model) *mean* (trend)." DCPS07 do not consider *any* uncertainty in b_o, and σ_{SE} is based solely on the inter-model variability of trends:

$$\sigma_{SE} = s\left\{\langle b_m\rangle\right\} / \sqrt{n_m} \tag{5.10}$$

To evaluate the performance of the DCPS07 test, we define the test statistic d^*:

$$d^* = \left(<< b_m >> - b_o\right) / \sigma_{SE} \tag{5.11}$$

If the DCPS07 test were valid, a large value of d^* would imply a significant difference between $<< b_m>>$ and b_o. However, the test is not valid. There are a number of reasons for this:

1. DCPS07 ignore the pronounced influence of interannual variability on the observed trend (see Fig. 5.2a). They make the implicit (and incorrect) assumption that the externally forced component in the observations is perfectly known (i.e., that the observed record consists only of $\phi_o(t)$ and that $\eta_o(t) = 0$).
2. DCPS07 ignore the effects of interannual variability on model trends—an effect which we consider in our "paired trends" test [see Eq. (5.3)]. They incorrectly assume that the forced component of temperature

change is perfectly known in each individual model (i.e., that each individual 20CEN realization consists only of $\phi_m(t)$ and that $\eta_m(t) = 0$).[11]

3. DCPS07's use of σ_{SE} is incorrect. While σ_{SE} is an appropriate measure of how well the multi-model mean trend can be estimated from a finite sample of model results, it is not an appropriate measure for deciding whether this trend is consistent with a single observed trend.

Practical consequences of these problems are discussed later in Sects. 5.5 and 5.6.

We can easily modify the DCPS07 d^* test to account for the factor neglected by DCPS07—the effects of interannual variability on the "trend signal" in $y_o(t)$. The resulting d_1^* test is similar in form to a t-test of the difference in means:

$$d_1^* = \left(<< b_m >> - b_o \right) / \sqrt{\frac{1}{n_m} s\left\{ \langle b_m \rangle \right\}^2 + s\left\{ b_0 \right\}^2} \tag{5.12}$$

where the term $\frac{1}{n_m} s\left\{ \langle b_m \rangle \right\}^2$ is a standard estimate of the variance of the mean (in this case, the variance of the model average trend $<< b_m >>$; see Storch and Zwiers 1999), and $s\{b_o\}^2$ is an estimate of the variance of the observed trend b_o [see Eqs. (5.4, 5.5, and 5.6)].

There are three underlying assumptions in the d_1^* test. The first assumption (which is also made by DCPS07) is that the uncertainty in $<< b_m >>$ is entirely due to inter-model differences in forcing and response, and not to differences in variability and ensemble size. The second assumption is that the uncertainties in the observed trend are due solely to the effects of interannual variability—i.e., that there are no residual errors in the observations being tested. The third assumption is that d_1^* has a Student's t distribution, and that the number of degrees of freedom associated with the estimated variances of $<< b_m >>$ and b_o are $n_m - 1$ and $n_e - 2$, respectively.

As noted above, the variances of $<< b_m >>$ and b_o are influenced by very different factors, and are unlikely to be identical. In this case, the degrees of freedom for the test DOF$\{d_1^*\}$ are approximated by

$$\mathrm{DOF}\left\{d_1^*\right\} = \left[1/n_m s\left\{\langle b_m\rangle\right\}^2 + s\left\{b_0\right\}^2\right]^2 / \frac{\left[1/n_m s\left\{< b_m >\right\}^2\right]^2}{n_m - 1} + \frac{\left[s\left\{b_0\right\}^2\right]^2}{n_e - 2} \tag{5.13}$$

(see Storch and Zwiers 1999). We will demonstrate in Sect. 5.6 that d_1^* and the DCPS07 d^* test exhibit very different behavior when applied to synthetic data.

5.5 Results of Significance Tests

Tropospheric Temperature Trends

Tests with Individual Model Realizations

Figure 5.3a shows trends in tropical T_{2LT} in the two satellite datasets (RSS and UAH) and in 49 realizations of the 20CEN experiment, together with their adjusted 2σ confidence intervals. Values of b_m vary substantially, not only between models, but also within the different 20CEN realizations of individual models. The adjusted 2σ confidence interval on the RSS T_{2LT} trend includes 47 of the 49 simulated trends. This strongly suggests that there is no fundamental inconsistency between modeled and observed trends.[12]

Results from the paired trends test [see Eq. (5.3)] are summarized in Table 5.2. For each of the two layer-averaged temperatures considered here (T_{2LT} and T_2), UAH and RSS trends were tested against trends from the 49 individual model simulations. Calculated p-values for the d statistic were compared with stipulated p-values of 0.05, 0.10, and 0.20. We then determined the number of tests in which hypothesis H_1 (see Sect. 5.4) is rejected at the 5%, 10%, and 20% significance levels.

If model and observed trends were in perfect agreement, we would still expect (for a very large number of tests) p% of the tests to show signifi-

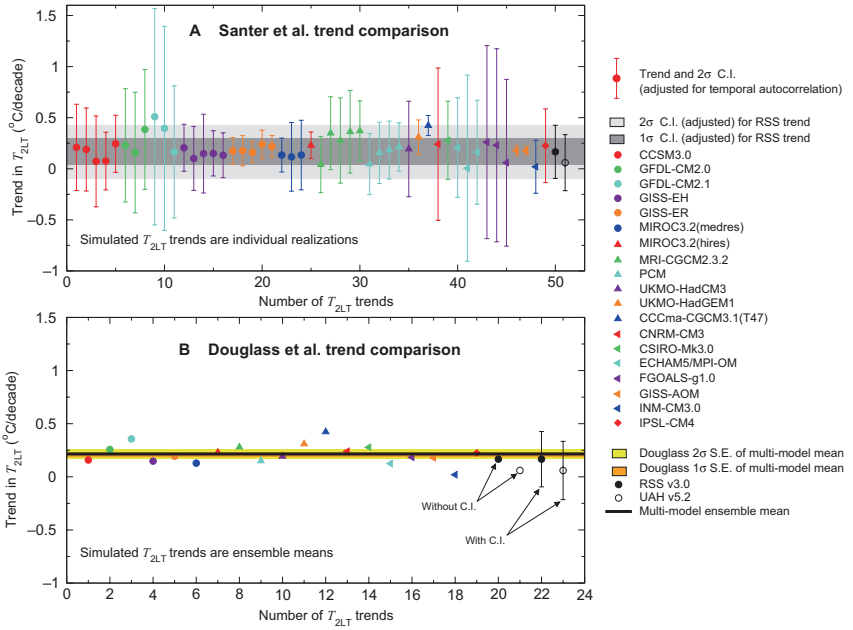

Fig. 5.3 Comparisons of simulated and observed trends in tropical T_{2LT} over January 1979 to December 1999. Model results in panel A are from 49 individual realizations of experiments with twentieth-century external forcings, performed with 19 different A/OGCMs. Observational estimates of T_{2LT} trends are from Mears and Wentz (2005) and Christy et al. (2007) for RSS and UAH data, respectively. The dark and light gray bands in panel A are the 1σ and 2σ confidence intervals for the RSS T_{2LT} trend, adjusted for temporal autocorrelation effects. In the paired trends test applied here, each individual model T_{2LT} trend is tested against each observational T_{2LT} trend (Sect. 5.4.1). Panel B shows the three elements of the DCPS07 "consistency test": the multi-model ensemble-mean T_{2LT} trend, $<< b_m >>$ (represented by the horizontal black line in panel B); σ_{SE}, DCPS07's estimate of the uncertainty in $<< b_m >>$; and b_o, the individual RSS and UAH T_{2LT} trends (with and without their 2σ confidence intervals from panel A). The 1σ and 2σ values of σ_{SE} are indicated by orange and yellow bands, respectively. The colored dots in panel B are either the ensemble-mean T_{2LT} trends for individual models or the trend in an individual 20CEN realization (for models that did not perform multiple 20CEN realizations). Statistical uncertainties in the observed trends are neglected in the DCSP07 test. If these uncertainties are accounted for, $<< b_m >>$ is well within the 2σ confidence intervals on the RSS and UAH T_{2LT} trends (Sect. 5.5.1.2)

Table 5.2 Significance of differences between modeled and observed tropospheric temperature trends: Results for paired trends tests

Sig. level (%)	RSS T_{2LT} (%)		UAH T_{2LT} (%)		RSS T_2 (%)		UAH T_2 (%)	
5	0	(0.0)	1	(2.0)	1	(2.0)	1	(2.0)
10	1	(2.0)	1	(2.0)	1	(2.0)	3	(6.1)
20	1	(2.0)	4	(8.2)	1	(2.0)	6	(12.2)

Statistical significance of differences between modeled and observed tropospheric temperature trends. Results are for the paired trends test described in Sect. 5.4.1. Model data employed in the test are tropical T_{2LT} and T_2 trends from 49 realizations of twentieth-century climate change performed with 19 different A/OGCMs (together with their associated adjusted standard errors). Observational trends and adjusted standard errors were estimated from RSS and UAH satellite data. There are 49 tests for each tropospheric layer and each observational dataset. Results are expressed as the number of rejections of hypothesis H_1 (see Sect. 5.4) at stipulated significance levels of 5%, 10%, and 20%. Percentage rejection rates of H_1 (out of 49 tests) are given in parentheses. All trends and standard errors were calculated over the period January 1979 to December 1999 from time series of spatially averaged (20 °N–20 °S) anomaly data

cant trend differences at the p% significance level. Our rejection rates are invariably lower than the theoretical expectation (Table 5.2). There are at least four possible explanations for this:

1. Not all 49 tests are statistically independent.
2. Tests are affected by differences between modeled and observed variability.
3. Results are influenced by the sampling variability arising from the relatively small number of tests performed.
4. Our method of adjusting standard errors for temporal autocorrelation effects is not reliable.[13]

Overall, however, our paired test results show broad agreement between tropospheric temperature trends estimated from models and satellite data. This consistency holds even if we account for errors in model variability (see Supporting Online Material).

Tests with Multi-Model Ensemble-Mean Trend

We now seek to understand why DCPS07 concluded that the multi-model ensemble-mean trend was inconsistent with observed trends, despite the fact that almost all of the individual b_m trends are consistent with observations (see Sect. 5.5.1.1).

Application of the DCPS07 test yields values of the test statistic d^* [see Eq. (5.11)] ranging from 2.25 for RSS T_{2LT} trends to 7.16 for UAH T_{2LT} trends (Table 5.3). In all four d^* tests,[14] hypothesis H_2 is rejected at the 5% level or better. This is why DCPS07 conclude that the multi-model ensemble-mean trend is inconsistent with observed T_{2LT} and T_2 trends. As will be shown below, this conclusion is erroneous.

It is obvious from Fig. 5.3b and Table 5.1 that for T_{2LT} data, $<< b_m >>$ lies within the adjusted 2σ confidence intervals for the RSS and UAH trends. As was noted in Sect. 5.4.2, however, DCPS07 ignore trend uncertainties arising from interannual variability, both for observational and model trends. If DCPS07 had accounted for these trend uncertainties, they would have obtained very different results.

This is evident when we apply our modified version of the DCPS07 test, which accounts for uncertainties in both the observational and model trend signals. For all four tests with d_1^*, hypothesis H_2 cannot be rejected at the nominal 5% level (Table 5.3). These findings differ radically from those obtained with DCPS07's "consistency test". We

Table 5.3 Significance of differences between modeled and observed tropospheric temperature trends: Results for tests involving multi-model ensemble-mean trend

Statistic type	RSS T_{2LT}	UAH T_{2LT}	RSS T_2	UAH T_2
d^*	2.25^{**}	7.16^{***}	2.48^{**}	6.78^{***}
d_1^*	0.37	1.11	0.44	1.19

Statistical significance of differences between modeled and observed tropospheric temperature trends. Results are the actual test statistic values for two different tests of the hypothesis H_2—the original DCPS07 "consistency test" [d^*; see Eq. (5.11)] and a modified version of the DCPS07 test [d_1^*; see Eq. (5.12)]. Both d^* and d_1^* involve the model average signal trend. The T_{2LT} and T_2 data used in the tests are described in Table 5.2. One, two, and three asterisks indicate model-versus-observed trend differences that are significant at the 10%, 5%, and 1% levels (respectively; two-tailed tests)

conclude, therefore, that when uncertainties in both observational and model trend signals are properly accounted for, there is no statistically significant difference between the model average trend signal and the observed trend in $\phi_o(t)$.

Trends in Lower Tropospheric Lapse Rates

Tests with Individual Model Realizations

Tests involving trends in the surface-minus-T_{2LT} difference series are more stringent than tests of trend differences in T_{L+O}, T_{SST}, or T_{2LT} alone. This is because differencing removes much of the common variability in surface and tropospheric temperatures, thus decreasing both the variance and lag-1 autocorrelation of the regression residuals (Wigley 2006). In turn, these twin effects increase the effective sample size and decrease the adjusted standard error of the trend, making it easier to identify significant trend differences between models and observations.

Despite these decreases in $s\{b_m\}$ and $s\{b_o\}$, however, 45 out of 49 trends in the simulated T_{SST} minus T_{2LT} difference series are still within the $\pm 2\sigma$ confidence intervals of the ERSST-v3 minus RSS difference series trend (Fig. 5.4a). Irrespective of which observational dataset is used for estimating surface temperature changes, each of the three T_{SST} minus T_{2LT} pairs involving RSS data (and the single T_{L+O} minus T_{2LT} pair) has a *negative* trend in the difference series, indicating larger warming aloft than at the surface, consistent with the model results (Table 5.4). Application of the paired trends test [Eq. (5.3)] reveals that there are very few statistically significant differences between the model difference series trends and observed lapse rate trends computed using RSS T_{2LT} data (Table 5.5).

For all four difference series "pairs" involving UAH T_{2LT} data, the warming aloft is smaller than the warming of the tropical surface, leading to a *positive* trend in the surface-minus-T_{2LT} time series—i.e., a trend of opposite sign to virtually all model results (Table 5.4 and Fig. 5.4a). Even in the UAH cases, however, not all models are inconsistent with the observed estimates of "differential warming" (despite DCPS07's claim to

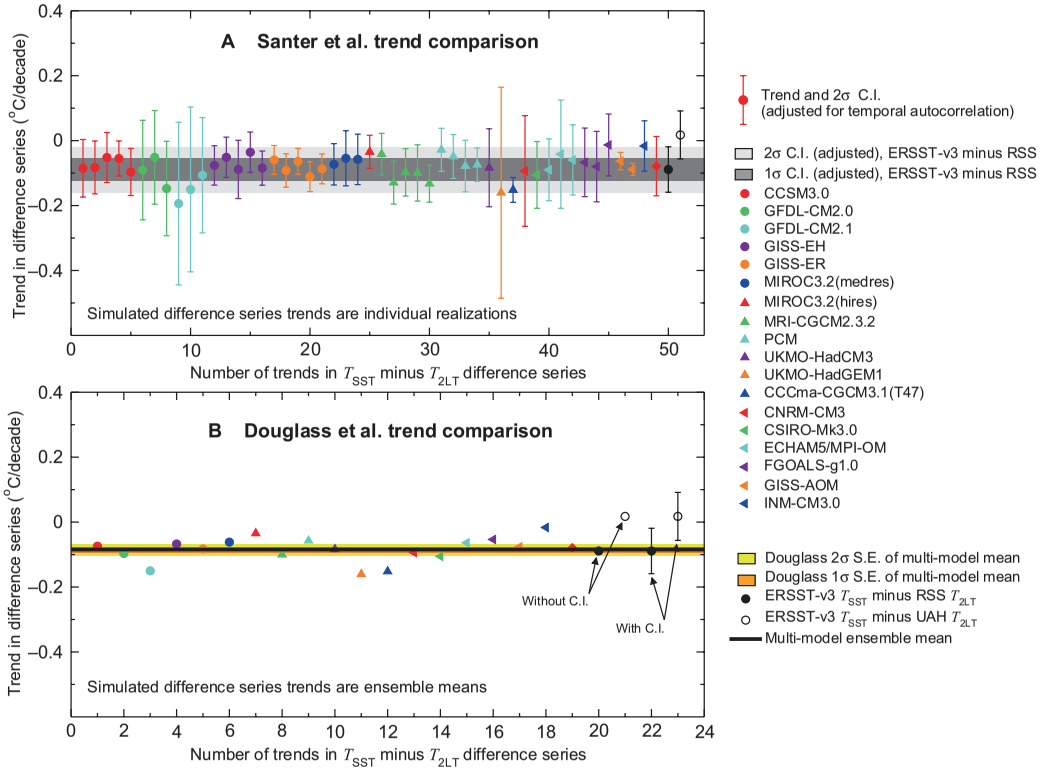

Fig. 5.4 As for Fig. 5.3, but for comparisons of simulated and observed trends in the time series of differences between tropical T_{SST} and T_{2LT}. The observed T_{SST} data are from NOAA ERSST-v3 (Smith et al. 2008). For trends and confidence intervals from other observed pairs of surface and T_{2LT} data, refer to Table 5.4

Table 5.4 Statistics for observed and simulated time series of differences between tropical surface temperature and lower tropospheric temperature

Dataset	Trend	1σ S.E.	Std. Dev.	r_1	n_e
HadCRUT3v T_{L+O} minus UAH T_{2LT}	0.061	0.036	0.165	0.642	55.0
HadCRUT3v T_{L+O} minus RSS T_{2LT}	−0.046	0.034	0.162	0.608	61.5
Multi-model mean T_{L+O} minus T_{2LT}	−0.069	0.040	0.164	0.614	62.5
Inter-model S.D. T_{L+O} minus T_{2LT}	0.032	0.031	0.057	0.137	27.3
HadISST1 T_{SST} minus UAH T_{2LT}	0.049	0.037	0.170	0.630	57.2
ERSST-v2 T_{SST} minus UAH T_{2LT}	0.041	0.040	0.172	0.665	50.7
ERSST-v3 T_{SST} minus UAH T_{2LT}	0.018	0.037	0.167	0.633	56.6
HadISST1 T_{SST} minus RSS T_{2LT}	−0.058	0.035	0.170	0.595	64.0
ERSST-v2 T_{SST} minus RSS T_{2LT}	−0.066	0.038	0.175	0.637	56.0
ERSST-v3 T_{SST} minus RSS T_{2LT}	−0.089	0.035	0.174	0.601	62.7
Multi-model mean T_{SST} minus T_{2LT}	−0.085	0.053	0.197	0.654	55.3
Inter-model S.D. T_{SST} minus T_{2LT}	0.038	0.036	0.064	0.146	28.4

As for Table 5.1, but for basic statistical properties of observed and simulated time series of differences between tropical surface and lower tropospheric temperatures. We use three datasets (HadISST1, ERSST-v2, and ERSST-v3) to characterize observed changes in T_{SST}, one dataset (HadCRUT3v) to describe changes in T_{L+O}, and two datasets (RSS and UAH) to estimate observed changes in tropical T_{2LT}. This yields eight different combinations of observed surface-minus-T_{2LT} difference series

the contrary). Rejection rates for paired trends tests with a stipulated 5% significance level range from 31% to 88%, depending on the choice of observed surface record (Table 5.5). The highest rejection rates are for lapse rate trends computed with the HadCRUT3v surface data, which has the largest surface warming.

Tests with the Multi-Model Ensemble-Mean Trend

Figure 5.4b shows that the multi-model ensemble-mean difference series trend is very close to the trend in the ERSST-v3 minus RSS difference series. In this specific case, even the incorrect, unmodified DCPS07 test yields a nonsignificant value of d^* (0.49; see Table 5.6). In seven of the other eight difference series pairs, however, use of the original DCPS07 consistency test leads to rejection of the H_2 hypothesis at the nominal 5% level (see Sect. 5.4).

Table 5.5 Significance of differences between modeled and observed trends in lower tropospheric lapse rates: Results for paired trends tests

Dataset pair	5% sig. level		10% sig. level		20% sig. level	
HadCRUT3v T_{L+O} minus UAH T_{2LT}	43	(87.8%)	45	(91.8%)	47	(95.9%)
HadISST1 T_{SST} minus UAH T_{2LT}	28	(57.1%)	39	(79.6%)	44	(89.8%)
ERSST-v2 T_{SST} minus UAH T_{2LT}	25	(51.0%)	33	(67.4%)	44	(89.8%)
ERSST-v3 T_{SST} minus UAH T_{2LT}	15	(30.6%)	24	(49.0%)	35	(71.4%)
HadCRUT3v T_{L+O} minus RSS T_{2LT}	1	(2.0%)	1	(2.0%)	3	(6.1%)
HadISST1 T_{SST} minus RSS T_{2LT}	1	(2.0%)	2	(4.1%)	3	(6.1%)
ERSST-v2 T_{SST} minus RSS T_{2LT}	1	(2.0%)	1	(2.0%)	2	(4.1%)
ERSST-v3 T_{SST} minus RSS T_{2LT}	0	(0.0%)	0	(0.0%)	2	(4.1%)

As for Table 5.2, but for paired tests involving trends in modeled and observed time series of differences between surface and lower tropospheric temperatures. Trends in T_{SST} minus T_{2LT} and T_{L+O} minus T_{2LT} provide simple measures of changes in lower tropospheric lapse rates in the tropics. For sources of data, refer to Table 5.4. Each of the eight observed difference series trends is tested against each of the 49 simulated difference series trends. Results are the number of rejections of hypothesis H_1 and the percentage rejection rates (in parentheses) for three stipulated significance levels. The analysis period and anomaly definition are as for the T_{2LT} and T_2 data described in Table 5.2

The modified DCPS07 test with d_1^* [see Eq. (5.12)] yields strikingly different results: there is no case in which the model average signal trend differs significantly from the four pairs of observed surface-minus-T_{2LT} trends calculated with RSS T_{2LT} data (Table 5.6). When the UAH T_{2LT} data are used to estimate lapse rate trends, however, H_2 is rejected at the nominal 5% level for all four of the observed surface-minus-T_{2LT} trends. This sensitivity of significance test results to the choice of RSS or UAH T_{2LT} data is qualitatively similar to that obtained for "paired trends" tests of the H_1 hypothesis (see Sect. 5.5.2.1).[15]

Summary of Tests with Lower Tropospheric Lapse Rates

On the basis of these new results, we conclude that considerable scientific progress has been made since the CCSP report, which described "*a potentially serious inconsistency*" between recent modeled and observed trends

Table 5.6 Significance of differences between modeled and observed trends in lower tropospheric lapse rates: Results for tests involving multi-model ensemble-mean trend

Statistic type		d^*	d^*
HadCRUT3v T_{L+O} minus UAH	T_{2LT}	17.05^{*}	3.50^{***}
HadISST1 T_{SST} minus UAH	T_{2LT}	14.94^{***}	3.52^{***}
ERSST-v2 T_{SST} minus UAH	T_{2LT}	14.01^{***}	3.04^{***}
ERSST-v3 T_{SST} minus UAH	T_{2LT}	11.43^{***}	2.68^{***}
HadCRUT3v T_{L+O} minus RSS	T_{2LT}	3.05^{***}	0.67
HadISST1 T_{SST} minus RSS	T_{2LT}	3.01^{***}	0.75
ERSST-v2 T_{SST} minus RSS	T_{2LT}	2.09^{**}	0.48
ERSST-v3 T_{SST} minus RSS	T_{2LT}	0.49	0.12

As for Table 5.3, but for tests of hypothesis H_2 involving trends in modeled and observed time series of differences between surface and lower tropospheric temperatures in the deep tropics

in tropical lapse rates (Karl et al. 2006, page 11). As described in Sects. 5.5.2.1 and 5.5.2.2, modeled trends in tropical lapse rates are now broadly consistent with results obtained using RSS T_{2LT} data. Why has this progress occurred?

There are at least two contributory factors. First, the new RSS tropical T_{2LT} trend is over 25% larger than the old trend (0.166 versus 0.130 °C/decade), primarily due to a change in RSS's procedure of adjusting for inter-satellite biases. Adjustments now incorporate a latitudinal dependence (as in Christy et al. 2003), which tends to increase trends in the tropics and decrease trends at mid-latitudes. Second, our work reveals that comparisons of modeled and observed tropical lapse rate changes are sensitive to structural uncertainties in the observed SST data and that these uncertainties may be larger than one would infer from the CCSP report. The tropical SST trends estimated here range from 0.077 °C to 0.108 °C/decade (see Table 5.1), with differences primarily related to different processing choices in the treatment of satellite and buoy data and in the applied infilling and filtering procedures (Smith and Reynolds 2005; Smith et al. 2008; Rayner et al. 2006; Brohan et al. 2006). The smaller observed SST changes in the ERSST-v2 and ERSST-v3 data[16] yield lapse rate trends that are in better accord with model results.

5.6 Experiments with Synthetic Data

The following section compares the performance of d, d^*, and d_1^* under controlled conditions, when the test statistics are applied to synthetic data. We use a standard lag-1 autoregressive (AR) model to generate the synthetic time series $x(t)$:

$$x(t) = a_1\left(x(t-1) - a_m\right) + z(t) + a_m \quad ; \; t = 1,\ldots,n_t \tag{5.14}$$

where a_1 is the coefficient of the AR-1 model, $z(t)$ is randomly generated white noise, and a_m is a mean term. Here, we set a_1 to 0.87 (close to the lag-1 autocorrelation of the monthly-mean UAH and RSS T_{2LT} and T_2 anomaly data; see Table 5.1) and a_m to zero. The noise $z(t)$ is scaled so that $x(t)$ has approximately the same temporal standard deviation as the UAH anomaly data. Each $x(t)$ series has the same length as the observational and model data (252 months), and monthly-mean anomalies were defined as for $y_m(t)$ and $y_o(t)$.

Rejection rate results for these idealized cases are shown in Fig. 5.5 as a function of N, the number of synthetic time series. Consider first the results for our "paired trends" test of hypothesis H_1 (see Sect. 5.4). For each synthetic time series, we calculate the trend b_x and its unadjusted and adjusted standard errors, and then compute the test statistic d for all unique combinations of time series pairs. In the $N = 19$ case, for example (which corresponds to the number of A/OGCMs used in our study), there are 171 unique pairs. Under the assumption that d has a Normal distribution, we determine rejection rates for H_1 at stipulated significance levels of 5%, 10%, and 20%. This procedure was repeated 1000 times, with 1000 different realizations of 19 synthetic time series, allowing us to obtain estimates of the parameters of the underlying rejection rate distributions. We followed a similar process for all other values of N considered.

The paired trend results obtained with adjusted standard errors are plotted as blue lines in Fig. 5.5a. The percentage rejections of hypothesis H_1 (averaged over all values of N) are close to the theoretical expectations:

the 5%, 10%, and 20% significance tests have rejection rates of ca. 6%, 11%, and 21%, respectively (see Supporting Online Material).

This bias of roughly 1% between theoretical and empirically estimated rejection rates is very small compared to the bias that occurs if the paired trends test is applied without adjustment for temporal autocorrelation effects. In the latter case, rejection rates for 5%, 10%, and 20% tests consistently exceed 60%, 65%, and 72% (respectively; see green lines in Fig. 5.5a). Clearly, ignoring the influence of temporal autocorrelation on the estimated number of independent time samples yields incorrect test results.

We now examine tests of hypothesis H_2 with the DCPS07 d^* statistic [Eq. (5.11)] and our d_1^* statistic [Eq. (5.12)]. Consider again the example of the $N = 19$ case. The first time series is designated as the "observations", from which we calculate the trend $b_x(1)$ and its adjusted standard error. With the remaining 18 time series, we compute the ensemble-mean "model" trend, $< b_x >$, and DCPS07's σ_{SE}. We then calculate the test statistics d^* and d_1^*. This is repeated with the trend in the second time series

Fig. 5.5 Performance of statistical tests with synthetic data. Results in panel A are for the "paired trends" test [d; see Eq. (5.3)], in which trends from "observed" temperature time series are tested against trends from individual realizations of "model" 20CEN runs. Two versions of the paired trends test are evaluated, with and without adjustment of trend standard errors for temporal autocorrelation effects. Panel B shows results obtained with the DCPS07 "consistency test" [d^*; see Eq. (5.11)] and a modified version of the DCPS07 test [d_1^*; see Eq. (5.12)] which accounts for statistical uncertainties in the observed trend. In the d^* and d_1^* tests, the "model average" signal trend is compared with the "observed" trend. Synthetic $x(t)$ time series were generated using the standard AR-1 model in Eq. (5.14). Rejection rates for hypotheses H_1 (for the "paired trends" test) and H_2 (for the d^* and d_1^* tests; see Sect. 5.4) are given as a function of N, the total number of synthetic time series, for $N = 5, 6, \ldots 100$. Each test is performed for stipulated significance levels of 5%, 10%, and 20% (denoted by dashed, thin, and bold lines, respectively). For each value of N, rejection rates are the mean of the sampling distribution of rejection rates obtained with 1000 realizations of N synthetic time series. The specified value of the lag-1 autocorrelation coefficient in Eq. (5.14) is close to the sample value of r_1 in the UAH and RSS T_{2LT} data (Table 5.1). Similarly, the noise component of the synthetic $x(t)$ data was scaled to ensure $x(t)$ had (on average) approximately the same temporal standard deviation as the observed T_{2LT} anomaly data. See Sect. 5.6 for further details

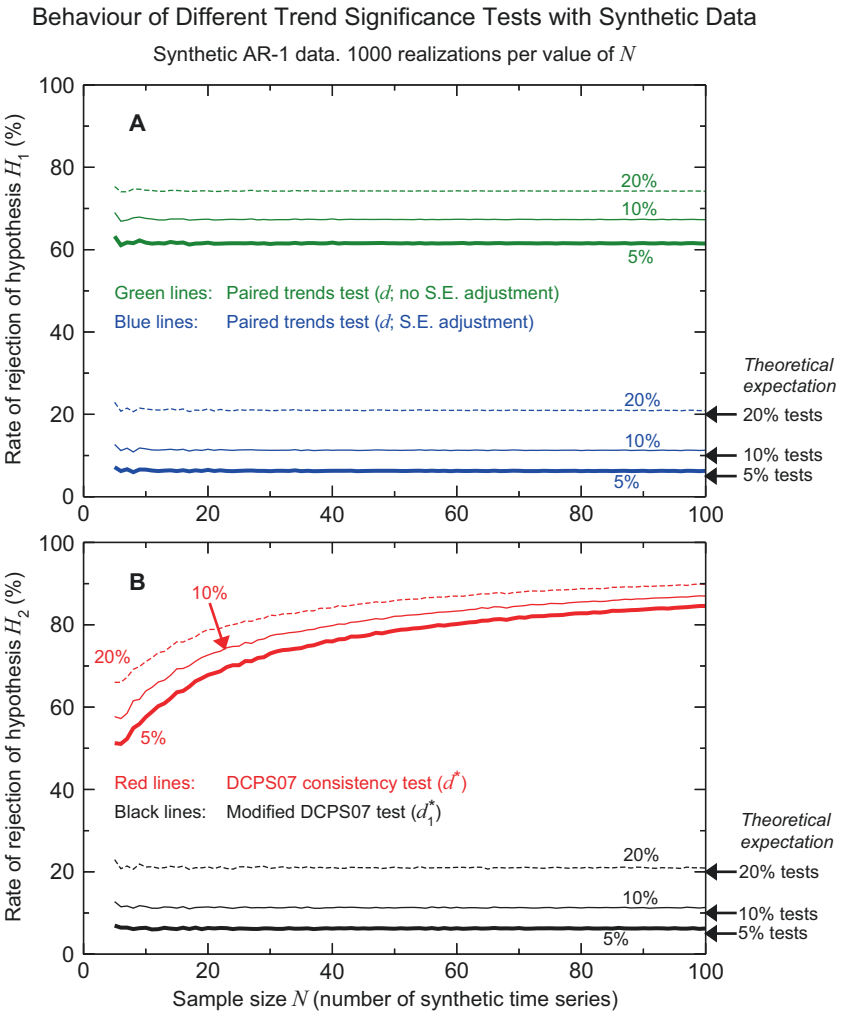

Fig. 5.5 (continued)

as the surrogate observations, and with $< b_x >$ and σ_{SE} calculated from time series 1, 3, 4,…19, etc. For each of the two test statistics, our procedure yields 19 separate tests of hypothesis H_2 (see Sect. 5.4). As for the paired trends test with synthetic data, we repeat this procedure 1000 times, generate distributions of rejection rates at the three stipulated significance levels, and then repeat the process for all other values of N.

Application of the unmodified DCPS07 test to synthetic data leads to alarmingly large rejection rates of H_2 (Fig. 5.5b; red lines). Rejection rates are a function of N. For 5% significance tests, rejection rates rise from 65% to 84% (for $N = 19$ and $N = 100$, respectively). Although DCPS07 refer to this as a "*robust statistical test*", it is clearly flawed, and robust only in its ability to incorrectly reject hypothesis H_2. When our modified version of this test is applied to the same synthetic data, results are strikingly different: rejection rates are within 1–2% of the theoretical expectation values (Fig. 5.5b; black lines).

The lesson from this exercise is that DCPS07's consistency test, when applied to synthetic data generated with the same underlying statistical model, yields incorrect results. It finds a very high proportion of significant differences between "modeled" and "observed" trends, even in a situation where we know a priori that trend differences should occur by chance alone and that the proportion of tests with significant differences should be small. Although these synthetic data simulations are not an exact analogue of the "real world" application of the d^* and d_1^* tests, a test that yields incorrect results under controlled conditions with synthetic data cannot be expected to produce reasonable results in a "real world" application.

5.7 Vertical Profiles of Atmospheric Temperature Trends

DCPS07 also use their consistency test to compare simulated vertical profiles of tropical temperature change with results from radiosondes. They conclude that the multi-model ensemble-mean trend profile, $<< b_m(z) >>$ (where z is a nominal height coordinate), is inconsistent with

the trends inferred from radiosondes. We have shown previously that their test is flawed and yields incorrect results when applied in controlled settings (Sects. 5.5 and 5.6).

A further concern relates to the observational data used by DCPS07. They rely on radiosonde data from HadAT2 (McCarthy et al. 2008), RATPAC version B (Free et al. 2005),[17] RAOBCORE version 1.2 (Haimberger 2007), and the Integrated Global Radiosonde Archive ("IGRA"; Durre et al. 2006). DCSP07 claim that these constitute "*the best available updated observations*". As noted in Sect. 5.1, there are large structural uncertainties in radiosonde-based estimates of atmospheric temperature change (see, e.g., Seidel et al. 2004; Thorne et al. 2005b; Mears et al. 2006). An important question, therefore, is whether DCSP07 accurately represented our best currently available estimates of structural uncertainties in radiosonde data.

To address this question, we first consider the RAOBCORE datasets developed at the University of Vienna ("UnV"). We use three versions of the RAOBCORE data: v1.2 and v1.3, which were described in Haimberger (2007), and v1.4, which was introduced in Haimberger et al. (2008). While RAOBCORE v1.2 shows little net warming of the tropical troposphere over the satellite era, v1.3 and v1.4 exhibit pronounced tropospheric warming, with warming maxima in excess of 0.6 °C/decade at 200 hPa, and cooling of up to 0.1 °C/decade between 700 and 500 hPa (Fig. 5.6a). These large differences in RAOBCORE vertical temperature profiles arise because of different decisions made by the UnV group in the data homogenization process. Although DCPS07 had access to all three RAOBCORE versions, they presented results from v1.2 only.

We also analyze two new radiosonde products, RICH and IUK, which were not available to DCPS07. RICH relies on the same procedure as the RAOBCORE datasets to identify inhomogeneities ("breaks") in radiosonde data. Unlike the RAOBCORE products, however (which use information from the ERA-40 background forecasts for break adjustment), RICH adjusts for breaks with homogeneous information from nearby radiosonde stations (Haimberger et al. 2008). IUK employs a new homogenization procedure in which raw radiosonde data are represented

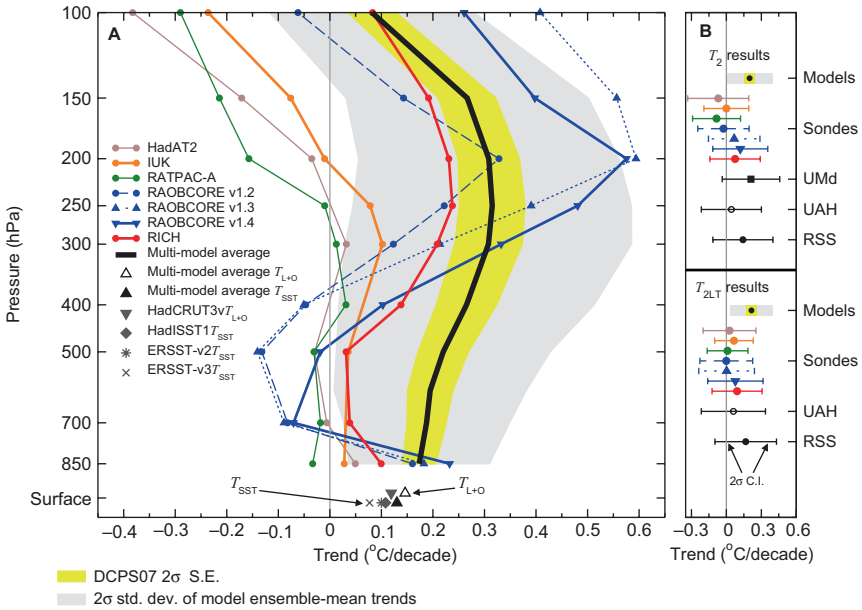

Fig. 5.6 Vertical profiles of trends in atmospheric temperature (panel A) and in actual and synthetic MSU temperatures (panel B). All trends were calculated using monthly-mean anomaly data, spatially averaged over 20°N–20°S. Results in *panel A* are from seven radiosonde datasets (RATPAC-A, RICH, HadAT2, IUK, and three versions of RAOBCORE; see Sect. 5.2.1.2) and 19 different climate models. Tropical T_{SST} and T_{L+O} trends from the same climate models and four different observational datasets (Sect. 5.2.1.3) are also shown. The multi-model average trend at a discrete pressure level, $<< b_m(z) >>$, was calculated from the ensemble-mean trends of individual models [see Eq. (5.7)]. The gray shaded envelope is $s\{< b_m(z) >\}$, the 2σ standard deviation of the ensemble-mean trends at discrete pressure levels. The yellow envelope represents $2\sigma_{SE}$, DCPS07's estimate of uncertainty in the mean trend. For visual display purposes, T_{L+O} results have been offset vertically to make it easier to discriminate between trends in T_{L+O} and T_{SST}. Satellite and radiosonde trends in panel B are plotted with their respective adjusted 2σ confidence intervals (see Sect. 5.4.1). Model results are the multi-model average trend and the standard deviation of the ensemble-mean trends, and gray and yellow shaded areas represent the same uncertainty estimates described in panel A (but now for layer-averaged temperatures rather than temperatures at discrete pressure levels). The *y*-axis in panel B is nominal, and bears no relation to the pressure coordinates in panel A. The analysis period is January 1979 through December 1999, the period of maximum overlap between the observations and most of the model 20CEN simulations. Note that DCPS07 used the same analysis period for model data, but calculated all observed trends over 1979–2004

by a model of step-function changes (associated with instrument biases) and natural climate variability (Sherwood 2007).[18] Both RICH and IUK do not display the prominent lower tropospheric cooling evident in the RAOBCORE, HadAT2, and RATPAC-A products. For comparisons over the period 1979–1999, the multi-model ensemble-mean trend profile in the tropical lower troposphere is closer to the IUK and RICH results than to the changes derived from the other five radiosonde datasets.

The results presented here illustrate that current structural uncertainties in the radiosonde data are substantially larger than one would infer from DCPS07. Different choices in the complex process of dataset construction and homogenization lead to marked differences in both the amplitude and vertical structure of the resulting tropical trends. Temperatures from the most recent homogenization efforts, however, invariably show greater warming in the tropical troposphere than is evident in the raw data upon which they are based. Climate model results are in closer agreement with these newer radiosonde datasets, which were not used by DCPS07.

The model average warming of the tropical surface over 1979–1999 is slightly larger than in the single realization of the observations, both for T_{SST} and T_{L+O} (Fig. 5.6a and Table 5.1). As discussed in Sect. 5.3, this small difference in simulated and observed surface warming rates may be due to the random effects of natural internal variability, model error, or some combination thereof.[19] One important consequence of this difference is that we *expect* the simulated warming in the free troposphere to be generally larger than in observations.

Figure 5.6b summarizes results from a variety of trend comparisons and shows trends in tropical T_{2LT} and T_2 from RSS and UAH, in synthetic MSU temperatures from the seven radiosonde products, and in the model average synthetic MSU temperatures. Results are also given for DCPS07's σ_{SE} and for $s\{<b_m>\}$, the inter-model standard deviation of trends. Application of the DCPS07 consistency test leads to the incorrect conclusion that the model average T_{2LT} and T_2 signal trends are significantly different from the observed signal trends in all radiosonde products. Modification of the test to account for uncertainties in the observed trends leads to very different conclusions. For T_{2LT}, for example, the d_1^*

test statistic [see Eq. (5.12)] indicates that the model average signal trend is not significantly different (at the 5% level) from the observed signal trends in three of the more recent radiosonde products (RICH, IUK, and RAOBCORE v1.4). Clearly, agreement between models and observations depends on both the observations that are selected and the metric used to assess agreement.

5.8 Summary and Conclusions

Several recent comparisons of modeled and observed atmospheric temperature changes have focused on the tropical troposphere (Santer et al. 2006; Douglass et al. 2007; Thorne et al. 2007). Interest in this region was stimulated by an apparent inconsistency between climate model results and observations. Climate models consistently showed tropospheric amplification of surface warming in response to human-caused increases in well-mixed greenhouse gases. In contrast, early versions of satellite and radiosonde datasets implied that the surface had warmed *by more* than the tropical troposphere over the satellite era. This apparent discrepancy has been cited as evidence for the absence of a human effect on climate (e.g., Singer 2008).

A number of national and international assessments have tried to determine whether this discrepancy is real and of practical significance, or an artifact of problems with the observational data (e.g., NRC 2000; Karl et al. 2006; IPCC 2007). The general tenor of these assessments is that structural uncertainties in satellite- and radiosonde-based estimates of tropospheric temperature change are currently large: we do not have an unambiguous observational yardstick for gauging true levels of model skill (or lack thereof). The most comprehensive assessment was the first report produced under the auspices of the U.S. Climate Change Science Program (CCSP; Karl et al. 2006). This report concluded that advances in identifying and adjusting for inhomogeneities in satellite and radiosonde data had helped to resolve the above-described discrepancies, at least at global scales.

In the tropics, however, important differences remained between the simulated and observed "differential warming". In climate models, the

tropical lower troposphere warmed by more than the surface. This amplification of surface warming was timescale-invariant, consistent across a range of models and in accord with basic theoretical considerations (Santer et al. 2005, 2006; Thorne et al. 2007). For month-to-month and year-to-year temperature changes, all satellite and radiosonde datasets showed amplification behavior consistent with model results and basic theory. For multi-decadal changes, however, only two of the then-available satellite datasets (and none of the then-available radiosonde datasets) indicated warming of the troposphere exceeding that of the surface (Karl et al. 2006).

Karl et al. noted that these findings could be interpreted in at least two ways. Under one interpretation, the physical mechanisms controlling real-world amplification behavior vary with timescale, and models have some common error in representing this timescale dependence. The second interpretation posited residual errors in many of the satellite and radiosonde datasets used in the CCSP report. In view of the large structural uncertainties in the observations, the consistency of model amplification results across a range of timescales, and independent evidence of substantial tropospheric warming (Santer et al. 2003, 2007; Paul et al. 2004; Mears et al. 2007; Allen and Sherwood 2008a, b), this was deemed to be the more plausible explanation.

DCPS07 reach a very different conclusion from that of the CCSP report, and claim to find significant differences between models and observations, both for trends in tropospheric temperatures and for trends in lower tropospheric lapse rates. Their claim is based on the application of a "consistency test" to essentially the same model and observational data available to Karl et al. (2006). Their test has two serious flaws: it neglects statistical uncertainty in observed temperature trends arising from interannual temperature variability, and it uses an inappropriate metric [σ_{SE}; see Eq. (5.10)] to judge the statistical significance of differences between the observed trend and the multi-model ensemble-mean trend, $<< b_m >>$.

Consider first the issue of statistical uncertainties. DCPS07 make the implicit assumption that the observed and simulated trends are unaffected by interannual climate variability, and provide perfect information on the true temperature response to external forcing. This assumption is

incorrect, as examination of Figs. 5.1 and 5.2a readily shows: the true response is *not* perfectly known in either observations or the model results. It can only be estimated from a single, noisy observational record and from relatively small ensembles of model results. Any meaningful consistency test must account for the effects of interannual variability, and for the uncertainties it introduces in estimating the underlying (but unknown) "trend signal" in observations. The DCPS07 test does not do this.

Second, DCPS07's σ_{SE} is not a meaningful basis for testing whether a highly uncertain observed trend signal is consistent with the average of imperfectly known model signal trends. This is readily apparent when one applies the DCPS07 test to synthetic data with approximately the same statistical properties as satellite T_{2LT} and T_2 data. In this case, we know a priori that the same statistical model generated the synthetic "observed" and synthetic "simulated" data and that application of the test should yield (on average) rejection of the hypothesis of "no significant difference in signal trends" approximately p% of the time at a stipulated p% significance level. The DCPS07 test, however, gives rejection rates that are many times higher than values expected by chance alone (see Fig. 5.5b).

In contrast to DCPS07, we explicitly account for the effects of interannual variability on observational trends. We do this using two different significance testing strategies. In the first, we use a "paired trends" test [with the d statistic; Eq. (5.3)] that compares each observational trend with the trend from each individual realization of each model. With this procedure, we are testing the hypothesis (H_1) that the trend in an individual model realization of signal plus noise is consistent with the single realization of signal plus noise in the observations. In our second approach, we use a modified version of DCPS07's consistency test [with the d_1^* statistic; Eq. (5.12)], to test the hypothesis (H_2) that the model average signal trend is consistent with the signal trend estimated from the single realization of the observations. With the d test, very few of the model trends in tropical T_{2LT} and T_2 over 1979–1999 are significantly different from RSS or UAH trends (Table 5.2). Similarly, when the d_1^* test is applied to T_{2LT} and T_2 trends, hypothesis H_2 cannot be rejected at the nominal 5% level (Table 5.3).

A more stringent test of model performance involves trends in the time series of differences between surface and lower tropospheric temperature anomalies. Trends in T_{SST} (or T_{L+O}) minus T_{2LT} provide a simple measure of changes in lapse rate. Differencing reduces the amplitude of the (common) unforced variability in surface temperature and T_{2LT}, and makes it easier to identify true model errors in the forced component of lapse rate trends.

While tests involving trends in T_{2LT} and T_2 time series almost invariably showed nonsignificant differences between models and satellite data (Sect. 5.5.1), results for lapse rate trends are more sensitive to structural uncertainties in observations (Sect. 5.5.1.1). If RSS T_{2LT} data are used for computing lapse rate trends, the warming aloft is larger than at the surface (consistent with model results). Very few simulated lapse rate trends differ significantly from observations in "paired trends" tests (Table 5.5). When the d_1^* test is applied, there is *no case* in which hypothesis H_2 can be rejected at the nominal 5% level (Table 5.6).

When UAH T_{2LT} data are used, the warming aloft is smaller than at the surface. Even in the UAH case, however, hypothesis H_1 is not rejected consistently. Rejection rates for "paired trends" tests conducted at the 5% significance level range from ca. 31% to 88%, depending on the choice of observational surface temperature dataset (Table 5.5). Alternately, our modified version of the DCPS07 test reveals that hypothesis H_2 is rejected at the nominal 5% level in all cases involving UAH-based estimates of lapse rate changes (Table 5.6).

Our findings do not bring final resolution to the issue of whether UAH or RSS provide more reliable estimates of temperature changes in the tropical troposphere. We note, however, that the RSS-based estimates of tropical lapse rate changes are in better accord with satellite datasets developed by the UMd and NOAA/NESDIS groups (Vinnikov and Grody 2006; Zho et al. 2006), with newer radiosonde datasets (e.g., Haimberger et al. 2008; Titchner et al. 2008; Allen and Sherwood 2008a, b; Sherwood et al. 2008), and with basic moist adiabatic lapse rate theory. Furthermore, RSS results show amplification of tropical surface warming across a range of timescales (consistent with model behavior), whereas UAH T_{2LT} data yield amplification for monthly and annual temperature

changes, but not for decadal changes. If the UAH results were correct, the physics controlling the response of the tropical atmosphere to surface warming must vary with timescale. Mechanisms that might govern such behavior have not been identified.

Model errors in forcing and response must also contribute to remaining differences between simulated and observed lapse rate trends. For example, only nine of the 19 models used in our study attempted to represent the climate forcing associated with the eruptions of El Chichón and Pinatubo (Forster and Taylor 2006). Statistical comparisons between modeled and observed temperature changes can be sensitive to the inclusion or exclusion of volcanic forcing (Santer et al. 2001; Wigley et al. 2005; Lanzante 2007).

Similarly, roughly half of the models analyzed here exclude stratospheric ozone depletion, which has a pronounced impact on lower stratospheric and upper tropospheric temperatures, and hence on T_2 (Santer et al. 2006). Even models which include some form of stratospheric ozone depletion do not correctly represent the observed profile of ozone losses below ca. 20 km in the tropics (Forster et al. 2007). The latter deficiency may have considerable impact on model-predicted temperature changes above the tropical tropopause and in the uppermost troposphere, and hence on agreement with observations.

In summary, considerable scientific progress has been made since the first report of the U.S. Climate Change Science Program (Karl et al. 2006). There is no longer a serious and fundamental discrepancy between modeled and observed trends in tropical lapse rates, despite DCPS07's incorrect claim to the contrary. Progress has been achieved by the development of new T_{SST}, T_{L+O}, and T_{2LT} datasets, better quantification of structural uncertainties in satellite- and radiosonde-based estimates of tropospheric temperature change, and the application of rigorous statistical comparisons of modeled and observed changes.

We may never completely reconcile the divergent observational estimates of temperature changes in the tropical troposphere. We lack the unimpeachable observational records necessary for this task. The large structural uncertainties in observations hamper our ability to determine how well models simulate the tropospheric temperature changes that

actually occurred over the satellite era. A truly definitive answer to this question may be difficult to obtain. Nevertheless, if structural uncertainties in observations and models are fully accounted for, a partial resolution of the long-standing "differential warming" problem has now been achieved. The lessons learned from studying this problem can and should be applied toward the improvement of existing climate monitoring systems, so that future model evaluation studies are less sensitive to observational ambiguity.

5.9 Acknowledgments

We acknowledge the modeling groups for providing their simulation output for analysis, PCMDI for collecting and archiving this data, and the World Climate Research Programme's Working Group on Coupled Modelling for organizing the model data analysis activity. The CMIP-3 multi-model dataset is supported by the Office of Science, U.S. Department of Energy. The authors received support from a Distinguished Scientist Fellowship of the U.S. Dept. of Energy, Office of Biological and Environmental Research (BDS); the joint DEFRA and MoD Programme (PWT; contracts GA01101 and CBC/2B/0417 Annex C5, respectively); grant P18120-N10 of the Austrian Science Funds (LH); and the NOAA Office of Climate Programs ("Climate Change, Data and Detection") grant NA87GP0105 (TMLW). We thank Mike MacCracken (Climate Institute), David Parker (U.K. Meteorological Office Hadley Centre), Dick Reynolds (National Climatic Data Center), Dian Seidel (NOAA Air Resources Laboratory), Francis Zwiers (Environment Canada), and an anonymous reviewer for useful comments and discussion. Dave Easterling and Imke Durre (National Climatic Data Center) and R. Dobosy and Jenise Swall (NOAA Air Resources Laboratory) provided helpful comments in the course of NOAA internal reviews. Observed MSU data were kindly provided by John Christy (UAH) and Konstantin Vinnikov (UMd). Observed surface temperature data were provided by John Kennedy at the U.K. Meteorological Office Hadley Centre (HadISST1) and by Dick Reynolds at the National Climatic Data Center (ERSST-v2 and ERSST-v3).

Appendix: Statistical Notation

Subscripts and indices	
m	Subscript denoting model data
o	Subscript denoting observational data
t	Index over time (in months)
i	Index over number of models
j	Index over number of 20CEN realizations
z	Index over number of atmospheric levels

Sample sizes	
n_t	Total number of time samples (usually 252)
n_e	Effective number of time samples, adjusted for temporal autocorrelation
n_m	Total number of models (19)
$n_r\ (i)$	Total number of 20CEN realizations for the i^{th} model
N	Total number of synthetic time series

Time series	
$y_m(t)$	Simulated T_{2LT} or T_2 time series
$\phi_m(t)$	Underlying signal in $y_m(t)$ in response to forcing
$\eta_m(t)$	Realization of internally generated noise in $y_m(t)$
$x(t)$	Synthetic AR-1 time series
$z(t)$	Synthetic noise time series

Trends	
b_m	Least-squares linear trend in an individual $y_m(t)$ time series
$< b_m(i) >$	Ensemble-mean trend in the i^{th} model
$<< b_m >>$	Multi-model ensemble-mean trend
$<< b_m(z) >>$	Multi-model ensemble-mean trend profile

Standard errors and standard deviations	
$s\{b_m\}$	Standard error of b_m
$s\{y_m(t)\}$	Temporal standard deviation of $y_m(t)$ anomaly time series
$s\{< b_m >\}$	Standard deviation of ensemble-mean trends
$s\{< b_m(z) >\}$	Standard deviation of ensemble-mean trends at discrete pressure levels
σ_{SE}	DCPS07 "*estimate of the uncertainty of the mean*"

Other regression terms	
$e(t)$	Regression residuals
r_1	Lag-1 autocorrelation of regression residuals

Test statistics	
d	Paired trends test statistic [Eq. (5.3)]
d^*	Test statistic for original DCPS07 consistency test [Eq. (5.11)]
d_1^*	Test statistic for modified version of DCPS07 consistency test [Eq. (5.12)]

Notes

1. See Table 3.4 in Lanzante et al. (2006). For the specific period 1979–2004, tropical (20 °N–20 °S) T_2 trends range from 0.05 °C/decade (UAH) to 0.19 °C/decade (UMd), while T_{2LT} trends span the range 0.05 °C/decade (UAH) to 0.15 °C/decade (RSS). The most important sources of uncertainty are likely to be "*due to inter-satellite calibration offsets and calibration drifts*" (Mears et al. 2006, page 78).
2. The UMd and NOAA/NESDIS groups do not provide a T_{2LT} product. Because of their calibration procedure, the NOAA/NESDIS T_2 data are only available for a shorter period (1987 to present) than the T_2 products of the three other groups.
3. A more recent version of the RSS T_2 and T_{2LT} datasets (version 3.1) now exists. RSS versions 3.0 and 3.1 are virtually identical over the primary analysis period considered here (1979–1999). For UAH data, a version 5.2 exists for T_{2LT} but not for T_2 data, for which only version 5.1 is available.
4. RAOBCORE stands for *RAdiosonde OBservation COrrection using REanalysis*.
5. All simulations included human-induced changes in well-mixed GHGs and the direct (scattering) effects of sulfate aerosols on incoming solar radiation. Other external forcings (such as changes in ozone, carbonaceous aerosols, indirect effects of aerosols on clouds, land surface properties, solar irradiance, and volcanic dust loadings) were not handled uniformly across different modeling groups. For further details of the applied forcings, see Santer et al. (2005, 2006).

6. DCPS07 used a larger set of 20CEN runs (67 simulations performed with 22 different models) and incorporated model results that were not available at the time of the Santer et al. (2005) study. This difference in the number of 20CEN models employed in the two investigations is immaterial for illustrating the statistical problems in the consistency test applied by DCPS07. All 49 simulations employed in our current work were also analyzed by DCSP07.
7. Amplification occurs due to the nonlinear effect of the release of latent heat by moist ascending air in regions experiencing convection.
8. The 20CEN experiments analyzed here were performed with coupled atmosphere-ocean General Circulation Models (A/OGCMs) driven by estimates of historical changes in external forcing. Due to chaotic variability in the climate system, small differences in the atmospheric or oceanic initial conditions at the start of the 20CEN run (typically in the mid- to late nineteenth century) rapidly lead to different manifestations of climate noise. Within the space of several months, the state of the atmosphere is essentially uncorrelated with the initial state. This means that even the same model, when run many times with identical external forcings (but each time from slightly different initial conditions), produces many different samples of $\eta_m(t)$, each superimposed on the same underlying signal, $\phi_m(t)$.
9. Our d_1^* test involving the multi-model ensemble-mean trend [see Eq. (5.12)], also relies on an AR-1 model to estimate n_e and adjust the observed standard error, and is therefore also likely to be too liberal.
10. We use < > to denote an ensemble average over multiple 20CEN realizations performed with a single model. Double angle brackets, << >>, indicate a multi-model ensemble average.
11. Under this assumption, the total uncertainty in $<< b_m >> - b_o$ is determined solely by inter-model trend differences arising from structural differences between the models [see Eqs. (5.9, 5.10, and 5.11)]. As discussed in Sect. 5.3, however, the total uncertainty in the magnitude of $<< b_m >> - b_o$ reflects not only these structural differences, but also inter-model differences in internal variability and ensemble size.
12. Inter-model differences in the size of the confidence intervals in Fig. 5.3a are due primarily to differences in the amplitude and temporal autocorrelation properties of $\eta_m(t)$, but are also affected by neglect or inclusion of the effects of volcanic forcing (see Santer et al. 2005,

2006). Models with large ENSO variability (such as GFDL-CM2.1 and FGOALS-g1.0) have large adjusted confidence intervals, while A/OGCMs with relatively coarse-resolution, diffusive oceans (such as GISS-AOM) have much weaker ENSO variability and smaller values of $s\{b_m\}$.

13. We have explored the sensitivity of our adjusted standard errors and significance test results to choices of averaging period ranging from two to 12 months. These choices span a wide range of temporal autocorrelation behavior. Results for the d test are relatively insensitive to the selected averaging period, suggesting that our adjustment method is reasonable.
14. Two layers (T_{2LT} and T_2) × two observational datasets (RSS and UAH).
15. One of the assumptions underlying the d_1^* test (and all tests performed here) is that structural uncertainty in the observations is negligible (see Sect. 5.4.2). We know this is not the case in the real world (see, e.g., Seidel et al. 2004; Thorne et al. 2005; Lanzante et al. 2006; Mears et al. 2006). In the present study, we have examined the effects of structural uncertainties in satellite and radiosonde data by treating each observational dataset independently, and assessing the robustness of our model-versus-observed trend comparisons to different dataset choices. An alternative approach would be to explicitly include a structural uncertainty term for the observations in the test statistic itself.
16. These datasets were not examined in DCPS07 or in Santer et al. (2005, 2006).
17. Note that RATPAC-B is unadjusted after 1997. RATPAC-A, which we use here, accounts for inhomogeneities before and after 1997.
18. Sherwood et al. (2008) argue that this procedure does not completely homogenize data from stations between 5°S and 20°N, since trends at these stations remained highly variable and (on average) unphysically low compared to those at neighboring latitudes that are much more accurately known. The implication is that gradual (rather than step-like) changes in bias at many tropical stations may not be reliably identified and adjusted by the IUK homogenization procedure. If this is the case, the IUK trends shown here are likely to be underestimates of the true trends.
19. An error in the model average surface warming is entirely likely given the neglect of indirect aerosol effects in roughly half of the models analyzed here.

References

Allen, R.J., and Sherwood, S.C. 2008a. Utility of Radiosonde Wind Data in Representing Climatological Variations of Tropospheric Temperature and Baroclinicity in the Western Tropical Pacific. *Journal of Climate* (in press).

———. 2008b. Warming Maximum in the Tropical Upper Troposphere Deduced from Thermal Winds. *Nature Geoscience* (in press).

Brohan, P., J.J. Kennedy, I. Harris, S.F.B. Tett, and P.D. Jones. 2006. Uncertainty Estimates in Regional and Global Observed Temperature Changes: A New Dataset from 1850. *Journal of Geophysical Research* 111: D12106. https://doi.org/10.1029/2005JD006548.

Christy, J.R., R.W. Spencer, and W.D. Braswell. 2000. MSU Tropospheric Temperatures: Data Set Construction and Radiosonde comparisons. *Journal of Atmospheric and Oceanic Technology* 17: 1153–1170.

Christy, J.R., R.W. Spencer, W.B. Norris, W.D. Braswell, and D.E. Parker. 2003. Error Estimates of Version 5.0 of MSU/AMSU Bulk Atmospheric Temperatures. *Journal of Atmospheric and Oceanic Technology* 20: 613–629.

Christy, J.R., W.B. Norris, R.W. Spencer, and J.J. Hnilo. 2007. Tropospheric Temperature Change Since 1979 from Tropical Radiosonde and Satellite Measurements. *Journal of Geophysical Research* 112: D06102. https://doi.org/10.1029/2005JD006881.

Douglass, D.H., B.D. Pearson, and S.F. Singer. 2004. Altitude Dependence of Atmospheric Temperature trends: Climate Models Versus Observations. *Geophysical Research Letters* 31: L13208. https://doi.org/10.1029/2004/GL020103.

Douglass, D.H., J.R. Christy, B.D. Pearson, and S.F. Singer. 2007. A Comparison of Tropical Temperature Trends with Model Predictions. *International Journal of Climatology* 27. https://doi.org/10.1002/joc.1651.

Durre, I., R. Vose, and D.B. Wuertz. 2006. Overview of the Integrated Global Radiosonde Archive. *Journal of Climate* 19: 53–68.

Forster, P.M., and K.E. Taylor. 2006. Climate Forcings and Climate Sensitivities Diagnosed from Coupled Climate Model Integrations. *Journal of Climate* 19: 6181–6194.

Forster, P.M., G. Bodeker, R. Schofield, and S. Solomon. 2007. Effects of Ozone Cooling in the Tropical Lower Stratosphere and Upper Troposphere. *Geophysical Research Letters* 34: L23813. https://doi.org/10.1029/2007GL031994.

Free, M., et al. 2005. Radiosonde Atmospheric Temperature Products for Assessing Climate (RATPAC): A New Dataset of Large-Area Anomaly Time Series. *Journal of Geophysical Research* 110: D22101. https://doi.org/10.1029/2005JD006169.

Gaffen, D., et al. 2000. Multi-decadal Changes in the Vertical Temperature Structure of the Tropical Troposphere. *Science* 287: 1239–1241.

Haimberger, L. 2007. Homogenization of Radiosonde Temperature Time Series Using Innovation Statistics. *Journal of Climate* 20: 1377–1403.

Haimberger, L., C. Tavolato, and S. Sperka. 2008. Towards Elimination of the Warm Bias in Historic Radiosonde Temperature Records – Some New Results from a Comprehensive Intercomparison of Upper Air Data. *Journal of Climate* (in press).

Hegerl, G.C., et al. 2007. Understanding and Attributing Climate Change. In *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, ed. S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor, and H.L. Miller. Cambridge/New York: Cambridge University Press.

IPCC (Intergovernmental Panel on Climate Change). 1996. Summary for Policy-Makers. In *Climate Change 1995: The Science of Climate Change*, Contribution of Working Group I to the Second Assessment Report of the Intergovernmental Panel on Climate Change, ed. J.T. Houghton et al. Cambridge/New York: Cambridge University Press.

———. 2001. Summary for Policy-Makers. In *Climate Change 2001: The Scientific Basis*, Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change, ed. J.T. Houghton et al. Cambridge/New York: Cambridge University Press.

———. 2007. Summary for Policy-Makers. In *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, ed. S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor, and H.L. Miller. Cambridge/New York: Cambridge University Press.

Karl, T.R., S.J. Hassol, C.D. Miller, and W.L. Murray (eds.). 2006. *Temperature Trends in the Lower Atmosphere: Steps for Understanding and Reconciling Differences.* A Report by the U.S. Climate Change Science Program and the Subcommittee on Global Change Research. National Oceanic and Atmospheric Administration, National Climatic Data Center, Asheville, p. 164.

Lanzante, J.R. 2005. A Cautionary Note on the Use of Error Bars. *Journal of Climate* 18: 3699–3703.

———. 2007. Diagnosis of Radiosonde Vertical Temperature Trend Profiles: Comparing the Influence of Data Homogenization Versus Model Forcings. *Journal of Climate* 20 (21): 5356–5364.

Lanzante, J.R., S.A. Klein, and D.J. Seidel. 2003. Temporal Homogenization of Monthly Radiosonde Temperature Data. Part II: Trends, Sensitivities, and MSU Comparison. *Journal of Climate* 16: 241–262.

Lanzante, J.R., T.C. Peterson, F.J. Wentz, and K.Y Vinnikov. 2006. What Do Observations Indicate About the Change of Temperatures in the Atmosphere and at the Surface Since the Advent of Measuring Temperatures Vertically? In *Temperature Trends in the Lower Atmosphere: Steps for Understanding and Reconciling Differences*, ed. T.R. Karl, S.J. Hassol, C.D. Miller, and W.L. Murray. A Report by the U.S. Climate Change Science Program and the Subcommittee on Global Change Research.

Manabe, S., and R.J. Stouffer. 1980. Sensitivity of a Global Climate Model to an Increase of CO2 Concentration in the Atmosphere. *Journal of Geophysical Research* 85: 5529–5554.

McCarthy, M.P., H.A. Titchner, P.W. Thorne, Tett SFB, L. Haimberger, and D.E. Parker. 2008. Assessing Bias and Uncertainty in the HadAT Adjusted Radiosonde Climate Record. *Journal of Climate* 21: 817–832.

Mears, C.A., and F.J. Wentz. 2005. The Effect of Diurnal Correction on Satellite-Derived Lower Tropospheric Temperature. *Science* 309: 1548–1551.

Mears, C.A., M.C. Schabel, and F.J. Wentz. 2003. A Reanalysis of the MSU Channel 2 Tropospheric Temperature Record. *Journal of Climate* 16: 3650–3664.

Mears, C.A., C.E. Forest, R.W. Spencer, R.S. Vose, and R.W. Reynolds. 2006. What Is Our Understanding of the Contributions Made by Observational or Methodological Un-certainties to the Previously-Reported Vertical Differences in Temperature Trends? In *Temperature Trends in the Lower Atmosphere: Steps for Understanding and Reconciling Differences,* ed. T.R. Karl, S.J. Hassol, C.D. Miller, and W.L. Murray. A Report by the U.S. Climate Change Science Program and the Subcommittee on Global Change Research.

Mears, C.A., B.D. Santer, F.J. Wentz, K.E. Taylor, and M.F. Wehner. 2007. Relationship Between Temperature and Precipitable Water Changes Over Tropical Oceans. *Geophysical Research Letters* 34: L24709. https://doi.org/10.1029/2007GL031936.

Mitchell, J.F.B., et al. 2001. Detection of Climate Change and Attribution of Causes. In *Climate Change 2001: The Scientific Basis*, Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change, ed. J.T. Houghton et al., 881. Cambridge, UK/New York: Cambridge University Press.

NRC (National Research Council). 2000. *Reconciling Observations of Global Temperature Change*. Washington, DC: National Academy Press. 85 pp.

———. 2005. *Radiative Forcing of Climate Change: Expanding the Concept and Addressing Uncertainties*, 168. Washington, DC: National Academy Press.

Paul, F., A. Kaab, M. Maisch, T. Kellenberger, and W. Haeberli. 2004. Rapid Disintegration of Alpine Glaciers Observed with Satellite Data. *Geophysical Research Letters* 31: L21402. https://doi.org/10.1029/2004GL020816.

Randel, W.J., and F. Wu. 2006. Biases in Stratospheric and Tropospheric Temperature Trends Derived from Historical Radiosonde Data. *Journal of Climate* 19: 2094–2104.

Rayner, N.A., et al. 2003. Global Analyses of Sea Surface Temperature, Sea Ice, and Night Marine Air Temperature Since the Late Nineteenth Century. *Journal of Geophysical Research* 108: 4407. https://doi.org/10.1029/2002JD002670. HadISST1 data are available at http://www.hadobs.org/

———. 2006. Improved Analyses of Changes and Uncertainties in Marine Temperature Measured in Situ Since the Mid-nineteenth Century: The HadSST2 Dataset. *Journal of Climate* 19: 446–469.

Santer, B.D., T.M.L. Wigley, T.P. Barnett, and E. Anyamba. 1996. Detection of Climate Change and Attribution of Causes. In *Climate Change 1995: The Science of Climate Change*, Contribution of Working Group I to the Second Assessment Report of the Intergovernmental Panel on Climate Change, ed. J.T. Houghton et al., 572. Cambridge, UK/New York: Cambridge University Press.

Santer, B.D., et al. 1999. Uncertainties in Observationally Based Estimates of Temperature Change in the Free Atmosphere. *Journal of Geophysical Research* 104: 6305–6333.

———. 2000a. Statistical Significance of Trends and Trend Differences in Layer-Average Atmospheric Temperature Time Series. *Journal of Geophysical Research* 105: 7337–7356.

———. 2000b. Interpreting Differential Temperature Trends at the Surface and in the Lower Troposphere. *Science* 287: 1227–1232.

———. 2001. Accounting for the Effects of Volcanoes and ENSO in Comparisons of Modeled and Observed Temperature Trends. *Journal of Geophysical Research* 106: 28033–28059.

———. 2003. Contributions of Anthropogenic and Natural Forcing to Recent Tropopause Height Changes. *Science* 301: 479–483.

———. 2005. Amplification of Surface Temperature Trends and Variability in the Tropical Atmosphere. *Science* 309: 1551–1556.

Santer, B.D., J.E. Penner, and P.W. Thorne. 2006. How Well Can the Observed Vertical Temperature Changes Be Reconciled with Our Understanding of the Causes of These Changes? In *Temperature Trends in the Lower Atmosphere: Steps for Understanding and Reconciling Differences,* ed. T.R. Karl, S.J Hassol., C.D. Miller, W.L. Murray. A Report by the U.S. Climate Change Science Program and the Subcommittee on Global Change Research.

Santer, B.D., et al. 2007. Identification of Human-Induced Changes in Atmospheric Moisture Content. *Proceedings of the National Academy of Sciences* 104: 15248–15253.

Seidel, D.J., et al. 2004. Uncertainty in Signals of Large-Scale Climate Variations in Radiosonde and Satellite Upper-Air Temperature Data Sets. *Journal of Climate* 17: 2225–2240.

Sherwood, S.C. 2007. Simultaneous Detection of Climate Change and Observing Biases in a Network with Incomplete Sampling. *Journal of Climate* 20: 4047–4062.

Sherwood, S.C., J.R. Lanzante, and C.L. Meyer. 2005. Radiosonde Daytime Biases and Late- 20th Century Warming. *Science* 309: 1556–1559.

Sherwood, S.C., C.L. Meyer, R.J. Allen, and H.A. Titchner. 2008. Robust Tropospheric Warming Revealed by Iteratively Homogenized Radiosonde Data. *Journal of Climate*. https://doi.org/10.1175/2008JCLI2320.1.

Singer, S.F. 2001. Global Warming: An Insignificant Trend? *Science* 292: 1063–1064.

———. 2008. In *Nature, Not Human Activity, Rules the Climate: Summary for Policymakers of the Report of the Nongovernmental International Panel on Climate Change*, ed. S.F. Singer. Chicago: The Heartland Institute.

Smith, T.M., and R.W. Reynolds. 2005. A Global Merged Land and Sea Surface Temperature Reconstruction Based on Historical Observations (1880–1997). *Journal of Climate* 18: 2021–2036.

Smith, T.M., R.W. Reynolds, T.C. Peterson, and J. Lawrimore. 2008. Improvements to NOAA's Historical Merged Land-Ocean Surface Temperature Analysis (1880–2006). *Journal of Climate* (in press).

Spencer, R.W., and J.R. Christy. 1990. Precise Monitoring of Global Temperature Trends from Satellites. *Science* 247: 1558–1562.

Storch, H., and F.W. Zwiers. 1999. *Statistical Analysis in Climate Research*, 484. Cambridge: Cambridge University Press.

Thiébaux, H.J., and F.W. Zwiers. 1984. The Interpretation and Estimation of Effective Sample Size. *Journal of Meteorology and Applied Climatology* 23: 800–811.

Thorne, P.W., et al. 2005a. Uncertainties in Climate Trends: Lessons from Upper-Air Temperature Records. *Bulletin of the American Meteorological Society* 86: 1437–1442.

———. 2005b. Revisiting Radiosonde Upper-Air Temperatures from 1958 to 2002. *Journal of Geophysical Research* 110: D18105. https://doi.org/10.1029/2004JD005753.

———. 2007. Tropical Vertical Temperature Trends: A Real Discrepancy? *Geophysical Research Letters* 34: L16702. https://doi.org/10.1029/2007GL029875.

Titchner, H.A., P.W. Thorne, M.P. McCarthy, S.F.B. Tett, L. Haimberger, and D.E. Parker. 2008. Critically Reassessing Tropospheric Temperature Trends from Radiosondes Using Realistic Validation Experiments. *Journal of Climate*. https://doi.org/10.1175/2008JCLI2419.1.

Trenberth, K.E., et al. 2007. Observations: Surface and Atmospheric Climate Change. In *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, ed. S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor, and H.L. Miller. Cambridge, UK/New York: Cambridge University Press.

Uppala, S.M., et al. 2005. The ERA-40 Reanalysis. *Quarterly Journal of the Royal Meteorological Society* 131: 2961–3012.

Vinnikov, K.Y., and N.C. Grody. 2003. Global Warming Trend of Mean Tropospheric Temperature Observed by Satellites. *Science* 302: 269–272.

Vinnikov, K.Y., et al. 2006. Temperature Trends at the Surface and the Troposphere. *Journal of Geophysical Research* 111: D03106. https://doi.org/10.1029/2005jd006392.

Wentz, F.J., and M. Schabel. 1998. Effects of Orbital Decay on Satellite-Derived Lower- Tropospheric Temperature Trends. *Nature* 394: 661–664.

———. 2000. Precise Climate Monitoring Using Complementary Satellite Data Sets. *Nature* 403: 414–416.

Wigley, T.M.L. 2006. Appendix A: Statistical Issues Regarding Trends. In *Temperature Trends in the Lower Atmosphere: Steps for Understanding and Reconciling Differences,* ed. T.R. Karl, S.J. Hassol, C.D. Miller, W.L. Murray. A Report by the U.S. Climate Change Science Program and the Subcommittee on Global Change Research.

Wigley, T.M.L., C.M. Ammann, B.D. Santer, and S.C.B. Raper. 2005. The Effect of Climate Sensitivity on the Response to Volcanic Forcing. *Journal of Geophysical Research* 110: D09107. https://doi.org/10.1029/2004/JD005557.

Wilks, D.S. 1995. *Statistical Methods in the Atmospheric Sciences*, 467 pp. San Diego: Academic Press.

Zho, C.-Z., et al. 2006. Recalibration of Microwave Sounding Unit for Climate Studies Using Simultaneous Nadir Overpasses. *Journal of Geophysical Research* 111: D19114. https://doi.org/10.1029/2005JD006798.

Zwiers, F.W., and H. von Storch. 1995. Taking Serial Correlation into Account in Tests of the Mean. *Journal of Climate* 8: 336–351.

6

The Role of "Complex" Empiricism in the Debates About Satellite Data and Climate Models

Elisabeth A. Lloyd

6.1 Introduction

In January 2015, climate scientist John Christy, an expert in satellite and weather balloon data, claimed in a US Congressional hearing that those data contradicted what the climate models said about the greenhouse effect, namely that it has had a significant impact on climate.[1] But this is a view held by few other climate scientists today (Thorne et al. 2011). In this chapter, I offer a case study to illustrate how different foundational approaches to data and models underpinned a two-decade-long debate about the existence of and evidence for the greenhouse effect. On one side, the climate models appeared to many, including Christy, to be

Much of the content of this chapter was taken from an article with the same title in *Studies in History and Philosophy of Science* 43: 390–401, 2012.

E.A. Lloyd (✉)
History and Philosophy of Science and Medicine Department,
Indiana University, Bloomington, IN, USA

E.A. Lloyd, E. Winsberg (eds.), *Climate Modelling*,
https://doi.org/10.1007/978-3-319-65058-6_6

falsified by the satellite and weather balloon data. The models all clearly predicted warming of the tropical troposphere, and there were not one but two kinds of datasets giving actual observations about the temperature that appeared to contradict the models' predictions. But both of these kinds of datasets were largely rejected by many modelers, with some data analysts on their side. The modelers did appeal to other observational evidence in their defense, but they seemed to be resisting the most obvious, forceful, and believable evidence concerning temperature trends. From outside the profession, it seemed indefensible. However, there were also some data analysts who, like the modelers, did not accept these data as falsifying the models, precisely because they would not accept the datasets as direct reflections of reality.[2]

Understanding this scientific episode requires some philosophical tools that go beyond comparing models with data in a straightforward way. It also requires a more sophisticated understanding of how observational evidence is put together in climate science. We are used to thinking of the data as either confirming or disconfirming the models in question, as is usually assumed under the hypothetico-deductive (H-D) approach to theory testing and confirmation, or its Popperian relative. The model is taken on one hand, predictions are made from it, and those predictions are compared to observations, resulting in the confirmation or disconfirmation of the model. In this case, the epistemic dynamics were very different, and thus very challenging philosophically. Recent philosophical work on models and measurement has investigated how data and observations are inevitably laden with assumptions and theory, all in the context of the actual practice of science (De Chadarevian and Hopwood 2004; Giere 2006; Morgan and Morrison 1999; van Fraassen 2008). I use some of this work as a springboard to elaborate here a widely applicable view I call "complex" empiricism, which also encompasses my approach to model testing and evaluation that differs significantly from the H-D view (Lloyd 1987, 1994, 2010, 2015). In the end (and in short), it now appears that the models were mostly right and the early data were mostly wrong, and therein lies an interesting story about data and their relations to scientists, models, and reality.

The tropical troposphere—the layer of atmosphere between the earth's surface and the stratosphere in the tropics—is a crucial piece of real estate

in the arguments about global warming. Physical theory and the global simulation models based on that theory predict that the tropical troposphere will warm faster than the surface as the greenhouse effect takes hold. Measurements of the temperature of the surface do indicate significant warming in the tropics, and have for some time. But when climate scientists first attempted to use satellites to measure atmospheric temperature trends in 1990, these satellite measurements indicated that the tropical troposphere, unlike the surface of the earth, was not warming (Spencer and Christy 1990). This provided empirical support for the view that there was no global warming occurring, and no greenhouse effect. The satellite data were trumpeted by global warming and greenhouse skeptics like Rush Limbaugh, who discussed the data on his conservative radio show as proof against global climate change.[3]

Consider the state of the science 10 years later, in the year 2000: The National Academy of Sciences had been brought in by Congress to address the issues of whether the satellite data and the radiosondes (weather balloons) really were in conflict with the global climate models, among other things. Did the apparent lack of a warming trend in the tropical troposphere indicate that the models really were untrustworthy? Were the satellite data themselves trustworthy, or not well enough developed and not firm enough to overthrow the climate community's trust in the global climate models? How large were the uncertainties (or errors) in the satellite data, the radiosonde data, and the models?

The National Academy of Sciences' (NAS) National Research Council Report, *Reconciling Observations of Global Temperature Change* (Wallace et al. 2000), was written by a distinguished array of modelers and data analysts on both sides of the argument. In the report, the climate scientists approached the disparity between the documented rise in temperature at the earth's surface since 1980 (on the order of 0.25–0.4 degree Celsius) and the apparent lack of a commensurate rise in temperature in the troposphere during that same time period, especially in the tropics (0–0.2 degree Celsius). The panel emphasized various natural and human-made causes that may have prevented the tropospheric temperature from its expected rise, including volcanic eruptions and ozone depletion in the stratosphere. Still, the report characterized the gap between the surface temperature and the tropospheric temperature trends as a "substantial

disparity" (2000, p. 2). Some interpreted this discrepancy as showing that the surface temperature trend was erroneous, while others concluded that the satellite dataset (or, as philosophers often say, "models of data" or "data models"), or the algorithms used to produce that dataset, must be erroneous.

The Microwave Sounding Units (MSUs) mounted in the satellites measure the microwave radiation emitted by oxygen molecules (called "radiance") at a number of different "channels," each of which covers a different set of elevations of the atmosphere. Radiance is then mapped onto temperature values at different elevations. The NAS report covered measurements from Channel 2, which included altitudes from just above the surface up to about 15 kilometers. To eliminate the influence of stratospheric radiation, complicated algorithms to process the microwave radiation into temperatures were required, and often revised. The report was working with the latest revision of these algorithms, "UAH," from 1999, authored by John Christy (University of Alabama at Huntsville) and Roy Spencer (NASA). The authors of the NAS report (which included both John Christy and Roy Spencer) noted that substantial uncertainties existed with the satellite datasets,[4] some of which arose from short overlaps in satellite intercalibration or sensor issues, as well as other "spacecraft biases and instabilities." Some of these weaknesses had been highlighted previously in work by data analysts Frank Wentz and Matthias Schabel (2000), and Wentz was on the National Academy of Sciences (NAS) panel. The panel also remarked: "Because there is, in effect, only one satellite-based temperature record for which most of the processing has been performed by a single group [UAH], efforts to independently verify the MSU temperature measurements have, of necessity, focused on comparisons with radiosonde data" (2000, p. 16).

It seemed like a good idea for data analysts like Christy and Spencer to try to correlate satellite radiance-based temperature measurements with actual temperature measurements taken from radiosondes, which provided measurements at specific altitudes. In fact, it was standard operating procedure to compare satellite measurements to radiosonde data at NASA, where Roy Spencer was based.[5] This was despite the fact that radiosondes had been found to be unusable to produce long-term

temperature trends, which raised multiple problems for the scientists. These problems, discussed below, would prove to be a huge issue for the climate scientists, as the basic purposes of the radiosondes are not thought to be truly compatible with their functioning as instruments for studying long-term climate trends.

Radiosondes were designed and intended for regional, meteorological purposes, rather than global, climatological purposes. The radiosondes are designed to support local weather forecasting, and, as Thorne et al. (2011) note, frequent changes and improvements in instrumentation have damaged the utility of observations for long-term climate study, by introducing arbitrary biases that vary over time. An additional problem is that it is difficult to extract a global mean temperature from radiosonde data, partly due to the irregular spacing of the radiosonde stations and large gaps over the oceans. Ben Santer, Tom Wigley, and others argued in 1999 that radiosonde data should not be taken as unproblematic independent confirmation of the satellite dataset, UAH, and emphasized the fact that the radiosonde data were very incomplete in their spatial and temporal coverage. They noted that two versions of the same raw radiosonde data, HadRT1.1 and HadRT1.2 have "markedly different lower tropospheric temperature trends over 1979–1996…primarily due to large differences in their spatial coverage" (Santer et al. 1999, p. 6331). They also found that different assumptions about the spatial representativeness of the same raw radiosonde data could even yield datasets with trends of opposite sign. In their words, "This provides a strong warning against overinterpreting apparent trend agreements between data sets" (1999, p. 6328).

Dian Gaffen (who was on the NAS panel) and colleagues, as well as other groups, had discussed these difficulties at length (Gaffen et al. 2000). The NAS panel authors emphasized that, even without any real temperature variation, the global mean temperature calculated from the radiosondes could change over years or decades, simply because of several stations going in or out of operation. Surface-based stations are subject to the same problem, but this is not as big an issue, because of the dense network of surface stations. Radiosonde stations are much rarer, thus the issue.

In the end, the levels of uncertainties of both the surface and tropospheric temperature trends were deemed almost as large as the disparity between them (Wallace et al. 2000, p. 22). In particular, the lack of validation of the satellite data as well as the algorithms used to process those data and the biases and coverage problems of the radiosonde datasets were all seen as contributing to the uncertainty of the temperature records. The panelists also wrote that the temperature records had been "partially, but not fully" reconciled with the climate model simulations. The report emphasized the uncertainties of the models, and the notion that, as the models included more realistic treatments of physical processes, including clouds, they might predict a cooler troposphere. This notion was reinforced by the results of recent experiments involving varying the climate models' initial conditions, while keeping the climate forcings (causal factors or forces) the same, in order to track a simulation of climate variability over a 20-year period. In addition, new climate forcings were added to the models, including volcanic aerosols, pollution, and ozone depletion. These computational experiments produced a wide variety of simulated results, indicating that an arbitrarily short period of record, such as the satellite record's period since 1979, was a risky and unwarranted foundation on which to base an understanding of how the climate changes. These model results were discussed by various modelers, but especially by Ben Santer, who leads the Lawrence Livermore National Laboratory's project on model intercomparison (see Santer et al. 1999, in which the authors discuss the satellite record and models; also continued in Santer et al. 2003).

The NAS panel wrote that climate models at that time were "not sufficiently reliable to provide a definitive assessment of whether the trends at the surface and troposphere are physically consistent" (2000, p. 18). Thus, the panel concluded that measurement uncertainties, modeling uncertainties, and sampling uncertainties were all possible causes of the disagreement between climate models and observations. The panelists recommended, among other things, public dissemination of the raw satellite data, so that other scientists could make their own judgments and decisions about handling those data and could develop them into alternative satellite datasets. And in fact, this approach was crucial to the eventual solution of the problems approached in this inquiry.

6.2 Philosophical Backdrop

Roy Spencer and John Christy pioneered the use of satellites to create temperature trend profiles of the atmosphere (Spencer and Christy 1990). As their UAH dataset was developed and corrected, radiosondes played an important role by providing another temperature trend dataset against which the satellites could be compared and tested (e.g., Christy et al. 2003). Christy and Spencer repeatedly, over the 20-year period of the debates over tropical temperature trends, appealed to radiosonde data as "independent" data against which their UAH dataset compares favorably. In the course of doing so, Christy, Spencer, and some skeptics of the greenhouse effect treated radiosonde data as if they straightforwardly and unproblematically represent the real state of the tropical troposphere. This is an example of what I call "direct empiricism"; the radiosonde data are treated as windows on the world, as reflections of reality, without any art, theory, or construction interfering with that reflection. This claim of a direct connection to reality is very important to their views.

These "direct empiricists" clashed with other climate scientists—both other data analysts and modelers—who took the radiosonde data to be more constructed than transparent.[6] All of the datasets, both satellite and radiosonde, were taken by these "complex empiricists" as theory-laden or heavily weighted with assumptions. Thus, they held that understanding the climate system and the temperature trends required a combination of tools, including models, theory, the taking of measurements, and manipulations of raw data. As I will show, the philosophical clash between "direct" and "complex" empirical approaches is one basis of this long disagreement over the status of climate models and the greenhouse effect.

The name "direct empiricism" is meant to capture an everyday, straightforward and relatively pure notion of how to think about data, measurement, and models. Most central for this discussion is an apparently sensible approach to extracting data from nature, wherein measurements are taken using instruments, and the resulting values are taken at face value, more or less, to represent the naked or unmediated truth about that particular aspect of the world. As pure or naked measurements, these values can then be compared to other values taken using different instruments, in order to calibrate them or compare to

them, as was attempted with the radiosondes and satellites. These values are also compared to the outputs of the models; if there is disagreement, the fault is usually attributed to the models alone, and not to any aspect of the datasets, a pattern of reasoning started in Spencer and Christy's very first paper on the topic (1990). Note that this follows precisely the pattern recommended by the H-D method of testing and confirmation.

This simple notion of data as a naked reflection of reality is contrasted by philosopher Bas van Fraassen with a notion of data as "representation," which is produced at the end of a scientific process that can involve theories or models and the decision-making of the scientists (van Fraassen 2008). On this view of scientific practice, data are never naked, and measurement does not occur without the imposition of framing or generating theories or models. Any measurement of, for example, temperature data from a radiosonde temperature sensor involves a number of adjustments invoked by the data analyst, such as corrections for solar heating, as well as for time of day, and so on (Gaffen et al. 2000).[7]

Van Fraassen offers a meteorological example of the development of a data model or dataset. The weather simulation model produces a "data model constructed" from an analysis of the raw data (van Fraassen 2008, p. 166). He points to the graph of the daily temperature in a region produced after much data processing from the different stations, processed through a statistical analysis, as the scientifically interesting or significant phenomenon. "What is important is that …the outcome must be regarded this way: *this is what the object looks like in this measurement set-up*" (van Fraassen 2008, p. 167). Van Fraassen is emphasizing that the observations are necessarily relativized to the measurement setup, whether that includes decisions about locations, models guiding the interpolation of values, or other decisions in measuring. Here, van Fraassen makes clear the depth of interdependence of data, theory, and model involved in the kind of complex empiricism he endorses, and contrasts it with the kind of direct empiricism I outlined above: "There is a long journey from the initial encounter with nature to the achievement of an even temporarily stable representation" (van Fraassen 2008, p. 91).[8]

Van Fraassen's picture of datasets and their interdependency with models and theories can be complemented by the picture of measurement

offered by philosopher Ronald Giere in his book *Scientific Perspectivism* (2006). Taking the Hubble telescope as the measurement instrument, Giere emphasizes the multiple assumptions about the processing of the raw data required to produce the final images, say, of deep space. "Each step in this process…in some way modifies the initial signal and contributes to the construction of the image…" (2006, p. 44). (This process involves multiple transmissions and retransmissions to satellites and bases, from the initial instruments on board the satellite telescope.)

In focusing on models and data, Giere emphasizes a fact recognized for many years in a model-based philosophical approach to science (Suppe 1962). He discusses the fit of a model to a real system, and how this is determined, emphasizing that that fit is never a direct comparison of model to reality, but rather a fit between a data model and a model derived from theories. The data model (from the observation side) and the model with which it is compared (from the theory side—in climate models, this is simply aspects of the simulation itself) are gradually built up toward one another, eventually converging toward structures that can be directly compared or matched. Thus, a great deal depends on how the data model (dataset) is derived from the raw data: "Of course there may be several different legitimate ways of analyzing the data to obtain a model of the data" (2006, pp. 68–69).[9]

Significantly, Giere notes, datasets are not *always* given first priority over theoretical models:

> The initial presumption is that the observational perspective has priority. The models of data generated within the observational perspective are to be used to decide on the fit of the model generated by theoretical principles; not the other way around. But this is only a strong methodological presumption. *The theoretical model might in some cases be used to question the reliability of the observational instrumentation.* (2006, p. 89; emphasis added)

Our tropical troposphere case is just such a case.

Historian of science and computer scientist Paul Edwards' book, *A Vast Machine* (2010), is primarily a history of climate science's measurement and modeling of Earth's climate. Edwards emphasizes throughout

the book that climate data, in virtue of various difficulties of both a practical and theoretical nature, are inevitably intertwined with climate models, a point that resonates nicely with our approach in this chapter. Edwards also writes on the controversy over the satellite data examined in this chapter (2010, pp. 413–418). He seems to be very sympathetic to a complex empiricist point of view when he concludes, "Neither models nor data alone can support a living understanding of physical phenomena" (2010, p. 418).

The complex empiricist approach advanced thus far is in need of more developed views on the testing and evaluation of the scientific theories and models it represents, and I propose basing such views partly on my analysis, given elsewhere, of the various forms of evidence supporting climate models (Lloyd 2009, 2010, 2015; Lloyd and Mearns 2011). This updated view of model evaluation focuses on independent avenues of theoretical and observational support for various aspects of the simulation models, as well as the accumulation of a variety of evidence for them. Seeing these features of data as bases of evidential support for the climate models directly conflicts, in some cases, with a H-D view of the evidence, as will be discussed below (see Sect. 6.4). Thus, parties in the debate committed to a H-D analysis can disagree strenuously with those taking a more modern approach to model evaluation compatible with complex empiricism. These standards of evidence have not, however, been created specifically with reference to climate models, as they have long been recognized in the biological sciences (Lloyd 1987, 1994; Rykiel 1996).

6.3 Christy and Spencer, the Skeptics, and Direct Empiricism at Work

When Spencer and Christy first published their analysis of the satellite temperature trends in 1990, they stated explicitly that the satellite data are needed for the evaluation of climate models, and also noted that they found no change in temperature trend for the 10-year period they examined, from 1979 to 1988, contrary to the predictions of the climate models.

In a 1992 paper, "Precision and Radiosonde Validation of Satellite Gridpoint Temperature Anomalies," Spencer and Christy used radiosonde data to "validate" the satellite data regarding the tropospheric readings both over the tropics and the rest of the globe. The radiosonde data were taken at face value, as representative measures of the true temperature of the atmosphere. Comparisons with models were not mentioned in this paper, and there is no indication that they intended to use this dataset in such a way, in contrast to their first paper on the satellite dataset. They claimed, "When the satellite measurements were compared to radiosonde measurements of 10 years of monthly anomalies, good agreement was found" (Spencer and Christy 1992, p. 858). Note that the reckoning of temperature from satellite measurements involves a number of variables and complicated adjustments involving the instrument, the MSU, which, to reiterate, does not directly measure temperature at all, but rather radiance, the mass-weighted averages of microwave emissions of oxygen molecules at different altitudes (see Karl et al. 2006). Thus, when they compared the calculations of the temperature from the satellites with the temperatures measured from the radiosondes, they were actually comparing two distinct physical variables measured by the MSUs and by thermometers, respectively.

In any case, Christy and Spencer repeatedly appealed to the radiosonde data to reinforce or "validate" their satellite measurements of the coolness of the tropical tropospheric trends. "Ours is the only dataset that has been compared to non-satellite data," touts Christy in a document prepared for the public on their University of Alabama website in 2003. "This gives us confidence in its results. Several different radiosonde-based products have been compared to the satellite data and the results of those studies have been published" (Christy and Spencer 2003, p. 5; see Christy et al. 1998, 2003a, for refereed versions of the same argument).

The direct empiricist reading of Christy and Spencer may appear to be too simplistic or unsympathetic. After all, isn't it reasonable to appeal to comparisons between the satellite and radiosonde datasets, when it appears that the radiosondes, while also potentially uncertain and heavy with assumptions, are still the best available resource? The problem is that the key question in this debate concerns whether the radiosonde datasets are indeed appropriate independent tests for the satellite datasets; the

complex empiricists rejected the direct empiricists' claim that the radiosondes provided reliable-enough datasets for any such comparison. As we reviewed in Sect. 6.1, meteorologists frequently change their instrumentation, which damages the utility of radiosonde measurements for long-term studies, because of a lack of calibration of those instruments. As we saw, very large errors can appear in long-term trends of radiosonde variables. For these and other reasons, the modelers and other data analysts (see Thorne et al. 2011) did not think the radiosonde datasets should be used as a basis for satellite comparisons (Wallace et al. 2000). They looked, instead, to other independent datasets such as sea surface temperature (SST), water vapor measures, and tropopause height. (See Sect. 6.1 for further discussion.) Thus, to treat the use of the radiosonde datasets as innocent is to beg the question against the modelers and the data analysts in their camp.

One example of this conflict in action can be seen in a debate in 1997 between Jim Hurrell and Kevin Trenberth, on the complex empiricist side, and Christy, Spencer, and William Braswell, on the direct empiricist side. Hurrell and Trenberth were challenging the validity of the quite cool tropical tropospheric temperatures that Christy et al. were producing in their UAH satellite dataset, that were in conflict with model predictions. Hurrell and Trenberth claimed that both the coldness and the downward trend in the satellite temperatures were spurious, and resulted from difficulties in the merging of satellite records, when they had to transition from one satellite to another, and from excessive noise from the radiation reflected from the surface (Hurrell and Trenberth 1997).

Using actual SST measurements and climate models, Hurrell and Trenberth reconstructed the tropical tropospheric temperature trend data, and found they were in agreement with the model expectations, while conflicting with the Christy et al. satellite datasets. In addition to denying the validity of the models' representation of the relationship between surface and tropospheric temperatures, the core of Christy and coauthors' defense of their dataset centered on its conformity with radiosondes: "We believe that lower-tropospheric temperatures measured directly by satellites have excellent long-term accuracy, as seen by comparisons with independent atmospheric measurements from weather balloons" (Christy et al. 1997, p. 342). Throughout the two decades of

debates, this has continued to be the bedrock of Christy et al.'s approach to their evidence.

As we see, Christy and Spencer emphasize repeatedly that radiosondes provide "independent" data against which to compare their interpretation of the satellite data. But Christy and Spencer make a particular choice of satellite intercalibration in building their UAH dataset based on how well it matches with radiosondes. They compare the results of each alternative path from satellites NOAA-6 to NOAA-9 to the results from radiosondes. It is clear that they are using the radiosondes as decisive: "Additionally, these comparisons support the choice of path C1 as the best route for satellite intercalibration, as certain versions of paths A and B would have produced much larger differences between the satellites and the radiosondes" (Christy et al. 1998, p. 2033). Because the radiosonde data are used in building the UAH dataset, it is somewhat circular to then claim support from the radiosonde data. Thus, their claim that their satellite dataset is supported by "independent observations," that is, the radiosonde data, is not accurate (Christy and Spencer 2003a, p. 1046). Critically, some of the radiosonde datasets that UAH is compared to are the exact same ones used in adjusting the data (Santer, Wigley, et al. 2003, p. 1048).

In any case, after more than a decade of improvements and modifications of the UAH dataset, some spurred by outside critiques, two new groups, Remote Sensing Systems (RSS) and the University of Maryland (UMd or VG2), produced their own independent analyses of the original NOAA satellite raw data, making different decisions than Christy and Spencer about how to process those raw data (Mears et al. 2003; Vinnikov and Grody 2003).

Both RSS and UMd yielded temperature trends (see Fig. 6.1) that were significantly warmer than the Christy and Spencer dataset, as we can see from the slopes of the trend lines in this figure. These warmer temperature trends were finally compatible with the models, so there was no longer a conflict between the satellite temperatures and those predicted by the models. Note again that all of these datasets are based on the same NOAA raw satellite data.

In confronting this situation, Christy and Spencer emphasized that the most thorough, and most challenging, of these new datasets, the RSS,

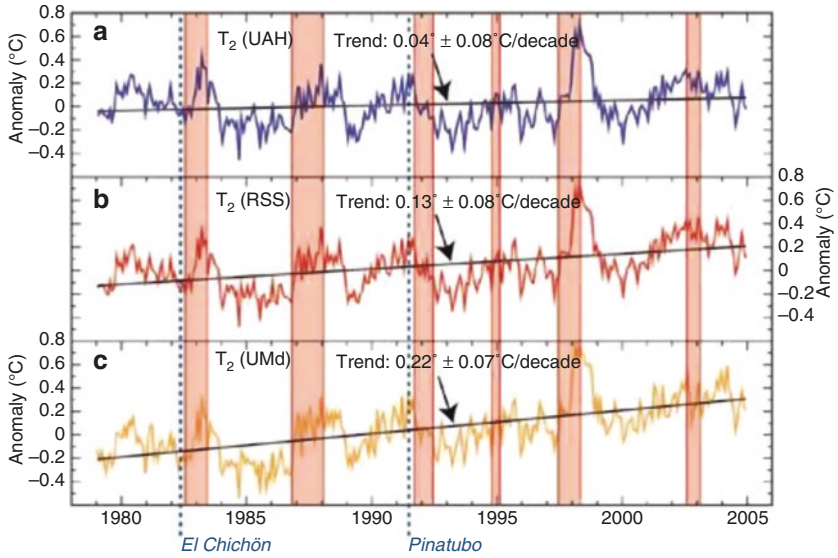

Fig. 6.1 The NOAA raw data as interpreted by three teams of analysts—UAH, RSS, and UMd—and their resulting trend lines. Note the difference in slopes of the trend lines (Karl et al. 2006)

was not compatible with the radiosonde data, and should therefore be considered inferior to the UAH interpretation of the raw NOAA satellite data. The fact that RSS was incompatible with the radiosonde data was foundational to Christy and Spencer, given how they treated these radiosonde data as representing "reality." According to them, a violation of the radiosonde data was not just a mismatch with an observational dataset, but rather, it was a mismatch with reality itself, a reason for definitive rejection of the RSS dataset. For example, here we can see that Christy thinks of the radiosondes as directly representing what he called the "real world," when he discusses the RSS datasets' performance relative to the radiosonde numbers:

> All this means is that [the RSS group's] three [datasets] don't line up properly in terms of the way *the real world* operates, but you don't know which one of those is off. It is pretty difficult to say, but we can say that those three [datasets] are not consistent in *the way balloons are* in describing the global

> atmosphere. (emphasis added; John Christy speaking at a round table at the Marshall Institute, April 17, 2006, 18, comparing RSS numbers with radiosonde numbers) (Christy and Spencer 2006)

In this passage, it is clear that Christy thinks of the radiosondes as directly representing what he called the "real world," when he discusses the RSS datasets' performance relative to the radiosonde, or balloon, numbers.

And look at Christy's recent congressional testimony:

> We, and others, have tested this specific signature [of tropospheric warming from the greenhouse effect], i.e. this hypothesis, against several *observational datasets* and conclude that this pervasive result from climate models has not been detected in the *real atmosphere*. (John Christy, Written testimony, Subcommittee on Energy and Power, Committee on Energy and Commerce, March 8, 2011; emphasis added)

Here we find Christy simply assuming that observational datasets—and he is referring to his UAH satellite dataset as well as some radiosonde datasets—do accurately represent the "real atmosphere," a perfect example of direct empiricism. Nowhere does he acknowledge the very severe problems with deriving trends from balloons, and, significantly, there is no acknowledgment in his testimony of the existence of other satellite datasets, such as RSS and UMd, which are compatible with the models, and which *did* "detect" the greenhouse signature in the atmosphere.[10]

Nevertheless, this perspective on both the satellite and radiosonde data was very influential in debunking the use of models for projecting future climate and for explaining current climate. Christy and Spencer were eager to drive home the consequences of their UAH dataset for the health of climate models: "So what can we say about UAH satellite data? It is the only satellite data that has been subjected to rigorous intercomparisons with independent data and found to be consistent, and the rates of atmospheric warming, both global and tropical, suggest less warming than the majority of model simulations" (Christy 2006, p. 23).

Christy, Spencer, et al.'s early and continuing critique of climate models on the basis of their mismatch with their satellite dataset was picked

up eagerly by climate deniers, both those who deny global warming in general and those who deny that carbon dioxide increases as the primary cause of recent increased global mean temperatures. The deniers were especially keen to note that the radiosonde data were consistent with the satellite data but inconsistent with the models. In skeptic S. Fred Singer's opinion piece in the newsletter of the American Geophysical Union, *EOS*, he emphasized the mismatch of model predictions with both satellite (UAH) and "independent" radiosonde data, citing Christy and coworkers. Singer also uses the lack of warming in the temperature trends in satellite and radiosonde datasets from 1979 to 1998 to argue against "an appreciable human contribution" to global temperature trends (Singer 1999, p. 187). Even in 2005, after the alternate datasets that show warming compatible with the models had become available, Singer argued, with coauthor and skeptic Douglass and Singer (2005): "The anthropogenic Greenhouse effect has been greatly exaggerated. The observational evidence does not support the climate models. But without such validation, there is little reason to trust model predictions of future global warming" (2005, p. 2; for similar arguments, see Baliunas 2002; Carter 2007; De Freitas 2002, p. 320; Green et al. 2004).

Climate scientist Chris De Freitas, emphasizing the damage to the predictions and power of the climate models represented by the conflicting satellite data interpreted by UAH, wrote that "the importance of the satellite data cannot be overestimated" (2002, p. 306). Moreover, according to him, "the satellite data is direct evidence against the IPCC global warming hypothesis" (2002, p. 306).[11] Those on all sides of the debate acknowledged that the mismatch between the predictions of the models and trends in the troposphere "has raised questions about the ability of current global climate models (GCMs) to predict climate changes, the reliability of the observational data used to derive temperature trends, and the reality of human-induced climate change" (Fu et al. 2004, p. 55). The attempt to undermine the status of models via the satellite data is summed up in a contribution from the World Climate Report: "The question remains, and since they didn't ask it, we will: Which do you believe, models or reality? We'll take reality every time" (May 19, 2003).

But however plausible this sort of approach to model evaluation taken by Singer, Christy, Spencer, and other skeptics at first appears, it is

problematic. They seem to view the models' performance in reproducing current tropical tropospheric temperature trends as some kind of crucial experiment or Popperian severe test. When the models fail to fit the available satellite and radiosonde datasets as processed by Christy et al., they are taken to be complete failures as models. In other words, they are then declared to be incompetent to do things that climate models are usually utilized for, including representing and explaining temperature trends, and making projections about future global climate states.

This outdated approach to model evaluation and confirmation is inappropriate for global climate models, and does not accurately reflect their accomplishments or empirical strengths (See Edwards 1999, 2010; Petersen 2006; Parker 2008; Randall et al. 2007; Gleckler et al. 2008; Lloyd 2009, 2010, 2015; Winsberg 2010). These models, in confronting the challenge of the satellite and radiosonde tropical data, have already been shown to be supported by empirical data and adequate for modeling any number of aspects of the environment, including global mean or large-scale distributions of precipitation, radiation, wind, oceanic temperatures, and currents. In addition, the models can simulate patterns of variability (where the model is compared to changes in a given variable over the seasons or months), such as the advance and retreat of major monsoon systems, seasonal shifts of temperatures, storm tracks, and rain belts. The models can also reproduce features of past climate and climate changes, such as the Mid-Holocene warming of 6000 years ago and the Last Glacial Maximum of 21,000 years ago, both of which have specific spatial patterns across the globe. This success in modeling the climate system from previous millennia is taken to show that the forces represented in models can handle values outside the ranges encountered recently, an important test of the applicability of a model to future centuries (Randall et al. 2007, p. 600).

Global climate models are properly judged on the weight of evidence in their favor or disfavor, just as with any scientific theory or model, not on the basis of a single test or performance in a single area, especially if that test depends on data that are contested and assumption-laden, as these satellite and radiosonde datasets are. The variety of evidence is a crucial source of support for climate models, and success among a variety of variables and independent tests of assumptions indicates that a model

is much better supported, and there are a number of such well-supported climate models (Lloyd 2009, 2010, 2015; Lloyd and Mearns 2011; Randall et al. 2007).

Against the background of the successes of global climate models outlined above, the failure of the models in the tropical troposphere was an anomaly, one that certainly needed an explanation, but perhaps not enough of one to require giving up using the models to make future projections or to give explanations of most current climate processes. Under this perspective, the modelers found little reason to follow the skeptics' conclusion to discard the models or to restrict their use. Rather, it seems that the modelers distrusted the satellite and radiosonde data, given the other successes of the models they had in hand and given the uncertainties surrounding the data themselves. After all, the modelers were being handed the multiply-corrected, assumption-laden radiance measurements from satellites, not actual temperature measurements at all. As Giere noted, sometimes the model might be used to "question the reliability of the observational instrumentation" (2006, p. 89).

6.4 Complex Empiricist Treatment of Radiosondes, Theory, and Model Evaluation

Santer, Wigley, et al. have shown that Christy and Spencer do not have a good grasp of the actual content and capacities of climate models, which undermines their claims concerning model evaluation; we should not follow their lead (Santer, Wigley, et al. 2003; e.g., Christy et al. 1997). Instead, these giant climate simulation models call for a novel treatment of evaluation or confirmation by philosophers of science, as Eric Winsberg has argued, echoing ecologist Eric Rykiel's analysis of complex ecological models (Rykiel 1996; Winsberg 2010; Edwards 2010). As Rykiel and Lloyd have shown, attention to complex modeling systems in the biological sciences has for some decades demanded more sophisticated understanding of model support and evaluation in both ecology and evolutionary biology (Lloyd 1987, 1994).

A complex empiricist understanding of climate modeling seems, in virtue of its emphasis on the interactions between models, researchers, and data, to demand a modern approach to model evaluation, one focused not just on the outcomes or predictions of a model, but also on the wide variety of evidence that might or might not be supporting its assumptions and laws. One striking difference between climate models and the physics models that have played leading roles in the H-D analyses, is that the climate models are built using and importing real empirical data. (Edwards calls this "model/data symbiosis" (2010, p. 281ff).) This compromises a philosophical desire for a pristine separation between hypothesis (simulation) and data (processed dataset), perhaps understandably upsetting more traditionally minded philosophers of science accustomed to the H-D approach (e.g., Petersen 2006; Edwards 1999).

These modern complex models involving the computational sciences present a fresh epistemological challenge, by incorporating empirical data into the models at the start. We can either see this as hopelessly bad, as it destroys the pristine separation of theory and data, presenting us with data-contaminated theories and models, or as very good, as it means that the simulations do not stray too far from our measurements built from nature, and are thus partially confirmed from the start (Bad: Edwards 1999; Good: Lloyd 2009; see Knutti this volume). The climate scientists simply see it as necessary in their building of the models and the datasets. They do, however, maintain a separation between the data they use to construct the models and the data used to test the models, often through "data-splitting," the practice of taking the lower or upper portions (or earlier or later portions) of datasets and setting them aside for later use in model verification. This should allay some philosophical worries about the threatened circularity of the procedures of model-building and testing (e.g., Edwards 1999, 2010; Petersen 2006; cf. Rykiel 1996; van Fraassen 2008; Giere 2006).

In addition, the derivation of aspects of model structure from physical laws adds to the modelers' convictions that some of the basic structure, proportions, and relations instantiated in models are fundamentally correct, and are unlikely to be challenged or undermined by datasets that themselves embody potentially arbitrary assumptions. Additionally, the

provision of independent observational evidence for various aspects and assumptions of the models—such as measuring parameter values and relations between variables—increases the credibility of claims made on behalf of models. This support can go beyond or replace the provision of empirical support that might otherwise be provided by matching the predictions with observational datasets. There are, in other words, many more ways to empirically support a model than through predictive success of a single variable such as global mean temperature, and all of these ways play significant roles in supporting climate models (Lloyd 2009, 2010, 2015). All of this is neglected by the direct empiricists, who, following an H-D approach, focus exclusively on predictive accuracy and matching of model predictions of a single variable with datasets with large uncertainties such as the satellite datasets or, even worse, the radiosonde datasets.

The complex empiricist vision must thus be complemented by a thorough understanding of how models are empirically supported through a variety of evidence that can include empirical evidence for specific parameter ranges, the derivation of specific aspects of the model from foundational physical laws, or the importation of empirical data in a micro-model embodying a parameterization of, for example, cloud behavior and its effects on model variables, as argued elsewhere (Lloyd 2010). Independent support for the assumptions and decisions made in processing raw data into datasets is an essential aspect of evidence, and demands for such evidence are often made by complex empiricists. Thus, understanding why the H-D account of confirmation taken by the direct empiricists, with its overemphasis on prediction, is inappropriate when examining climate science models, is essential.

Compare the direct empiricist approach outlined in Sect. 6.3 with the way that the modelers and other data experts approach the radiosonde and satellite data: "We have used basic physical principles as represented in current climate models, for interpreting and evaluating observational data" (Santer et al. 2005, p. 1555). Note that this is the opposite of direct empiricism; the data are seen in terms of the theory and its assumptions, just as Giere described. The evaluation of datasets is one where raw data are evaluated as plausible or acceptable based on their compatibility with certain theoretical or dynamic processes. On this view, Santer and the

other modelers and data experts co-authoring that paper are not taking even the "raw" data as transparently representing the real world, but rather as constructed already, seen best through a theoretical lens through which it was also created or mined out of the world. On this approach, all data are interpreted.

Santer et al. claim that their work involves using a combination of observations, theory, and models. They see the data as enmeshed with theory and models and their assumptions, and as constructed, rather than found and reflecting reality in a straightforward way (Santer et al. 2005, p. 1555. The cooperation between and complementarity of data analysts and theoreticians are embodied in this paper, as is seen upon examination of the list of authors, which includes top leaders among both specializations). As noted by van Fraassen, a complex empiricist approach "does not presuppose an impossible god-like view in which nature and theory and measurement practice are all accessed independently of each other and compared to see how they are related 'in reality'" (2008, p. 139). There is, in other words, no pristine separation of model and data.

Santer, other modelers, and Mears, Wentz, and Schabel, the data analysts, published a joint paper in 2003 comparing the new satellite dataset (RSS) to the models and to the UAH dataset (Santer et al. 2003; Mears et al. 2003). Because the UAH dataset was claimed by Christy and Spencer to be independently confirmed by the radiosonde data, the opponents of the direct empiricists needed to address the problem of the radiosondes head-on (Mears et al. 2003). Far from seeing the radiosonde datasets as a transparent reflection of the true temperature of the atmosphere, as the direct empiricists represent them, Mears and coauthors rejected the radiosondes as a decent source of information about the temperature of the atmosphere. Citing the difficulties raised in Gaffen et al. (2000) and Lanzante et al. (2003), they claimed that radiosondes are "subject to a host of complications, including changing instrumentation types, configurations, and observation practices… making long-term climatological studies difficult" (Mears et al. 2003, p. 3650). Like Gaffen, they noted that trends for individual radiosonde observation stations can vary as much as by 0.1 degree Celsius per decade, which is as large as the overall temperature trend in the tropics

under debate (Santer et al. 1999). "Based on these results we think it is *inappropriate* to use radiosonde comparisons as the single method for validating satellite-derived temperature trends, and studies, such as ours, that are primarily based on internal consistency should be considered on equal footing" (Mears et al. 2003, p. 3664; emphasis added). The RSS authors thus challenged the radiosonde temperature datasets, and did not support their use in interpreting climate trends. They suggest, rather, that their dataset be tested for compatibility with a variety of evidence from the independent observational datasets of SST, water vapor measures, and tropopause height.

In the 2003 paper coauthored by both leading modelers and the RSS data analysts, the authors argue that the differences between models and the Christy and Spencer UAH dataset may be artifactual, based on the fact that the RSS dataset fits model expectations but differs significantly from the UAH dataset (Santer, Sausen, et al. 2003). In this paper, the authors show that these two independent datasets constructed out of the same raw satellite-based emission data differ significantly in the tropospheric temperature trends. In the UAH, the tropospheric temperature remains constant, while in the RSS, there is a trend of increasing temperature. This is significant, because only the RSS is compatible with the models' prediction of a warming tropospheric fingerprint of combined anthropogenic and natural effects. However, the authors note that "we cannot say definitively whether RSS or UAH provides a better estimate of the 'true' tropospheric temperature changes" (Santer, Sausen, et al. 2003, p. 1283).

Like Mears et al. (2003), Santer, Sausen, et al. (2003) proposed that this dilemma be resolved by looking at "complementary data sets" related to tropospheric temperature, for example, other observational data such as change in tropopause height, water vapor, and SST (e.g., Santer, Sausen, et al. (2003), a paper on the changing height of the tropopause). They also refer to Wentz and Schabel, who claimed that satellite measurements of SST and water vapor can independently validate an MSU temperature measurement, noting that the water vapor measurement is highly sensitive to temperature changes (2000). This set of approaches, appealed to by Santer, Sausen, et al. (2003), had been endorsed before in the 2000 report from the National Academy of Sciences, in which the panel

pointed to the independent evidence of tropospheric warming arising from the melting of tropical glaciers in the Andes and high mountains of Africa, and the water vapor loading increase in the tropical troposphere (Wallace et al. 2000).[12] Note that the types of evidence supported by the complex empiricists are not predictions of global mean temperature from the climate models, the actual variable under dispute, but rather, other variable outcomes and parameter values of the models, including SST or water vapor, whose values can be measured accurately, and then used to calculate global mean temperature of the troposphere using the models. This strategy of gaining independent evidence for aspects of the global models is emphasized by my proposed modern approach to model evaluation.

In a reply to the Santer, Wigley, et al. (2003) paper in *Science*, Christy and Spencer claimed that the radiosonde data clinched the superiority of their UAH dataset (Christy and Spencer 2003b). They emphasized that they had already put the UAH through "appropriate, observationally based tests," emphasizing that these are "independent observations, not model output" (2003b, p. 1046). However, Santer et al. quickly responded, reciting the many problems with using the radiosonde datasets discussed above, and concluded: "For these and other reasons, radiosondes are not an *unambiguous 'gold standard'* for the evaluation of satellite data" (Santer et al. 2003, p. 1047; emphasis added). Thus, they viewed the radiosonde datasets as very much dependent on a series of decisions regarding how to handle the data, that is, as constructed through a combination of theory, model, and measurement.

This complex empiricist approach to the radiosonde data is even more visible in a 2005 paper coauthored by Santer, Wigley, Mears, Thorne, Wentz, and a host of other modelers and data analysts. This paper compares model results from multiple model runs to both radiosonde and satellite datasets, finding that only the RSS dataset fits the models on the decadal time scale. The evaluation of datasets referred to here is one where raw measurements are evaluated as plausible or acceptable based on their compatibility with certain theoretical or dynamic processes. This is the opposite of direct empiricism; the resulting datasets are processed or constructed, laden with the theoretical understanding of climate processes and causes. The scientists in question are not taking the raw measure-

ments as directly reflecting the real world, but rather as passing on some number or measurement already processed through a set of assumptions about the world, which need further scrutiny before they are to be accepted.[13] As Santer et al. put it, their work

> illustrates that progress toward an improved understanding of the climate system can best be achieved by combined use of observations, theory and models. The availability of a large range of model and observational surface and atmospheric temperature datasets has been of great benefit to this research, and highlights the dangers inherent in drawing inferences on the agreement between models and observations without adequately accounting for uncertainties in both. (Santer et al. 2005, p. 1555)[14]

This complex empiricist approach to the constructed nature of the datasets and the processes involved in theory and model evaluation is echoed by the authors of a third satellite dataset. Each dataset is produced through a succession of decisions involving various physical processes themselves, all culminating in a set of values usable by the scientific community. Vinnikov et al., the analysts behind the VG2 (UMd) satellite dataset based on the NOAA raw satellite data, claim that "the fact that the model and observations agree so closely gives us more confidence in both the observational record and in the model projections of future climate change" (Vinnikov et al. 2006, p. 12). The models, physical theory, and observational datasets are thus seen as mutually supporting. Here we have data analysts claiming that their observations are partially evaluated in terms of their compatibility with the models, instead of just simply testing the models with their data, as a traditional H-D approach to theory evaluation would advise.

Further Complex Empiricist Treatment of Satellite Data

As we have seen, the complex empiricists raised a number of worries about the radiosonde temperature readings. Despite its unsuitability and lack of design for climate research, complex empiricists, like direct empiricists, are responsible for utilizing the radiosonde data in everyday reasoning about satellite data, and in endorsing certain conclusions

based partially on those data involving the merits of the particulars of the satellite datasets. This is illustrated by a paper by Qiang Fu and his colleagues that challenged both the RSS and UAH datasets' interpretation of tropospheric temperature trends, arguing that both datasets suffered from significant contamination from stratospheric cooling, and were thus too cool (Fu et al. 2004). Fu and his colleagues used both the stratospheric satellite raw data and careful reconstructions using the latest radiosonde data to construct a new interpretation of the middle tropospheric temperature profile. Under Fu et al.'s new analysis, the global middle tropospheric trends for the UAH and RSS datasets are each 0.08 degree Celsius per decade warmer than the original calculations. Thus corrected, the datasets are well within the simulation model projections of tropical tropospheric warming. Fu and Johanson also later analyzed the inconsistency of their radiosonde-reconstructed tropospheric and stratospheric temperature trends with the UAH datasets (2005). This made it much more difficult for Christy et al. to claim that their satellite datasets were uniquely supported by consistency with the radiosonde datasets.[15]

Another large correction involved solar heating of the instrument carried by the radiosondes, such that there was probably a spurious cooling trend in the radiosonde dataset amounting to 0.16 degree Celsius per decade in the tropics in the mid troposphere, spanning the satellite period from 1979 to 1997 (Sherwood et al. 2005). This means that if the satellites were found to agree with the radiosondes during this period, they were also found to be agreeing to a spurious trend of sizable magnitude (see IPCC 2007, p. 267). Allen and Sherwood later calculated warmer temperature trends in the tropical troposphere compatible with the models using the thermal wind equation, a method not subject to the difficulties with the instrumental biases of the thermometers (Allen and Sherwood 2008). Mitch Goldberg, Zou Cheng-Zhi and their colleagues are, most recently, clarifying and correcting the temperature profiles and trends, using updated intercomparisons of satellites and their calibration, the latest tools for handling the problems with the radiosonde and satellite data, and a complex empiricist approach to the problems (Zou et al. 2009; Trenberth Pers. comm. August 2010).

6.5 Skeptics and the True Data with No Error Bars

We have explored the many reasons that the complex empiricists reject many radiosonde datasets as being direct reflections of the "real world," and as decisive evidence against the models, as claimed by the direct empiricists (e.g. Douglass and Singer 2005). But the direct empiricist approach is also manifest in a 2008 attack on the climate models, which Christy and colleagues claimed clearly falsified the greenhouse effect. This skirmish arose despite the fact that, to the US Climate Change Science Program and nearly all of the participants (except, apparently, Christy and possibly Spencer), the debate over whether models were consistent with observational data was considered settled by 2006 (Karl et al.), after the establishment of the alternative datasets to UAH. To Christy and his skeptical coauthors—David Douglass, Benjamin Pearson, and S. Fred Singer—it is a direct empirical matter to check the models against the straightforward observational data, which represent the real world (Douglass et al. 2008; see Fig. 6.2). These observational data, according to their 2008 paper, include the radiosonde datasets and satellite datasets.

What is fascinating about their argument with the complex empiricists is that Christy, Singer, and the skeptics present their observational datasets, both radiosonde and satellite, without any error bars at all, but rather as single values, as if the observations are to be taken as simple, clearly correct values representing reality (2008, p. 1697). The point of the paper is to argue that the satellite data—uniquely as UAH interpret them, without any suggestion that there are other interpretations in other datasets—do not fit the models. The temperature datasets used to show this point, and against which the models are compared, are presented as some kind of fixed reality, absent of error, rather than as representations resulting from scientific processes. This objectification of the temperature values, through omitting the error bars, is a startling abandonment of normal scientific practice, all in the context of presenting the temperatures as firm counterevidence for the models. This paper received wide attention and acceptance not only at Fox News but also within the Department of Energy and at NOAA.[16] An additional problem is that the radiosonde datasets presented in this Figure were outdated, and much

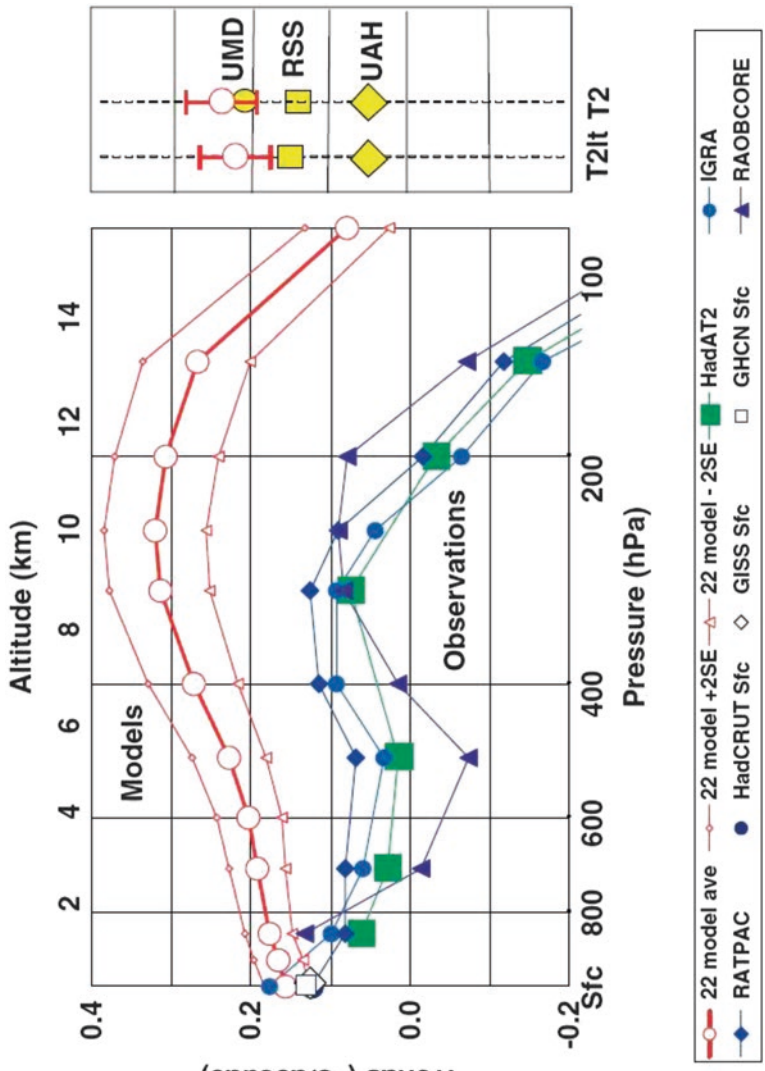

Fig. 6.2 Note that the models are presented within the bounds of two standard errors at the top of the figure, while the four observational radiosonde datasets below are presented as lone points, as are the satellite datasets on the side (Douglass et al. 2008)

cooler than the more recent, corrected radiosonde datasets available then. The RSS and UMd satellite dataset values were also outdated. Here we see the essence of the direct empiricist view of the processes of measurement of temperature: they capture nature's reality transparently and unproblematically. And this is precisely the direct empiricist view of evidence and data that Santer, Wigley, and the various modelers, as well as data analysts Mears, Wentz, Sherwood, Fu, and others, reject.

This Douglass et al. paper was decisively, if not savagely, refuted by Santer, Wigley, Mears, Lanzante, Sherwood, Karl, Gleckler, Thorne, and a host of other data analysts and modelers in a paper published the same year, in the same journal (Santer et al. 2008; see Fig. 6.3; see Chaps. 3 and 4). Figure 6.3 shows that all of the updated and corrected observational datasets are within the confidence intervals of the models. The authors emphasized the "structural uncertainties," of both radiosonde and satellite datasets, which result from the different choices that analysts make when processing the raw data to adjust for inhomogeneities (Santer et al. 2008, p. 1704). The complicated and active role of the data analyst envisioned by these complex empiricists, one where measurement itself is a scientific process, seems quite alien to the direct empiricist approach assumed in the Douglass et al. paper. Thus, we can see why the two parties are unlikely to agree about even the most fundamental question in

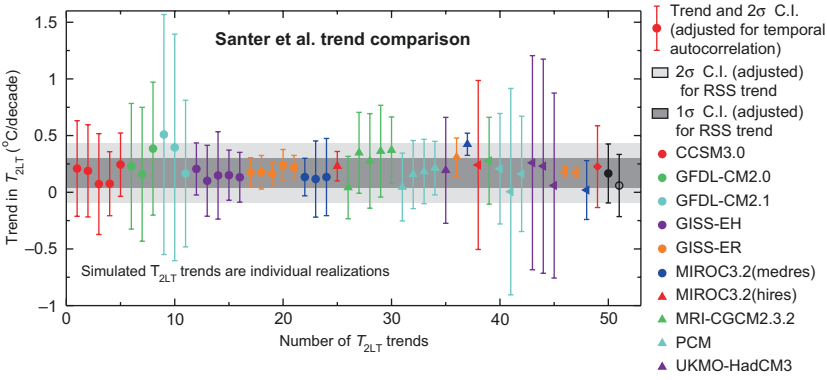

Fig. 6.3 Note that the model realizations are all found within two standard deviations of the RSS trend, thus demonstrating the compatibility of the satellite data and various models (Santer et al. 2008)

this dispute, namely, what the data say. Hence, we witnessed Christy testifying to Congress in 2011, using the results from the Douglass et al. 2008 paper, despite the fact that Santer et al. (2008) had demonstrated that it contained significant and fatal statistical flaws.

In sum, we can see how the Christy, Spencer, and skeptics' direct empiricist approach to much of the observational data has conflicted with the complex empiricist approach of their critics. Christy et al. view the radiosonde data and datasets as clearly representing the "real world," against which their UAH satellite dataset is compared and found to also represent that same "real world" (Christy 2006, p. 18). This direct empiricist approach to the datasets is also manifest in their most recent attack on the climate models, in which the observational data are represented as single values, with no error bars, all in the context of presenting the temperatures as firm counterevidence for the models (Douglass et al. 2008). This paper was presented to Congress on March 8, 2011, by John Christy as proof that the greenhouse effect is not occurring, and is not a danger to the United States.[17]

6.6 Conclusion

I have contrasted the direct empiricist approach to a much more complex one taken by the modelers and many of the data analysts involved in this debate about the satellite data. The analysis here thus documents both the scientific and philosophical utility of the kind of complex empiricist understanding of evidence, models, and theory advanced by Giere, van Fraassen, and others, and further developed here. Under this view, data, theory, models, and scientific practice are deeply intertwined and play complementary roles in producing the scientific values in any dataset. In this case, those scientists taking this approach have refused to follow a simple line of reasoning in which the radiosonde data provide a firm empirical grounding for the satellite datasets, which in turn show that the climate models are false and thereby useless for predicting future climate.

I have also provided a sketch of a complex empiricist approach to theory and model evaluation and testing. Understanding how these climate models are evaluated requires close attention to the independent

theoretical and empirical support for various aspects of the models, such as laws, parameterizations, parameters, and variable values, as well as for the assumptions embodied in the observational datasets. Correlatively, we need less focus on deductive predictions, which are the main focus of the traditional H-D approach to model confirmation and evaluation taken by the direct empiricists. Understanding the modelers' and data analysts' apparently stubborn rejection of the proffered counterevidence requires an enriched awareness of the interdependent relations of theory, model, and data, a complex empiricist view contrary to a direct empiricist account that takes data as a straightforward window onto the world. While the contrast between the approaches I've drawn is naturally somewhat simplistic, this illustration of their differences enriches our understanding of this significant episode in climate science. Because many of the sciences are adopting more and more uses for computational models and simulations, it is imperative that philosophers of science adopt and develop something like a complex empiricist understanding of those models and how they are built, used, and evaluated.

Acknowledgment I am indebted to climate scientists Caspar Ammann, Jim Hurrell, Jeffrey Kiehl, Ricky Rood, Ben Santer, Peter Thorne, Kevin Trenberth, and Tom Wigley for their assistance. I thank philosophers Ron Giere, Peter Gildenhuys, Alex Klein, Steven Lawrie, Helen Longino, Gordon McOuat, Rudy Raff, Paul Teller, Trin Turner, Sean Valles, and Eric Winsberg. Bas van Fraassen and Isabelle Peschard organized "The Experimental Side of Modeling" workshop at San Francisco State University (March 2010), where this paper got its start, and I thank them, as well as the attendees, for their guidance and comments on this project and its issues.

Notes

1. John Christy, Written testimony, US Senate Committee on Commerce, Science, and Transportation, Subcommittee on Space, Science, and Competitiveness, January 8, 2015, convened by Chairman US Senator Ted Cruz.
2. This includes Peter Thorne, John Lanzante, Thomas Peterson, Dian Seidel, and Keith Shine, data analysts involved in the debate, who offer

a new detailed, technical review of the decades of debate, which balances my focus on the satellites with that on the radiosonde datasets, which remain relatively neglected in my study (Thorne et al. 2011). Their study offers independent support for my analysis regarding the philosophical approaches to the satellite and radiosonde datasets.

3. R. Raff, Pers. comm. February 14, 2010.
4. In climate science, what philosophers such as Patrick Suppes call "models of data" are called "datasets (1962)."
5. Ricky Rood, Pers. comm. August 30, 2010.
6. When I talk of "construction" of datasets, this is not meant to imply anything arbitrary or fanciful about the process. Rather, it refers to the necessary, rational, and scientific decision processes that are required in the production of the final measurements that make up the datasets.
7. Paul Edwards gives a detailed account of this process in his book, *A Vast Machine* (2010).
8. Van Fraassen himself does not use the name "complex empiricism" for his view but has agreed to my attribution.
9. For a contemporary review of some of these ways of analyzing data, see Thorne et al. (2011) and Edwards (2010, esp. pp. 256–273).
10. As a coauthor, Christy admitted that the question of which dataset—RSS, UAH, or UMd—is closest to the true tropospheric temperature was unknown (Karl et al. 2006).
11. The Intergovernmental Panel on Climate Change (IPCC) is a United Nations panel, part of which is dedicated to summarizing the state of climate science at a given time. The conclusion of the reports of the IPCC had been that global warming existed, and was significantly affected by human causes (Houghton et al. 2001; IPCC 2007).
12. The NAS panel itself was split on the importance of the radiosonde data. It noted that "[t]hose more inclined to take the MSU [satellite] measurements at face value cite the high degree of consistency with radiosonde measurements (Figs. 2.3, 9.2, and 9.3) [based on data from Christy et al. 2000], whereas those less inclined to do so note the retreat of the tropical glaciers and the increasing burden of water vapor" (Wallace et al. 2000, p. 65).
13. See, for another example, Jeffrey Kiehl et al. (2005), "On using global climate model simulations to assess the accuracy of MSU retrieval methods for tropospheric warming trends."
14. This attitude is also reflected in the IPCC 2007 report when discussing the radiosonde and satellite datasets. There, the authors of the chapter on

climate observations write: "It is difficult to make quantitatively defensible judgments as to which, if any, of the multiple, independently derived estimates is closer to the true climate evolution. This. . . points to the need for future network design that provides the reference sonde-based ground truth" (Solomon et al. IPCC 2007, p. 265).

15. Spencer et al. (2006) subsequently attacked the Fu and Johanson methods, and rejected their conclusions about radiosonde temperature trends. Johanson and Fu (2006) addressed the issues raised in the Spencer et al. (2006) paper.
16. Ben Santer, pers. comm. March 10, 2011.
17. John Christy, Written testimony, Subcommittee on Energy and Power, Committee on Energy and Commerce, March 8, 2011.

References

Allen, Robert J., and Steven C. Sherwood. 2008. Warming Maximum in the Tropical Upper Troposphere Deduced from Thermal Winds. *Nature Geoscience* 1 (6): 399–403. https://doi.org/10.1038/ngeo208.

Baliunas, Sallie. 2002. New Scientific Advances: The Human Impact on Global Climate Change. Testimony before the Senate Committee on Environment and Public Works, March 13.

Carter, R.M. 2007. *The Myth of Dangerous Human-Caused Climate Change*. Australasian Institute of Mining and Metallurgy, New Leader's Conference, Brisbane.

Christy, John R., and Roy W. Spencer. 2003a. *Global Temperature Report: 1978–2003*. Huntsville: Earth System Science Center, University of Alabama in Huntsville.

———. 2003b. Reliability of Satellite Data Sets. *Science* 301 (5636): 1046–1049. https://doi.org/10.1126/science.301.5636.1046.

———. 2006. Satellite Temperature Data. In *Washington Roundtable on Science & Public Policy*, pp. 1–37. George Marshall Institute. http://marshall.wpengine.com/wp-content/uploads/2013/08/Christy-and-Spencer-Satellite-Temperature-Data.pdf

Christy, John R., Roy W. Spencer, and William D. Braswell. 1997. How Accurate Are Satellite 'Thermometers'? *Nature* 389 (6649): 342–342. https://doi.org/10.1038/38640.

Christy, John R., Roy W. Spencer, and Elena S. Lobl. 1998. Analysis of the Merging Procedure for the MSU Daily Temperature Time Series. *Journal of Climate* 11 (8): 2016–2041. https://doi.org/10.1175/1520-0442(1998)011<2016:AOTMPF>2.0.CO;2.

Christy, John R., Roy W. Spencer, and William D. Braswell. 2000. MSU Tropospheric Temperatures: Dataset Construction and Radiosonde Comparisons. *Journal of Atmospheric and Oceanic Technology* 17 (9): 1153–1170. https://doi.org/10.1175/1520-0426(2000)017<1153:MTTDCA>2.0.CO;2.

Christy, John R., Roy W. Spencer, William B. Norris, William D. Braswell, and David E. Parker. 2003. Error Estimates of Version 5.0 of MSU–AMSU Bulk Atmospheric Temperatures. *Journal of Atmospheric and Oceanic Technology* 20 (5): 613–629. https://doi.org/10.1175/1520-0426(2003)20<613:EEOVOM>2.0.CO;2.

De Chadarevian, Soraya, and Nick Hopwood. 2004. *Models: The Third Dimension of Science*. Stanford: Stanford University Press.

de Freitas, C.R. 2002. Are Observed Changes in the Concentration of Carbon Dioxide in the Atmosphere Really Dangerous? *Bulletin of Canadian Petroleum Geology* 50 (2): 297–327.

Douglass, D.H., and S.F. Singer. 2005. *Climate Data Disagree with Climate Models: Policy Dilemma: Should We Believe in Atmosphere or in Models?* AGU Fall Meeting 2005: American Geophysical Union.

Douglass, David H., John R. Christy, Benjamin D. Pearson, and S. Fred Singer. 2008. A Comparison of Tropical Temperature Trends with Model Predictions. *International Journal of Climatology* 28 (13): 1693–1701. https://doi.org/10.1002/joc.1651.

Edwards, Paul N. 1999. Global Climate Science, Uncertainty and Politics: Data-laden Models, Model-Filtered Data. *Science as Culture* 8 (4): 437–472. https://doi.org/10.1080/09505439909526558.

———. 2010. *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. Cambridge, MA: MIT Press.

Fu, Qiang, and Celeste M. Johanson. 2005. Satellite-Derived Vertical Dependence of Tropical Tropospheric Temperature Trends. *Geophysical Research Letters* 32 (10): L10703. https://doi.org/10.1029/2004GL022266.

Fu, Qiang, Celeste M. Johanson, Stephen G. Warren, and Dian J. Seidel. 2004. Contribution of Stratospheric Cooling to Satellite-Inferred Tropospheric Temperature Trends. *Nature* 429 (6987): 55–58. https://doi.org/10.1038/nature02524.

Gaffen, Dian J., Michael A. Sargent, R.E. Habermann, and John R. Lanzante. 2000. Sensitivity of Tropospheric and Stratospheric Temperature Trends to Radiosonde Data Quality. *Journal of Climate* 13 (10): 1776–1796. https://doi.org/10.1175/1520-0442(2000)013<1776:SOTAST>2.0.CO;2.

Giere, Ronald. 2006. *Scientific Perspectivism*. Chicago: University of Chicago Press. http://www.press.uchicago.edu/ucp/books/book/chicago/S/bo4094708.html.

Gleckler, P.J., K.E. Taylor, and C. Doutriaux. 2008. Performance Metrics for Climate Models. *Journal of Geophysical Research: Atmospheres* 113 (D6): D06104. https://doi.org/10.1029/2007JD008972.

Green, Kenneth, Tim Ball, and Steven Schroeder. 2004. The Science Isn't Settled: The Limitations of Global Climate Models. *Public Policy Sources* 80: 1–32.

Houghton, J., Y. Ding, D. Griggs, M. Noguer, P. van der Linden, X. Dai, K. Maskell, and C. Johnson. 2001. *Climate Change 2001: The Scientific Basis*. Cambridge: Cambridge University Press.

Hurrell, J.W., and K.E. Trenberth. 1997. Spurious Trends in Satellite MSU Temperatures from Merging Different Satellite Records. *Nature* 386 (6621): 164.

IPCC. 2007. Climate Change 2007: The Physical Science Basis. In ed. S. Solomon, D. Qin, M. Manning, M. Marquis, K. Averyt, M.M.B. Tignor, H. Ljr Miller, and Chen Zhenlin. http://agris.fao.org/agris-search/search.do?recordID=XF2016025238

Johanson, Celeste M., and Qiang Fu. 2006. Robustness of Tropospheric Temperature Trends from MSU Channels 2 and 4. *Journal of Climate* 19 (17): 4234–4242. https://doi.org/10.1175/JCLI3866.1.

Karl, Thomas, Susan Hassol, Christopher Miller, and Murray. 2006. *Temperature Trends in the Lower Atmosphere: Steps for Understanding and Reconciling Differences*. A Report by the U.S. Climate Change Science Program and the Subcommittee on Global Change Research. Asheville: National Oceanic and Atmospheric Administration, National Climatic Data Center.

Kiehl, Jeffrey T., Julie M. Caron, and James J. Hack. 2005. On Using Global Climate Model Simulations to Assess the Accuracy of MSU Retrieval Methods for Tropospheric Warming Trends. *Journal of Climate* 18 (14): 2533–2539. https://doi.org/10.1175/JCLI3492.1.

Lanzante, John R., Stephen A. Klein, and Dian J. Seidel. 2003. Temporal Homogenization of Monthly Radiosonde Temperature Data. Part II: Trends, Sensitivities, and MSU Comparison. *Journal of Climate* 16 (2): 241–262. https://doi.org/10.1175/1520-0442(2003)016<0241:THOMRT>2.0.CO;2.

Lloyd, Elisabeth A. 1987. Confirmation of Evolutionary and Ecological Models. *Biology and Philosophy* 2: 277–293.

———. 1994. *The Structure and Confirmation of Evolutionary Theory*. Princeton: Princeton University Press.

———. 2009. I—Varieties of Support and Confirmation of Climate Models. *Aristotelian Society Supplementary Volume* 83 (1): 213–232. https://doi.org/10.1111/j.1467-8349.2009.00179.x.

———. 2010. Confirmation and Robustness of Climate Models. *Philosophy of Science* 77 (5): 971–984.

———. 2015. Model Robustness as a Confirmatory Virtue: The Case of Climate Science. *Studies in History and Philosophy of Science Part A* 49: 58–68. https://doi.org/10.1016/j.shpsa.2014.12.002.

Lloyd, Elisabeth Anne, and Linda O. Mearns. 2011. *The Principle of the Variety of Evidence and Its Significance to Climate Science.* AGU Fall Meeting Presentation.

Mears, Carl A., Matthias C. Schabel, and Frank J. Wentz. 2003. A Reanalysis of the MSU Channel 2 Tropospheric Temperature Record. *Journal of Climate* 16 (22): 3650–3664. https://doi.org/10.1175/1520-0442(2003)016<3650:AROTMC>2.0.CO;2.

Morgan, Mary S., and Margaret Morrison. 1999. *Models as Mediators: Perspectives on Natural and Social Science*. Cambridge: Cambridge University Press.

Parker, Wendy S. 2008. Computer Simulation Through an Error-Statistical Lens. *Synthese* 163 (3): 371–384. https://doi.org/10.1007/s11229-007-9296-0.

Petersen, Arthur C. 2006. *Simulating Nature: A Philosophical Study of Computer-Simulation Uncertainties and Their Role in Climate Science and Policy Advice*. Apeldoorn: Het Spinhuis.

Randall, David A., Richard A. Wood, Sandrine Bony, Robert Colman, Thierry Fichefet, John Fyfe, Vladimir Kattsov, et al. 2007. Climate Models and Their Evaluation. In *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, ed. Susan Solomon and Others. New York: Cambridge University Press.

Rykiel, Edward J. 1996. Testing Ecological Models: The Meaning of Validation. *Ecological Modelling* 90 (3): 229–244. https://doi.org/10.1016/0304-3800(95)00152-2.

Santer, B.D., R. Sausen, T.M.L. Wigley, J.S. Boyle, K. AchutaRao, C. Doutriaux, J.E. Hansen, G.A. Meehl, E. Roeckner, R. Ruedy, and G. Schmidt. 2003. Behavior of tropopause height and atmospheric temperature in models, reanalyses, and observations: Decadal changes. *Journal of Geophysical Research:*

Atmospheres 108(D1). http://www.academia.edu/13425635/Behavior_of_tropopause_height_and_atmospheric_temperature_in_models_reanalyses_and_observations_Decadal_changes. Accessed 29 May 2017.

———. n.d. Response to Christy and Spencer 2003. *Science* 301: 1047–1049.

Santer, Benjamin D., J.J. Hnilo, T.M.L. Wigley, J.S. Boyle, C. Doutriaux, M. Fiorino, D.E. Parker, and K.E. Taylor. 1999. Uncertainties in Observationally Based Estimates of Temperature Change in the Free Atmosphere. *Journal of Geophysical Research: Atmospheres* 104 (D6): 6305–6333. https://doi.org/10.1029/1998JD200096.

Santer, Benjamin D., T.M.L. Wigley, G.A. Meehl, M.F. Wehner, C. Mears, M. Schabel, F.J. Wentz, et al. 2003. Influence of Satellite Data Uncertainties on the Detection of Externally Forced Climate Change. *Science* 300 (5623): 1280–1284. https://doi.org/10.1126/science.1082393.

Santer, Benjamin D., T.M.L. Wigley, C. Mears, F.J. Wentz, S.A. Klein, D.J. Seidel, K.E. Taylor, et al. 2005. Amplification of Surface Temperature Trends and Variability in the Tropical Atmosphere. *Science* 309 (5740): 1551–1556. https://doi.org/10.1126/science.1114867.

Santer, Benjamin D., P.W. Thorne, L. Haimberger, K.E. Taylor, T.M.L. Wigley, J.R. Lanzante, S. Solomon, et al. 2008. Consistency of Modelled and Observed Temperature Trends in the Tropical Troposphere. *International Journal of Climatology* 28 (13): 1703–1722. https://doi.org/10.1002/joc.1756.

Sherwood, Steven C., John R. Lanzante, and Cathryn L. Meyer. 2005. Radiosonde Daytime Biases and Late-20th Century Warming. *Science* 309 (5740): 1556–1560.

Singer, S. Fred. 1999. Human Contribution to Climate Change Remains Questionable. *Eos, Transactions American Geophysical Union* 80 (16): 183–187. https://doi.org/10.1029/99EO00132.

Spencer, Roy W., and John R. Christy. 1990. Precise Monitoring of Global Temperature Trends from Satellites. *Science* 247 (4950): 1558–1562.

———. 1992. Precision and Radiosonde Validation of Satellite Gridpoint Temperature Anomalies. Part I: MSU Channel 2. *Journal of Climate* 5 (8): 847–857. https://doi.org/10.1175/1520-0442(1992)005<0847:PARVOS>2.0.CO;2.

Spencer, Roy W., John R. Christy, William D. Braswell, and William B. Norris. 2006. Estimation of Tropospheric Temperature Trends from MSU Channels 2 and 4. *Journal of Atmospheric and Oceanic Technology* 23 (3): 417–423. https://doi.org/10.1175/JTECH1840.1.

Suppes, Patrick. 1962. Models of Data. In *Logic, Methodology, and Philosophy of Science: Proceedings of the 1960 International Congress*, ed. Ernest Nagel, Patrick Suppes, and Alfred Tarski. Stanford: Stanford University Press.

Thorne, Peter W., John R. Lanzante, Thomas C. Peterson, Dian J. Seidel, and Keith P. Shine. 2011. Tropospheric Temperature Trends: History of an Ongoing Controversy. *Wiley Interdisciplinary Reviews: Climate Change* 2 (1): 66–88. https://doi.org/10.1002/wcc.80.

van Fraassen, Bas. 2008. *Scientific Representation: Paradoxes of Perspective*. Oxford: Oxford University Press.

Vinnikov, Konstantin Y., Norman C. Grody, Alan Robock, Ronald J. Stouffer, Philip D. Jones, and Mitchell D. Goldberg. 2006. Temperature Trends at the Surface and in the Troposphere. *Journal of Geophysical Research* 111 (D3). http://cat.inist.fr/?aModele=afficheN&cpsidt=17645256.

Vinnlkov, Konstantin Y., and Norman C. Grody. 2003. Global Warming Trend of Mean Tropospheric Temperature Observed by Satellites. *Science* 302 (5643): 269–272.

Wallace, John, John R. Christy, Dian J. Gaffen, Norman C. Grody, James Hansen, David Parker, Thomas C. Peterson, et al. 2000. *Reconciling Observations of Global Temperature Change*. Washington, DC: Panel on Recording Temperature Observations, National Research Council, National Academy of Sciences. https://www.nap.edu/read/9755/chapter/1. Accessed 29 May 2017.

Wentz, Frank J., and Matthias Schabel. 2000. Precise Climate Monitoring Using Complementary Satellite Data Sets. *Nature* 403 (6768): 414–416. https://doi.org/10.1038/35000184.

Winsberg, Eric. 2010. *Science in the Age of Computer Simulation*. Chicago: University of Chicago Press.

World Climate Report. 2003. Structure of Scientific Devolution 8 (18). http://www.worldclimatereport.com/archive/previous_issues/vol8/v8n18/feature.htm.

Zou, Cheng-Zhi, Mei Gao, and Mitchell D. Goldberg. 2009. Error Structure and Atmospheric Temperature Trends in Observations from the Microwave Sounding Unit. *Journal of Climate* 22 (7): 1661–1681. https://doi.org/10.1175/2008JCLI2233.1.

7

Reconciling Climate Model/Data Discrepancies: The Case of the 'Trees That Didn't Bark'

Michael E. Mann

One way scientists attempt to validate theoretical models of Earth's climate is to measure their predictions against real-world observations. There is always the danger in this process, however, that the models may be artificially tuned, directly or indirectly, to get key climate attributes right. For example, there may be a tendency for scientists to choose values of uncertain parameters governing both the sensitivity of the climate to increasing greenhouse gas concentrations and the offsetting cooling impacts of industrial aerosol emissions in such a way that models correctly reproduce the observed warming trend of the past century. There is some evidence that such "compensation" may have led to artificially small spreads in the estimated uncertainty ranges in key climate parameters (Andreae et al. 2005).

It is therefore useful to employ a variety of observations from both the present and past, as independent constraints on climate model behavior. This is particularly true of efforts to estimate the equilibrium climate

M.E. Mann (✉)
Department of Meteorology, The Pennsylvania State University, University Park, PA, USA

E.A. Lloyd, E. Winsberg (eds.), *Climate Modelling*,
https://doi.org/10.1007/978-3-319-65058-6_7

sensitivity ("ECS")—a key measure of our impact on the climate that is defined by the eventual warming we expect in response to a doubling of CO_2 concentrations relative to pre-industrial levels—levels we will see in a matter of decades under business-as-usual fossil fuel emissions. Various independent lines of evidence that can be brought to bear on the problem of estimating ECS include (Fig. 7.1) the ability of models to reproduce modern-day climatology, the cooling response of the climate to modern volcanic eruptions, the temperature changes during the last glacial

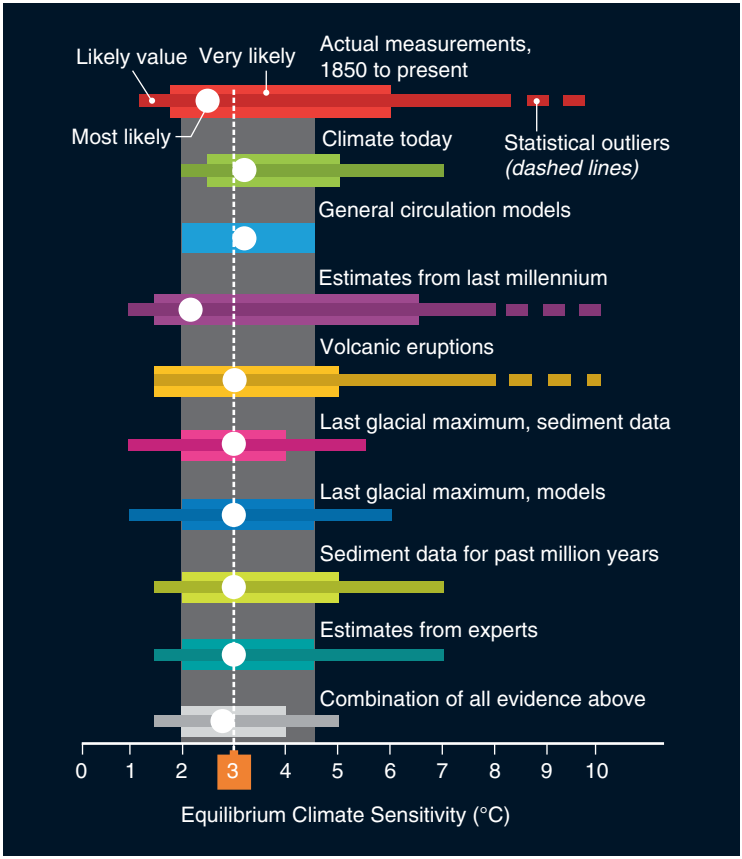

Fig. 7.1 Estimates of the equilibrium climate sensitivity ("ECS") based on various independent lines of evidence summarized by Knutti and Hegerl (2008) (Modified from Mann 2014 Scientific American)

maximum period, and the changes in temperature associated with geological variations in greenhouse gas concentrations, among others. These different constraints point to a range for ECS of somewhere between 1.5 °C and 5 °C warming, with a mid-range/most likely value close to 3 °C. While most lines of evidence are broadly consistent with each other, there is at least one notable discrepancy: comparisons of simulations of temperature changes over the past millennium with paleoreconstructions of past temperature (the reconstructions are typically based primarily on tree rings, but they are often supplemented by information from corals, ice cores, lake sediments, and other climate "proxy" data). These comparisons (e.g., Hegerl et al. 2006) tend to suggest an ECS value toward the lower end of the range, closer to 2 °C than the mid-range of 3 °C.

This discrepancy is conspicuous enough to demand some level of additional scrutiny. In particular, it is important to consider what is driving the ECS estimate in these comparisons. In the centuries leading up to the industrial area of anthropogenic influence, the primary forcing of climate was from natural changes in radiative forcing associated with factors such as the gradual changes in the distribution of solar insolation associated with millennial-scale earth orbital variations, modest (small fraction of a percent) estimated changes in solar output on multidecadal and centennial timescales, and small but non-negligible natural fluctuations in greenhouse gas concentrations. The cooling effect of stratospheric aerosols (particles such as sulfates which reflect incoming sunlight) associated with intermittent but sizeable explosive volcanic eruptions, however, yields the greatest pre-anthropogenic radiative forcing of climate over the past millennium. The eruption of Tambora in 1815, for example, is estimated to have been twice as large, in terms of radiative forcing (~−4 W/m^2), as the largest eruptions recorded in the historical period (e.g., Krakatoa in 1883 and Pinatubo in 1991, both ~−2 W/m^2). The tropical eruption of AD 1258 is estimated as somewhere between three and four times as large (between −8 and −12 W/m^2). Volcanic forcing turns out to be by far the largest climate forcing in the pre-industrial era of the past millennium (see, e.g., Jansen et al. 2007). Hence, climate models driven by estimated natural radiative forcing changes over the past millennium yield temperature changes that are largely representative of the response to volcanic forcing. If either the model simulations or the

paleoreconstructions misestimate the amplitude of this signal, estimates of ECS from those comparisons will accordingly be biased. Indeed, any errors in (a) the volcanic radiative forcing used to drive the climate models, (b) the model-estimated responses to that forcing, (c) the volcanic cooling as estimated by the paleoreconstructions, or (d) any combination thereof, will lead to biased estimates of ECS as inferred from model/data comparisons over the past millennium.

In this article, I summarize evidence that such biases do indeed exist. Specifically, I show that the paleoreconstructions may selectively underestimate the cooling signal associated with large explosive volcanic eruptions of the past millennium. I discuss my previously posed hypothesis (see Mann et al. 2012a) that the underestimation of volcanic cooling arises from a problem specific to the reliance of paleoreconstructions on tree-ring data from treeline-proximal environments, which leads to potential loss of sensitivity to large summer cooling events associated with major explosive volcanic eruptions. This loss of sensitivity potentially results in chronological errors in some subset of tree-ring records used to reconstruct past temperatures.

Requiring that model simulations match the resulting artificially muted volcanic cooling signal may lead to low-biased estimates of ECS. I review the challenges to our hypothesis that have been published, the additional work that we have done in response to those challenges that substantiates the viability of the hypothesis, and a recently proposed test that both proponents and critics of the hypothesis appear to agree would objectively determine whether chronological errors do compromise the integrity of tree-ring-based estimates of past volcanic cooling. Finally, I show that, regardless of the precise reason for the discrepancy, the mismatch between the paleoreconstructed and model-simulated volcanic cooling for a small number of large pre-industrial volcanic eruptions drives the anomalously low apparent values of ECS derived from comparisons of the past millennium. We demonstrate that there are ways to alleviate the impact of these events on the process of estimating ECS from model/data comparisons of the past millennium, and that doing so yields inferences more consistent with other independent lines of evidence.

7.1 Hypothesis Posed

Back in 2012, my co-authors and I published an article (Mann et al. 2012a—henceforth "MFR12") providing a new hypothesis for the enigmatic discrepancy between the tree-ring reconstructed and climate model-predicted magnitude of volcanic cooling in the Northern Hemisphere (NH) mean temperatures during the pre-industrial era of the past millennium. Most notable among the discrepancies is the virtual absence of cooling in tree-ring reconstructions of NH mean temperatures during what ice core and other evidence suggest is the largest explosive volcanic eruption of the past millennium—the AD 1258 eruption (see Emile-Geay et al. 2008 for a review of evidence for a wide-spread global climate impact of this eruption). We suspected that the discrepancy (the trees that didn't bark) might have something to do with the particular types of tree-ring information that were used to reconstruct past temperatures.

Tree rings are used as proxies for climate because trees create unique rings each year that often reflect the weather conditions that influenced the growing season that year. When seeking to reconstruct past temperature changes, tree-ring researchers (dendroclimatologists) typically seek trees growing at the boreal or alpine tree line, since temperature is most likely to be the limiting climate variable in that environment. This choice may prove problematic under certain conditions however. Trees in such environments are close enough to the summer temperature minimum threshold for growth that a lowering of temperatures by just a couple of degrees during the growing season may yield little or no growth and a consequent loss of sensitivity of tree growth to further cooling. In extreme cases, there may be no growth ring at all. If no ring is formed in a given year, that creates a further complication, introducing an error in the chronology established by counting rings back in time.

We investigated the potential impact of this problem by comparing a tree-growth model driven with climate model simulations of the past millennium with the model-simulated temperatures and tree-ring reconstructions of temperatures. The tree-growth model simulates the dependence of the thickness of growth rings on growing season temperature,

based on an empirical growth response curve that accounts for the temperature thresholds governing tree growth (see Mann et al. 2012a for further details). Climate models were driven with estimated natural (volcanic+solar) and anthropogenic forcings over the past millennium. We employed two different climate model simulations: (1) the simulation of the NCAR CSM 1.4 coupled atmosphere-ocean General Circulation Model (GCM) analyzed by Ammann and Wahl (2007) and (2) simulations of a simple Energy Balance Model (EBM). While the GCM provides a more comprehensive and arguably realistic description of the climate system, the computational simplicity of the EBM lends itself to extensive sensitivity tests. As the target for our comparison, we used a state-of-the-art tree-ring-based NH mean temperature reconstruction of D'Arrigo et al. (2006—henceforth "D06"). The reconstruction was based on a composite of tree-ring annual ring width series from boreal and alpine tree-line sites across the NH, and made use of a very conservative ("RCS") tree-ring standardization procedure designed to preserve as much low-frequency climatic information as possible.

Interestingly, the long-term variations indicated by the model simulations compared remarkably well with those documented by the tree-ring reconstruction (Fig. 7.2), showing no obvious sign of the potential biases in the estimated low-frequency temperature variations that have been the focus of some previous work (see e.g., Jones and Mann 2004 for a discussion). Instead, the one glaring inconsistency was in the high-frequency variations, specifically, the cooling response to the largest few tropical eruptions, AD 1258/1259, 1452/1453 and the 1809 + 1815 double pulse of eruptions, which is sharply reduced in the reconstruction relative to the model predictions. Indeed, this was found to be true for any of several different published volcanic forcing series for the past millennium, regardless of the precise geometric scaling used to estimate radiative forcing from volcanic optical depth, and regardless of the precise climate sensitivity assumed.

Following the AD 1258 eruption, the climate model simulations predict a drop of 2 °C, but the tree-ring-based reconstruction shows only about a 0.5 °C cooling. Equally vexing, the cooling in the reconstruction occurs several years late relative to what is predicted by the model. The other large eruptions showed similar discrepancies. An analysis using

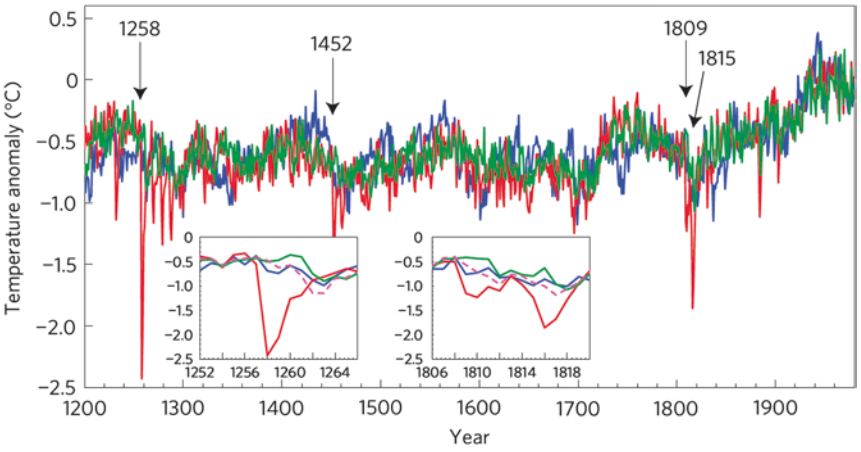

Fig. 7.2 Shown in the above is the D'Arrigo et al. tree-ring-based NH reconstruction (*blue*) along with the climate model (NCAR CSM 1.4) simulated NH mean temperatures (*red*) and the "simulated tree-ring" NH temperature series based on driving the biological growth model with the climate model-simulated temperatures (*green*). The two insets focus on the response to the AD 1258 and AD 1809+1815 volcanic eruption sequences. Also shown in the insets are the results (dashed magenta) when the volcanic diffuse-light impact is ignored (From Mann et al. (2012a))

synthetic proxy data with spatial sampling density and proxy signal-to-noise ratios equivalent to those of the D06 tree-ring network (see MFR12 for further discussion) suggest that these discrepancies cannot be explained in terms of either the spatial sampling/extent or the intrinsic "noisiness" of the network of proxy records. However, using a tree-growth model that accounts for the temperature growth thresholding effects discussed above, combined with the complicating effects of chronological errors due to potential missing growth rings, explains the observed features remarkably well (see green curve in Fig. 7.2).

The attenuation of the response is produced primarily by the loss of sensitivity to further cooling for eruptions that place growing season temperatures close to the lower threshold for growth. The smearing and delay of the cooling, however, arises from another effect: when growing season lengths approach zero, we assume that no growth ring will be detectable for that year. That means that an age model error of one year will be introduced into the chronology counting back in time. As multiple large eruptions are encountered further back in time, these age model errors

accumulate. This factor would lead to a precise chronological error, rather than smearing of the chronology, if all tree-line sites experienced the same cooling. However, stochastic weather variations will lead to differing amounts of cooling for synoptically distinct regions. That means that in any given year, some regions might fall below the "no ring" threshold, while other regions do not. That means that different chronological errors accumulate in synoptically distinct regions of the NH. In forming a hemispheric composite, these errors thus lead to a smearing out of the signal back in time as slightly different age model errors accumulate in the different regions contributing to the composite.

Accounting for this effect, our model accounts not only for the level of attenuation of the signal, but the delayed and smeared out cooling as well. This is particularly striking in comparing the behavior following both the AD 1258 and AD 1809 eruptions (compare the green and blue curves in the insets of the figure). Our model, for example, predicts the magnitude of the reduction of cooling following the eruptions and the delay in the apparent cooling evidence in the tree-ring record (i.e., in AD 1262 rather than AD 1258). We have also included a minor additional effect in these simulations. While volcanic aerosols cause surface cooling due to decreased shortwave radiation at the surface, they also lead to *increased* indirect, scattered light at the surface. Plant growth benefits from indirect sunlight, and past studies show that, e.g., a Pinatubo-sized eruption (roughly −2 W/m^2 radiative forcing) can result in a 30% increase in carbon assimilation by plants. This effect turns out to be relatively small because it is proportional in nature, and thus results in a very small absolute increase when growth is suppressed in the first place by limited growing seasons. However, *not* including this effect results in a slightly worse reproduction (purple dashed curves in the two insets of the figure) of the observed behavior.

As shown in MFR12, the central conclusions discussed above are insensitive to the precise details of the forcing estimates used, the volcanic scaling assumptions made, and the precise assumed climate sensitivity. They are also insensitive to the details of the biological tree-growth model over a reasonable range of model assumptions. Our conclusions would nonetheless soon be challenged by other scientists.

7.2 Hypothesis Challenged

The conclusion that tree-ring temperature reconstructions might suffer from age model errors due to missing rings is controversial, and it is important to recognize that it is only a working hypothesis for explaining some enigmatic features of tree-ring temperature reconstructions, more specifically, the *attenuation*, and the increasing (back in time) *delay* and *temporal smearing* in association with the response to past volcanic forcing. Were an equally successful and more parsimonious hypothesis to be provided for these features, we would be the first to concede to this alternative explanation. It was my hope that our hypothesis as presented in MFR12 would encourage a healthy discussion within the paleoclimate community, whether or not it ultimately stands up to additional scrutiny. In particular, it was my hope that dendroclimatologists might, in response to our work, go back and reassess their raw tree-ring chronologies more carefully, and critically assess the extent to which the artifacts we predicted might indeed be present in the underlying tree-ring data.

Initially, however, we instead encountered what might be considered a blanket dismissal of our hypothesis. A group comprised of the majority of leading tree-ring researchers in the United States and Europe published a comment (Anchukaitas et al. 2012—henceforth "A12") that criticized various aspects of our analysis, but did not provide a plausible alternative explanation for the vexing problem we had identified. Our response (Mann et al. 2012b) appeared along with the comment. A12 suggested that our study represented a fundamental challenge to the validity of large-scale tree-ring-based reconstructions in general, but that is certainly not the case. As we noted in our response, in MFR12 we showed that tree-ring reconstructions effectively capture long-term temperature trends. We were simply questioning the ability of tree-ring width proxies to detect the short-term cooling associated with the largest few volcanic eruptions of the past millennium.

A12 criticized our study for not using more elaborate tree-growth models that include other influences (e.g., precipitation), but this rather misses the point. The fundamental assumption underlying tree-ring-based temperature reconstructions such as those we analyzed is that

annual growth at temperature-limited tree-line locations yields an unbiased estimate of temperature changes exclusively. A12 further criticized our tree-growth parameter choices, and suggested that these parameter values yield an unrealistic prediction of missing twentieth-century tree rings. However, as we noted in our response, our analysis predicted no missing tree rings for the twentieth century. Our value of 10 °C as a threshold temperature for growth is at the upper end of the accepted 3–10 °C range, but this choice yields the closest fit to the observed tree-ring response, and we see qualitatively similar results for a lower temperature threshold value.

Addressing A12's criticism over the specifics of our tree-growth model, we demonstrated that similar results are obtained using the simplest possible (growing degree day) model, which involves a linear growth response above a threshold temperature. Using that model, we showed that the underestimation of volcanic cooling by tree rings is substantial for threshold values spanning the entire upper half of the 3–10 °C range, even using a conservative assumption of what constitutes a missing ring (a growing season of less than one week). Including the effect of increased diffuse light caused by volcanic aerosols—an important factor neglected by A12—leads to better agreement between our growth model and existing tree-ring reconstructions. For growth-model assumptions substantially different from those we adopted, however, the effect produces offsetting and spurious warming responses in the first few years following an eruption (see Mann et al. 2012a).

A12 sought to reconcile the lack of the expected cooling response to the AD 1258/1259 in the D06 tree-ring reconstruction by arguing that the radiative forcing might have been smaller than generally assumed. However, as we showed in MFR12, our findings are robust with respect to which of the various published volcanic forcing reconstructions or volcanic scaling assumptions are used. Moreover, changing the estimated radiative forcing associated with the AD 1258/1259 eruption would not explain other problematic features in the tree-ring reconstructed response. Our analysis, by contrast, provides a plausible explanation for why cooling is observed four years later than expected, and is greatly diminished in magnitude. Our hypothesis also explains a similar discrepancy between the tree-ring reconstruction and the cooling associated with the 1815

Tambora eruption. Importantly, this latter eruption is constrained by observational surface temperature data (Rohde et al. 2013). These data (a) confirm the model-estimated cooling and (b) contradict the muted/absent cooling in the tree-ring estimates.

Perhaps most importantly, we did not argue, as A12 seemed to suggest, that tree rings are uniformly recording the wrong year of the eruption in a way that can be diagnosed just by looking at composite series. Instead, we suggest that sufficiently many individual tree-ring records within the composites are likely to have dating errors due to potential missing/undetected rings following the largest volcanic eruptions that the cooling signal is muted and smeared in the large-scale averages.

One argument against the specific conclusion of missing growth rings is that trees are carefully cross-dated when forming regional chronologies, and this precludes the possibility of chronological errors. That, however, assumes that there are at least some trees within a particular region that will not suffer a missing ring during the years where our model predicts it. Yet our prediction is that *all trees* within a region of synoptic or lesser scale where growing season temperatures lie below the growth threshold will experience a missing ring. Thus, cross-dating within that region, regardless of how careful, cannot resolve the lost chronological information.

As we noted in our response, it should be possible to further investigate this hypothesis through a careful analysis of the detailed patterns of response to the largest eruptions among individual tree-ring chronologies distributed over the globe.

7.3 Additional Evidence

As we have seen, subsequent to the publication of MFR12 there was a vigorous debate about the viability of our hypothesis for the muted, delayed volcanic cooling signal in tree-ring composite-based reconstructions of hemispheric temperature change. Chief among the criticisms is that our hypothesis was based entirely on theoretical modeling, and that we had provided no empirical evidence for the claim of missing tree rings—an important component of our mechanism for the underestimation,

smearing and delay of the volcanic cooling signal in tree-ring-based temperature reconstructions. In subsequent work (Mann et al. 2012b), we attempted to provide precisely that evidence.

It is necessarily more challenging to prove that something is missing than to prove it is present. Though local cross-dating of trees can be used to identify missing rings in individual cores contributing to local chronologies developed from nearby trees, it cannot reliably identify a coherent large-scale pattern of missing rings across an entire climatic region experiencing sub-growth limit summer temperatures, as MFR12 predicts to be the case following the largest few tropical volcanic eruptions. A more nuanced approach is required to detect the influence of missing rings.

We instead attempt to account for the effects of missing rings in some subset of the underlying tree-ring chronologies. We employed the original tree-ring data used by D06, which consists of a maximum of 66 distinct site chronologies representing 19 different regions back to 1686, decreasing to eight regions back to AD 1190 (we used the conventionally standardized tree ring series of D06, but broadly similar results were obtained using the alternative "RCS" standardization; see Mann et al. 2013). We performed Monte Carlo simulations using the MFR12 estimates of the timing and probabilities for a missing ring in a given year, yielding alternative versions of the D06 tree-ring series consistent with estimated chronological (age model) errors. Using these surrogate tree-ring series, we generated an ensemble of alternative *regional* composites consistent with estimated tree-age model uncertainties (e.g., the chance of a given region missing a ring in any particular realization is 90% in AD 1258, and 55% in AD 1816 as prescribed by MFR12—note that our net estimated age model errors amount to <1%, i.e., no more than 6 years out of 700+). This procedure was used to generate a large ensemble of surrogate hemispheric temperature reconstructions based on averaging the surrogate regional series emulating the procedures of D06 (see Mann et al. 2013 for further details). In principle, some subset of these surrogates should correct for the age model errors (i.e., missing rings).

As shown in Fig. 7.3, some of the surrogate reconstructions indeed suggest significantly greater cooling in association with the major volcanic eruptions. For the AD 1258 eruption, a large number of Monte Carlo

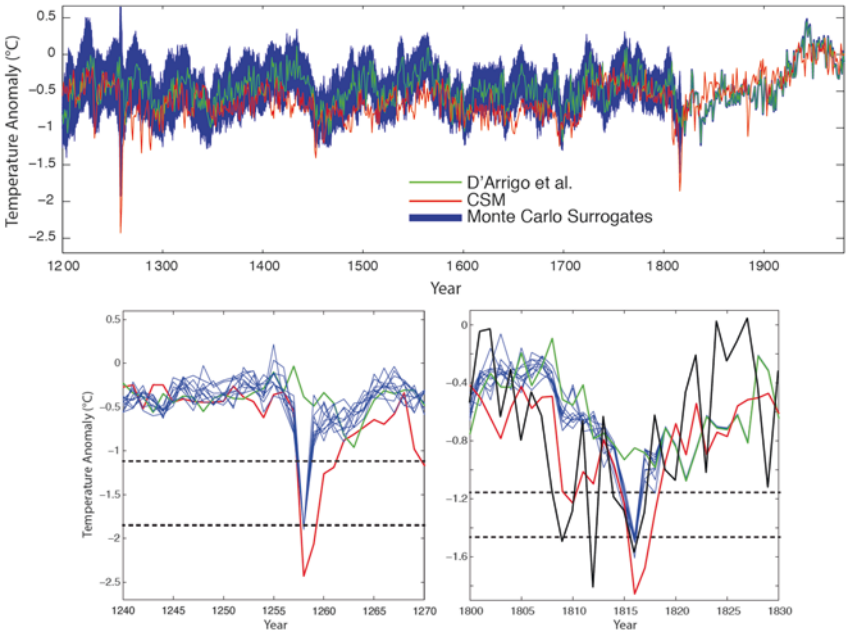

Fig. 7.3 Ensemble of hemispheric tree-ring temperature reconstructions derived from available regional tree-ring composites resampled to account for predicted age model errors. Shown are the raw composite based on the D'Arrigo et al. (2006) tree-ring data (*green*), Monte Carlo surrogate reconstructions (8000 in total—blue curves), and GCM simulation (*red*). Insets: Expanded views of the response to the AD 1258/1259 and AD 1815 eruptions responses showing the 10 coldest surrogates (*blue*) for each eruptions and the 2 and 4 sigma significance thresholds for cooling (*dashed black*). Shown also for AD 1815 eruption is the recently back-extended instrumental NH land temperature record of Rohde et al. (2013) (*black*). Centering of all series is based on a 1961–1990 modern base period (From Mann et al. (2013))

surrogates point toward a distinct ~2 °C cooling in AD 1258 (lacking the enigmatic delayed and reduced 1260–62 cooling signal seen in the raw reconstruction). The increased AD 1258 cooling and disappearance of (likely spurious) AD 1260–62 cooling is seen to arise from a realignment of much larger cooling signals that are present in individual tree-ring series but interfere destructively before they are brought into alignment (see Mann et al. 2013). The year AD 1816 is far more consistent with its moniker as the "Year Without a Summer," with surrogates showing

cooling of up to ~1.6 °C. The amplified cooling is not only far more consistent with the model-predicted cooling, but agrees far better with the available instrumental temperature record. These enhanced cooling responses that arise from permuting the tree-ring data within estimated age model errors are highly significant relative to the null hypothesis of chance occurrence due to random sampling variations from the Monte Carlo procedure (see Mann et al. 2013).

We thus argue that the missing rings in regional tree-ring temperature composites as hypothesized in MFR12 are not only plausible from a theoretical perspective, but appear to be detectable in the actual underlying regional tree-ring series and resulting hemispheric composites. Attempts to correct for the estimated chronological errors yield far greater post-volcanic cooling responses that agree with model predictions.

7.4 Wider Implications

I return now to the issue of why a seemingly technical and mundane matter involving tree rings and volcanic eruptions actually matters. As noted earlier, the apparent weak response of surface temperatures to the few largest eruptions of the past millennium as inferred from proxy temperature reconstructions is what drives estimates of relatively low ECS as derived from proxy reconstructions based either entirely or substantially upon tree-ring data (Hegerl et al. 2006). Hegerl et al. (2006) for example used comparisons during the pre-industrial period of EBM simulations and proxy temperature reconstructions based entirely or partially on tree-ring data to estimate ECS. Hegerl et al. (2006) ended up arguing for a substantially lower 5–95% range of ECS (1.5–6.2 °C) than is evident from other lines of evidence (see Fig. 7.1). As the primary radiative forcing during the pre-industrial period is from volcanic forcing, their conclusions were leveraged by the muted apparent response to very large past volcanic eruptions. If that muted response is an artifact, as our work suggests it to be, the resulting estimates of ECS are almost certainly downwardly biased. Moreover, this one potentially biased constraint on ECS (central value about 2.1 °C—see Fig. 7.1) is enough of an outlier (nearly all other lines of evidence point to an ECS value at or slightly

above 3.0° C) that it ends up downwardly biasing the "combined" estimate of ECS (Fig. 7.1), taking it from 3.2 °C to roughly 2.8 °C, a non-trivial lowering of nearly 0.5 °C. Our findings therefore suggest that prevailing estimates of ECS from combinations of various lines of evidence (e.g., Knutti and Hegerl 2008) have likely underestimated the true climate sensitivity.

In Mann et al. (2013), we assessed the impact that the underestimation of volcanic cooling from tree-ring reconstructions as estimated by MFR12 would have on inferred values of ECS. Our analysis employed EBM simulations where the actual value of ECS is precisely known (it was set to the canonical mid-range value of 3 °C) and is then estimated using the simulated tree-ring response. We found that the truncation of volcanic cooling alone led to a decrease in apparent ECS from 3.0 °C to 1.7 °C in simulations of the pre-industrial interval AD 1200–1849. That calculation did not take into account the additional degradation by estimated chronological errors. When chronological errors are accounted for, the estimated ECS value drops to less than 1.0 °C—similar to the ECS value estimated using the D'Arrigo et al. (2006) tree-ring reconstruction. Using a later period AD 1300–1849, which eliminates the influence of the AD 1258 eruption, leads to a lesser but still large impact on ECS values (ECS ~2.0 °C without considering chronological errors, and ECS ~1.0 °C with chronological errors accounted for). These estimates pertain only to tree-ring-based temperature reconstructions. Most proxy-based reconstructions of past temperature instead use a mix of proxy data, including corals, ice cores, sediments, and other types of proxy information. For such reconstructions, we might expect a smaller underestimation of volcanic cooling than estimated for tree-ring only temperature reconstructions, and potentially a smaller bias in ECS estimates derived from the reconstructions. However, even if the estimated impact is reduced by a factor of two or three, it is large enough to explain the discrepancy between "last millennium" estimate of ECS and ECS estimates derived from the remaining lines of evidence (Fig. 7.1).

It is reasonable to ask whether our principal conclusions hold up even if the specifics of our hypothesis about the underestimation of volcanic cooling by tree-ring temperature reconstructions do not. We addressed that matter in additional work (Schurer et al. 2013) using an alternative

approach. We employed a method wherein a large ensemble of state-of-the-art climate model simulations of the past millennium—the Coupled Model Intercomparison Project 5 (CMIP5) "past millennium" simulations—were used to estimate the "fingerprints" of the various natural radiative forcings of climate which include solar irradiance, Earth orbital changes, natural variations in greenhouse gas concentrations, and explosive volcanic eruptions. The amplitudes of those fingerprints were then estimated (via total least squares regression) from nine different proxy-based reconstructions of NH mean temperature spanning all or most of the past millennium. The amplitudes estimated from the paleoclimate reconstructions were then compared against the model-predicted amplitudes. The ratio of the two ("β") measures whether the reconstruction indicates a greater ($\beta > 1$), comparable ($\beta \sim 1$), or lesser ($\beta < 1$) amplitude than predicted by the models.

The procedure was performed using a variety of sub-intervals of the period 851–1950 as well as the full interval and the full pre-industrial interval AD 851–1850. With only one exception (a controversial reconstruction that exhibits far greater variability than all others), the reconstructions yielded estimates of β that are systematically less than unity (i.e., the entire uncertainty range for β lies below unity). However, if the few largest eruptions (which include the AD 1258, the AD 1453 Kuwae, and 1815 Tambora eruptions) are simply masked from the analysis (so that the analysis is based on the response to all other radiative forcing, i.e., moderate eruptions, solar irradiance changes, greenhouse gas concentrations, and Earth-Orbital changes), and the procedure is repeated, then remarkably, most of the β values are consistent with a value of unity within the associated error bars. In other words, if the largest eruptions of the past millennium are included in the analysis, the reconstructions indicate a response to forcing that is systematically smaller than predicted by the models. Yet if just that handful of eruptions is masked out, the reconstructions indicate a response that is consistent with the model simulations.

It is important to recognize that there are a number of sources of potential uncertainty and bias that contribute to these model/data comparisons in addition to potential biases in the proxy reconstructions. These include uncertainties or biases in the estimates of radiative forcings,

and uncertainties or biases in the models' response to radiative forcings. This latter uncertainty/bias is tied in part to the uncertainty in the associated ECS, though there are also potential uncertainties and/or biases in climate responses that are specific to the way particular forcings are represented in the models. For example, in the case of volcanic radiative forcing there is some uncertainty in how volcanic aerosol size distributions are represented (see, e.g., MFR12; Mann et al. 2012b, 2013). Any combination of these uncertainties or biases can contribute to the model/data misfit.

That notwithstanding, the simplest interpretation of the above findings is that the climate models, including the ECS values that characterize their response to radiative forcing, are consistent with the paleoreconstructions if the response to the few largest volcanic eruptions are masked out in the analysis. That implies that the reduced apparent response to forcing in the reconstructions overall arises entirely from the discrepancy between the apparent and predicted response to volcanic radiative forcing. That finding, in turn, is consistent with the proposition that it is the specific discrepancy between the model-predicted and proxy reconstruction-estimated response to the few largest volcanic eruptions of the past millennium that leads to anomalously low values of apparent ECS in studies using paleoreconstructions of the past millennium such as Hegerl et al. (2006). That conclusion does not establish that the source of this discrepancy is the tree-growth saturation mechanism proposed by MFR12, but it provides independent support for the existence of some source of bias that is limited to the apparent response of the climate to the few largest volcanic eruptions of the past millennium.

7.5 The Gauntlet Is Laid Down

In a recent comment, Büntgen et al. (2014) provide a potential way forward to resolve definitively whether or not the specific tree-ring age model errors predicted by MFR12 (and further supported by Mann et al. 2013) can be established in the actual data. The authors demonstrate the existence of a distinct radiocarbon event during AD 774–775, which has consistently been recorded by trees in disparate locations including Japan,

Germany, and the Alps, thus establishing that the dating of these trees is consistent and accurate.

Our hypothesis, as presented in MFR12, is that some trees growing near their thermal limits, as is the case with many trees selected for paleotemperature reconstructions which lie at the boreal or alpine tree line, can fail to produce an annual ring during unusually cold growing seasons following particularly large volcanic eruptions. The missing ring causes the year preceding the eruption to masquerade as the eruption year. Thus, the resulting chronology would not record the effects of the eruption because the ring from that year is missing, and all previous years in the chronology are shifted forward in time by the number of missing rings. This means that, even if the tree produced a growth ring following an older eruption, that ring would appear in the wrong year. The radiocarbon event of AD 774–775 provides a globally synchronous signature that ought to provide a unique, independent time marker that can be used to test our hypothesis.

As described by Rutherford and Mann (2014), we can make very specific predictions based on our hypothesis that can be tested using the radiocarbon event and existing tree-ring chronologies. With regard to the Alps series, the results from Mann et al. (2013) predict that there will be no missing rings in this region. The D'Arrigo et al. (2006) Alps regional series begins in AD 1350, and was included in our analysis of the climate response to the 1815/16 Tambora eruption sequence. Our "best match" surrogate ensembles for this eruption (Fig. 7.2 of Mann et al. 2013) use the Alps series on its original time scale. Our results are therefore consistent with the Büntgen et al. (2014) finding that there is no age model error with this series.

Of the 19 regional series used in D'Arrigo et al. (2006) and Mann et al. (2013), only three (Coastal Alaska, Tornestraesk, and Taymir) begin before AD 774 and can thus be directly tested using the AD 774/775 radiocarbon event. The results from Mann et al. (2013) predict the following minimum offsets for the event in these three series: the Coastal Alaska series should be four years too young, the Tornestraesk series should be one to five years too young, and the Taymir series should be one year too young (Fig. 7.4). In addition, the Mann et al. (2013) results predict that the "Icefields" series dates correctly, but as it begins in AD

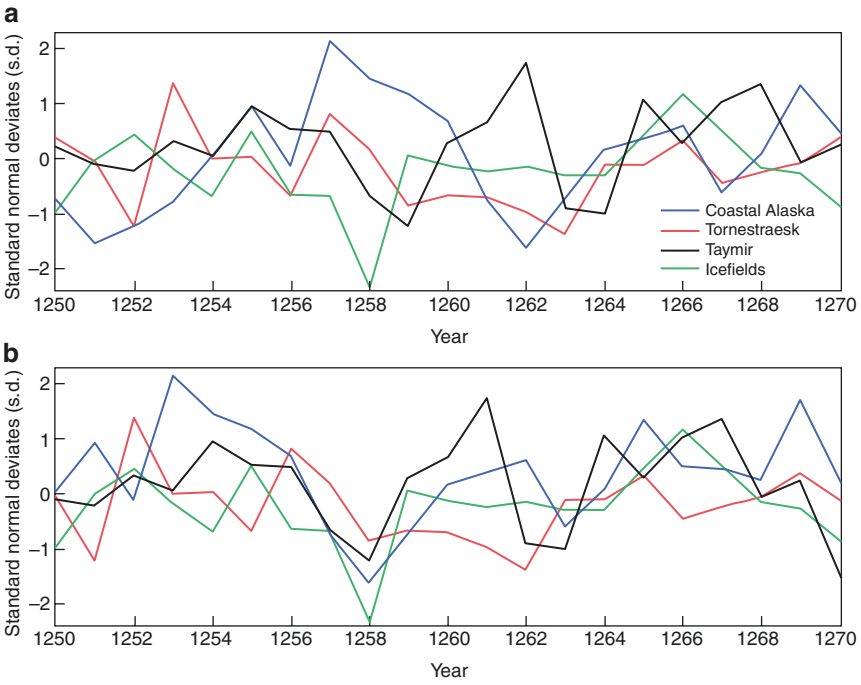

Fig. 7.4 Tree-ring records across the AD1258 eruption. The three D'Arrigo et al. regional series that begin before AD774 (Coastal Alaska, Tornestraesk, and Taymir), along with the Icefields series for reference, are shown on their original time scale (**a**) and age-adjusted (**b**) in a way consistent with our hypothesis. The Icefields series is unaltered, the Coastal Alaska series is shifted four-years older (~0.6%), and the Tornestraesk and Taymir series are both shifted one year older (~0.1%) (From Rutherford and Mann (2014))

918, its age model cannot be validated with the AD 774/775 radiocarbon event.

Thus, the MFR12 hypothesis that missing growth rings due to unusually cold summers at tree line following the few largest volcanic eruptions of the past millennium is now testable. It will be up to dendroclimatologists and/or dendrochronologists to go back and examine the specific chronologies mentioned above which we predict to contain missing rings and check, using the AD 774/775 radiocarbon date to assess whether there are any age model errors in these chronologies.

7.6 Closing Thoughts

As alluded to by the title of this piece, what led to the hypothesis explored in this article isn't what was evident in paleoclimatic reconstructions of the past millennium, but instead, what *wasn't* evident. Much as with Sherlock Holmes and the "curious incident of the dog in the night-time [that didn't bark]," it is sometimes those things that we inexplicably can't see in the data that points to gaps in knowledge or understanding.

The scientific investigations summarized in this article grew out of an enigmatic observation that had bothered me for some time: paleoclimate reconstructions based partly or entirely on tree-ring data fail to show any evidence of large-scale cooling following what various lines of evidence indicate was the largest (from a radiative forcing standpoint) eruption of the past millennium, the AD 1258 tropical eruption. More generally, we found that the paleoclimate reconstructions indicate systematically less cooling following the largest volcanic eruptions than is predicted by climate models.

We are able to reproduce these observations based on simulations using a model of tree growth forced with climate model simulations of temperature over the past millennium. For values of the relevant parameters (i.e., the minimum temperature threshold for tree growth) within the cited range, we are able to reproduce the muted, delayed, and smeared cooling response to very large volcanic eruptions seen in tree-ring-based temperature reconstructions. These features are seen, in the simulations, to be an artifact of a maximum threshold on the cooling that can be recorded by tree-line-proximal trees, combined with the introduction of chronological age model errors in some subset of chronologies associated with a lack of growth during the growing season. The chronological errors accumulate differentially in different regions, leading to a smearing out of temperature signals in hemispheric composites that increases back in time.

While other researchers have raised various objections with our hypothesis and findings, we have been able to provide independent, indirect evidence that missing rings/chronological errors are indeed present in some subset of tree-ring chronologies based on Monte Carlo simulations that show that much larger volcanic cooling signals can be found in

hemispheric composites when the estimated age model errors are taken into account. Ours is just one potential hypothesis for the model/data discrepancies in question, and as discussed in this article, at least one aspect of our hypothesis—the existence of chronological errors in some subset of tree-ring chronologies—can now potentially be tested based on the radiocarbon event of AD 774/775. We await with great interest the results of these tests.

Whether or not our specific hypothesis is correct, however, we have shown that some of our key conclusions appear to be robust. In particular, there is very compelling evidence that the discrepancies between model simulations and paleoclimate reconstructions over the past millennium appear to be associated almost exclusively with the response to the few largest volcanic eruptions of the past millennium. It is clear that if one simply masks these eruptions from any model/data comparisons, then the model simulations and reconstructions are consistent. A corollary of this conclusion is that previous studies arguing for relatively low (~2 °C) ECS based on model/data comparisons over the past millennium likely suffer from a bias related to the underestimation of volcanic cooling in the reconstructions. That would explain why this one line of evidence for ECS gives a substantially lower estimate of ECS than essentially every other line of evidence. Finally, these findings provide additional support for the contention that the most likely value of ECS is in the range of 3.0 °C, and that previous assessments that consider, even partly, evidence from the last millennium, may have underestimated ECS. This conclusion is hardly a trivial one, as it provides support for the contention that the climate system is substantially sensitive to carbon emissions, and that business-as-usual fossil fuel burning may have a profound impact on Earth's climate.

References

Ammann, Caspar M., and Eugene R. Wahl. 2007. The Importance of the Geophysical Context in Statistical Evaluations of Climate Reconstruction Procedures. *Climatic Change* 85 (1–2): 71–88. https://doi.org/10.1007/s10584-007-9276-x.

Anchukaitis, Kevin J., Petra Breitenmoser, Keith R. Briffa, Agata Buchwal, Ulf Büntgen, Edward R. Cook, Rosanne D. D'Arrigo, et al. 2012. Tree Rings and Volcanic Cooling. *Nature Geoscience* 5: 836–837. https://doi.org/10.1038/ngeo1645.

Andreae, Meinrat O., Chris D. Jones, and Peter M. Cox. 2005. Strong Present-Day Aerosol Cooling Implies a Hot Future. *Nature* 435 (7046): 1187–1191.

Büntgen, Ulf, Lukas Wacker, Kurt Nicolussi, Michael Sigl, Dominik Güttler, Willy Tegel, Paul J. Krusic, and Jan Esper. 2014. Extraterrestrial Confirmation of Tree-Ring Dating. *Nature Climate Change* 4: 404–405. https://doi.org/10.1038/nclimate2240.

D'Arrigo, Rosanne, Rob Wilson, and Gordon Jacoby. 2006. On the Long-Term Context for Late Twentieth Century Warming. *Journal of Geophysical Research (Atmospheres)* 111: D03103. https://doi.org/10.1029/2005JD006352.

Emile-Geay, J., R. Seager, M.A. Cane, E.C. Cook, and G.J. Jaug. 2008. Volcanoes and ENSO Over the Past Millennium. *Journal of Climate* 21: 3134–3148.

Hegerl, Gabriele C., Thomas J. Crowley, William T. Hyde, and David J. Frame. 2006. Climate Sensitivity Constrained by Temperature Reconstructions over the Past Seven Centuries. *Nature* 440 (7087): 1029–1032. https://doi.org/10.1038/nature04679.

Jansen, Eystein, Jonathan Overpeck, and Keith R. Briffa. 2007. Paleoclimate. In *Climate Change 2007: The Physical Science Basis.* Working Contribution of Working Group 1 to the Fourth Assesment Report of the Intergovernmental Panel on Climate Change. Cambridge/New York: Cambridge University Press.

Jones, P.D., and M.E. Mann. 2004. Climate over Past Millennia. *Reviews of Geophysics* 42 (2): RG2002. https://doi.org/10.1029/2003RG000143.

Knutti, Reto, and Gabriele C. Hegerl. 2008. The Equilibrium Sensitivity of the Earth's Temperature to Radiation Changes. *Nature Geoscience* 1: 735–743. https://doi.org/10.1038/ngeo337.

Mann, Michael E. 2014. False Hope: The Rate of Global Temperature Rise May Have Hit a Plateau, but a Climate Crisis Still Looms in the Near Future. *Scientific American* 310: 78–81. https://doi.org/10.1038/scientificamerican0414-78.

Mann, Michael E., Jose D. Fuentes, and Scott Rutherford. 2012a. Underestimation of Volcanic Cooling in Tree-Ring-Based Reconstructions of Hemispheric Temperatures. *Nature Geoscience* 5: 202–205. https://doi.org/10.1038/ngeo1394.

———. 2012b. Reply to 'Tree-Rings and Volcanic Cooling'. *Nature Geoscience* 5 (12): 837–838.

Mann, Michael E., Scott Rutherford, Andrew Schurer, Simon F.B. Tett, and Jose D. Fuentes. 2013. Discrepancies Between the Modeled and Proxy-Reconstructed Response to Volcanic Forcing over the Past Millennium: Implications and Possible Mechanisms. *Journal of Geophysical Research: Atmospheres* 118 (14): 7617–7627. https://doi.org/10.1002/jgrd.50609.

Rohde, R., R.A. Muller, R. Jacobsen, E. Muller, S. Perlmutter, et al. 2013. A New Estimate of the Average Earth Surface Land Temperature Spanning 1753–2011. *Geoinfor Geostat: An Overview* 1: 1. http://static.berkeleyearth.org/papers/Results:Paper-Berkeley-Earth.pdf

Rutherford, Scott, and Michael E. Mann. 2014. Missing Tree Rings and the AD 774-775 Radiocarbon Event. *Nature Climate Change* 26: 648–649.

Schurer, Andrew P., Gabriele C. Hegerl, Michael E. Mann, Simon F.B. Tett, and Steven J. Phipps. 2013. Separating Forced from Chaotic Climate Variability over the Past Millennium. *Journal of Climate* 26 (18): 6954–6973. https://doi.org/10.1175/JCLI-D-12-00826.1.

8

Downscaling of Climate Information

L.O. Mearns, M. Bukovsky, S.C. Pryor, and V. Magaña

8.1 Introduction

Awareness of the potential inadequacy of the spatial scale of coupled atmosphere-ocean general circulation models (AOGCMs), for a variety of purposes, has been with us for a long time. When model projections were first used to determine the impacts of future climate on important resources such as crop yields and water resources (e.g., Liverman et al. 1986; Rosenzweig 1985; White 1985) the so-called mismatch of scale issue gained prominence.

L.O. Mearns (✉) • M. Bukovsky
National Center for Atmospheric Research (NCAR), Boulder, CO, USA

S.C. Pryor
Cornell University, Ithaca, NY, USA

V. Magaña
Universidad Nacional Autónoma de México, México City, Mexico

E.A. Lloyd, E. Winsberg (eds.), *Climate Modelling*,
https://doi.org/10.1007/978-3-319-65058-6_8

> Most GCMs neither incorporate nor provide information on scales smaller than a few hundred kilometers. The effective size or scale of the ecosystem on which climatic impacts actually occur is usually much smaller than this. We are therefore faced with the problem of estimating climate changes on a local scale from the essentially large-scale results of a GCM. (Gates 1985)

This concern has thus been registered for over 25 years, and has been reiterated numerous times (e.g., Carter et al. 1994; Wilby and Fowler 2012). However, the mismatch of scale between AOGCMs and impacts models (e.g., watershed modeling for water quality and quantity (Johnson et al. 2012)) is only one motivation for downscaling. The other major motivation for applying regionalization techniques is the need to resolve important processes at scales finer than those represented in AOGCMs that are important for simulating regional climate. Such processes may include local conditions such as narrow jet cores, sea breeze type circulations, lake effects, and the atmospheric response to complex topography and/or landscape heterogeneity. These purposes often go hand in hand, that is, they are far from mutually exclusive. However, it is important to differentiate these goals, since some downscaling techniques produce higher resolution data that may be adequate for deriving inputs for impacts models, but do not necessarily add information about finer-scale atmospheric processes.

Regardless of motive, the solution to the scale problem requires the application of one (or more) of a variety of so-called downscaling techniques. Downscaling refers to methods for developing regional or local information from coarser resolution information, usually generated from global climate models (discussed in Chap. 6). Another term that is sometimes used is "right-scaling," which refers to developing the appropriate spatial scale of information for a particular purpose.

Downscaling techniques, while available for more than a quarter century, are recently experiencing more intensive use, as finding solutions to the challenges presented by climate variability and change has become more urgent. This is particularly true in the case of adaptation research, planning, and implementation, which occur on regional to local scales (Wilby et al. 2009).

There have been a number of reviews of downscaling methods (e.g., Giorgi and Mearns 1991, 1999; Wilby and Wigley 1997; Giorgi et al.

2001; Fowler et al. 2007; Wilby and Fowler 2012), and there are a variety of means of categorizing the methods. In this chapter, we use three categories: simple downscaling and interpolation methods, statistical downscaling, and dynamical downscaling. These methods vary a great deal in terms of complexity, the computational and human resources needed to develop them, and what kind of "added value" they can produce.

Simple downscaling techniques, as the name implies, are generally the least complex, the least expensive, and have primarily been developed for producing higher resolution information from AOGCMs for driving impacts models. These techniques involve relatively simple manipulation of the coarser results from global models, particularly temperature and precipitation (Mearns et al. 2001). The simplest is the so-called "delta" method, whereby changes in climate (future vs. current) are applied to finer resolution observed data sets, thus producing a higher resolution changed climate data set that is also bias corrected. Another popular approach is the much more complex Bias Correction Spatial Disaggregation Method (BCSD) (Wood et al. 2002), wherein global climate model results are first bias corrected and then the corrected results are spatially disaggregated to a higher resolution.

Statistical downscaling generally refers to methods that statistically relate (often through regression techniques) larger scale atmospheric features from global climate models (the predictors), such as 500-mb geopotential heights, to local (typically point estimates) climate (predictand), for example, monthly temperature or precipitation. However, there are a number of different types of statistical downscaling that use different statistical techniques, such as neural networks, weather classification typing, weather generators, and so on (Giorgi et al. 2001; Fowler et al. 2007).

Dynamical downscaling refers broadly to all techniques that use some form of deterministic climate model. The main categories here include: high-resolution global atmospheric model time-slice experiments, stretched grid global models, and regional climate models. All of these methods are discussed in detail for the region of North America.

While the literature on downscaling is quite large, interestingly, a number of the central issues surrounding downscaling have not been resolved. The most important is whether, for dynamical downscaling methods in particular, but also statistical downscaling, greater confidence in the downscaled future climate has been robustly demonstrated. In the

many reviews and discussions of these methods, rarely are assertions of comparative value made; rather discussions tend to center on "advantages and disadvantages" of the various methods (e.g., Giorgi et al. 2001; Mearns et al. 2001; Wilby et al. 2009; Wilby and Fowler 2012), but a "value neutral" stance is usually taken.

In this chapter, we review these different techniques of downscaling from a methodological point of view, and assess their application over North America. We also compare the results across the different methods and attempt to draw conclusions regarding the value of these techniques for increasing our confidence in regional projections of climate change. Finally, we attempt to make some recommendations for research that would help to resolve some of the outstanding issues regarding downscaling.

8.2 Simple Downscaling and Interpolation Techniques

Delta Approach

As mentioned in the introduction, simple downscaling generally refers to the application of relatively straightforward techniques for creating greater spatial resolution, usually motivated by the higher resolution data requirements of climate impact models.

The simplest is the so-called "delta" method. Changes in climate determined by comparing the future climate and current climate simulated by a global or regional climate model are calculated and then these differences are combined with higher resolution observed data sets. Typically changes in temperature (often maximum and minimum) are added to observed temperature records, and ratios of precipitation (future divided by the current) are combined with observations through multiplication. Since the scenarios are constructed by modifying observations, the model biases are inherently corrected. The correction, however, only applies to the mean of the climate change. Higher-order moments are not corrected.

This method has been used over several decades in virtually every area of application (water resources, agriculture, ecology, etc.). Very often, impact models require daily time series of variables, so the monthly changes in the required climate variables are combined with the daily observed data. In this case, the mean of the observed time series of temperature is changed (by the amount of the "delta") but the variability (on daily to interannual time scales) remains the same. The multiplicative method of combining the change in precipitation results in a mean change, but it also affects the variance. A ratio greater than one increases the variance, while a ratio less than one decreases it. However, the frequency of precipitation and the sequence of dry and wet days are not altered. This method has been used for decades to downscale climate change information from GCMs (Rosenzweig 1985), as well as to further downscale and bias-corrected information from regional climate models (RCMs, e.g., Mearns et al. 2003). This method remains in use today, for example, the delta method was employed to develop data sets for the New York City Adaptation Planning efforts (Horton et al. 2010).

Bias Correction Spatial Disaggregation (BCSD) and Related Methods

Methods of downscaling that involve interpolation use some form of disaggregation to transform coarse resolution climate model data to higher resolutions. One of the best-known methods is the Bias Correction Spatial Disaggregation Method (BCSD) (Wood et al. 2002, 2004). This method is considerably more sophisticated than the delta method described above. We separate it from the statistical downscaling discussion that follows this section, since it does not involve production of "new" information about the future climate, but rather redistributes (interpolates) the information contained in the coarser resolution model simulations. BCSD involves: (climate) trend removal, bias correction via mapping between empirical cumulative distribution functions of observed and modeled variables, and spatial disaggregation by interpolation of the bias-corrected anomalies and imposition of finer scale climatological means. An important feature is the method of bias correction,

which results in a quantile-quantile correction, so that the entire distribution is corrected, not only the mean (as is the case with the "delta" approach described above). The method was specifically developed to aid in the determination of climate impacts on hydrology and water resources. As a service to the impacts community, the entire World Climate Research Programme's (WCRP's) Coupled Model Intercomparison Project phase 3 (CMIP3) data set (at least one realization for each GCM run), which was developed for the IPCC 2007 Report (Solomon et al. 2007), was downscaled to a 1/8 degree resolution using this method (temperature and precipitation) (Maurer et al. 2007). This data set has been used widely by impacts researchers in hydrology (Wood et al. 2004; Payne et al. 2004) and other impacts areas (e.g., ecology (Lawler et al. 2009)). Figure 8.1 provides a sample of results from the CMIP 3 data set for change in winter precipitation. Note that the BCSD change in precipitation

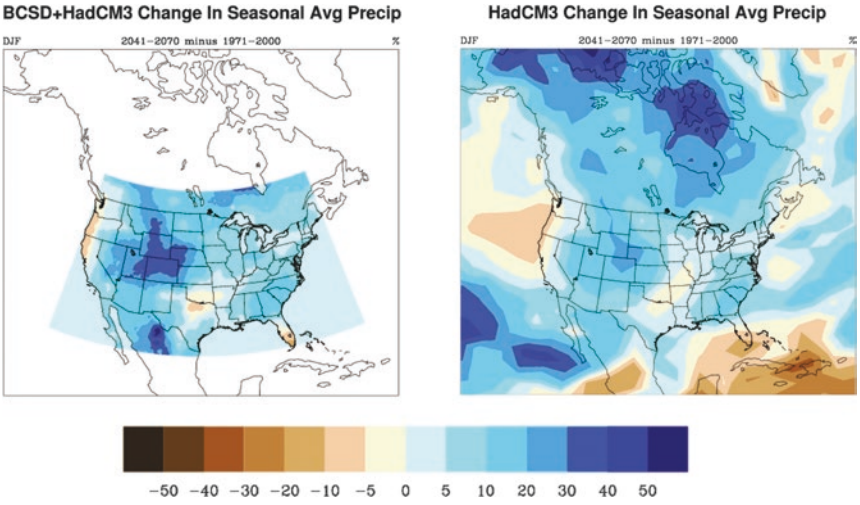

Fig. 8.1 Change (%) in winter precipitation mid-twenty-first century (2041–2070) vs. late-twentieth century (1971–2000) from simulations with the HadCM3 AOGCM (**a**) (*left*) downscaled using the BCSD method (1/8° resolution) and (**b**) (*right*) in the original HadCM3 model which was run at a spatial resolution of 2.5° latitude by 3.5° longitude (Graphics by Seth McGinnis and Joshua Thompson, NCAR, using data acquired from: https://esgcet.llnl.gov:8443/index.jsp for raw HadCM3 data; http://gdo-dcp.ucllnl.org/downscaled_cmip3_projections/dcpInterface.html for BCSD data)

bears some resemblance to that of the coarser resolution global climate model, the U.K. Met Office Hadley Centre Climate Model, version 3 (HadCM3), but there are distinct differences, for example in the central Rocky Mountain area where the increases in precipitation in BCSD are much larger than those in the HadCM3. This approach primarily has been used to downscale monthly mean values of temperature and precipitation, and of course, numerous impacts models require daily data.

Other methods have been developed to bias-correct and generate high-resolution daily data from coarser spatial resolution GCM output. One main approach is the Bias Correction Climate Analogue method (BCCA) (Maurer et al. 2010). While this method can also be considered in the category of statistical downscaling (discussed in Sect. 8.3) we include it here since it has some similarity with the BCSD approach and has been compared to it. BCCA relies on a fundamentally different concept—constructed analogues (CA)—for the downscaling part. This approach relates model-simulated variables (e.g., anomalies of daily temperature and precipitation) to observed large-scale patterns (of the same daily variables). BCCA uses a bias correction approach very similar to that of BCSD, but the quantile mapping (used for bias correction) is applied to the daily data within a particular month. The climate analogue approach relies on a library of coarse resolution and high-resolution observed climate anomaly patterns (of temperature and precipitation). A subset of observed large-scale pattern anomalies is selected, and then the linear combination of those patterns that best match the given (target) pattern is determined. The next step is the derivation of the high-resolution pattern by applying the linear fit developed from the subset of most suitable, coarse resolution historical patterns. The regression coefficients derived for each coarse resolution pattern in the diagnosis step are applied directly to the corresponding fine-resolution weather patterns for the same days (Maurer and Hidalgo 2008; Maurer et al. 2010).

Maurer et al. (2010) compared three downscaling methods including BCCA and BCSD and found that the BCCA method was somewhat better when used to generate important hydrologic variables using a hydrologic model for a number of stations in California. Gutmann et al. (2013) compared five different downscaling methods: BCSD on a daily and

monthly scale, two variants of BCCA, and an asynchronous regression technique for precipitation at three different temporal and multiple spatial scales over the contiguous US. Results were mixed, depending on the metric and scale of comparison, but BCAA tended to perform most poorly.

8.3 Empirical/Statistical Downscaling (ESD)

Methods

Empirical/statistical downscaling (ESD) is the process of developing mathematical links between the state (value) of some variable(s) representing large spatial scales and the state (value) of some variable(s) representing a much smaller spatial (local) scale. ESD thus assumes an implicit and fundamental dynamical link between the two scales (e.g., that the air temperature, wind speed, or occurrence or amount of precipitation at a specific location is determined, at least in part, by processes manifest at a scale that are well-described by global or regional climate models) (Benestad et al. 2008; Maraun et al. 2010). ESD may thus be used whenever the specific application requires local-scale climate projections, provided suitable observational data are available to develop the statistical models.

ESD techniques typically fall into one or more of the following three categories:

- Transfer functions. Typically these approaches involve development and application of linear or non-linear equations that link the local variable(s) (predictand(s)) (e.g., daily or monthly temperature or precipitation) of interest to large-scale predictors (e.g., 500-mb geopotential heights) drawn from output from AOGCMs or regional climate models (Li and Sailor 2000; Schoof and Pryor 2001). Some approaches within this class are referred to as probabilistic since they focus on simulating descriptors of the probability distribution of either or both of the predictors and predictands rather than a time series thereof (Pryor et al. 2006).

- Weather typing (Schoof and Pryor 2001). Typically these approaches involve sub-sampling of the local variable of interest by the prevailing synoptic-scale conditions, as categorized into defined classes, often based on the atmospheric circulation. They have also been adopted for hybrid downscaling applications wherein dynamical and statistical downscaling are combined (see below) (Wetterhall et al. 2012).
- Stochastic weather generators (SWGs) are models that produce synthetic time series of local climate variables with empirically determined statistical properties (i.e., parameters). Application of these approaches is often based on perturbation of the parameters according to climate changes projected by climate models (see Katz et al. 2003; Semenov et al. 1998; Wilks 2012).

ESD and dynamical downscaling can be applied independently or in combination (Manning et al. 2009). Hybrid ESD approaches that cross the boundaries implied by these categories are increasingly being applied (Li et al. 2012; Schoof et al. 2007; Vrac et al. 2007; Wetterhall et al. 2012), and new techniques are being developed coupling weather typing with signals from distant teleconnection indices (Canon et al. 2011).

Skill and Uncertainty

Implicit in the fundamental foundations and assumptions of ESD techniques are the following limitations:

(i) The statistical models are based on historical data. Application of ESD relies upon stationarity in the relationships codified within transfer functions, but there is no guarantee of stationarity in relationships between the local-scale variable and the large-scale forcing.
(ii) They need a robust and large training sample for use in model calibration.
(iii) There is high uncertainty in extrapolation of values outside the range experienced in the calibration data sets.
(iv) There is a tendency for many techniques to suppress the variance in the predictand.

(v) Many ESD techniques do not or cannot account for changes in temporal autocorrelation.
(vi) Many ESD techniques have greatest validity where the predictands and predictors exhibit (approximately) Gaussian distributions (*Semenov* 2008). In the case of highly nonlinear distributions it is sometimes desirable to transform the variable to conform more closely to a normal distribution.
(vii) ESD cannot "correct" for "aphysical" realizations from the climate model from which the predictors are drawn. While AOGCMs exhibit skill at larger spatial scales, their treatment of the synoptic-scale climatology remains imperfect and highly variable from model to model (Sheridan and Lee 2010).
(viii) The predictors must: significantly contribute to variability in the predictand, should represent important processes associated with climate evolution, and be "skillfully" simulated by the driving climate model.

It should be noted that while uncertainty/errors in climate projections are not necessarily propagated or amplified through impact analyses, the downscaling process is identified as an important source of uncertainty in hydrological impact studies (Stoll et al. 2011). Thus, there is continued need for evaluation and improvement of different downscaling methods and for verification/evaluation relative to independent data.

The "skill" and uncertainty of ESD show a high degree of sensitivity to the ESD model applied, the variable under consideration, a priori assumptions applied, the climate of the region under study, and the degree of temporal averaging (Dibike et al. 2008; Fowler et al. 2007; Khan et al. 2006; Maurer and Hidalgo 2008; Qian et al. 2008; Schoof and Pryor 2008; Wang and Zhang 2008; Wilby and Wigley 2000). Skill is typically demonstrated by withholding part of the historical training data from the construction of the ESD model, and then applying the model to that sub-set (Harpham and Wilby 2005; Schoof et al. 2010). This type of assessment provides useful information regarding the stability of the model, but does not fully address issues pertaining to the ability of the model to downscale conditions not (or under-) sampled in the training period and unless conducted with a very wide time span the

results cannot be offered as evidence that the downscaling model will necessarily provide robust results under an evolving climate.

Uncertainty in climate projections derived from ESD primarily originates from one of the following sources:

1. Boundary (or predictor) uncertainty due to the architecture and/or resolution of the climate model.
2. Initial conditions. Each climate model simulation represents only one realization of possible climate states.
3. Sampling uncertainty from use of short temporal windows to consider future conditions and integration over a finite number of years presuming that transient simulation output is not available.
4. The specific emission scenario or representative concentration pathway used and thus degree of climate forcing applied.
5. The specific ESD model applied and assumptions implicit thereto.

One probabilistic ESD of wind climates over northern Europe evaluated the relative roles of 1–4 and found that the AOGCM used to provide the downscaling predictors dominated uncertainty in downscaled 90th percentile wind speed for the end of the twenty-first century. Variations in initial conditions, climate forcing (as manifest in the IPCC's Special Report on Emissions Scenarios (SRES)), and stochastic influences within individual AOGCM simulations made lesser (but non-negligible) contributions to uncertainty in these projections (Pryor and Schoof 2010). A further study of uncertainty sources in ESD for hydrological impacts in Quebec considered uncertainty sources 1, 4, and 5, and found that when used to simulate discharge for a single river basin, the range of realizations from six ESD techniques was approximately comparable to the spread of realizations derived from seven AOGCMs and three emission scenarios (Chen et al. 2011a).

Results from Applications of ESD over North America

In the following, we describe the results of recent applications of ESD to development of climate projections over North America. Relative to

Europe, comparatively few studies have applied ESD over North America; nevertheless, due to space constraints this summary is not intended to be fully comprehensive of the array of prior research but rather has been selected to focus principally on downscaling of the CMIP3 AOGCM suite and convey the range of approach applied and the consistency (or otherwise) of the inferences drawn.

Temperature

Based on simple ESD downscaling of mean temperatures across the western US from 18 CMIP3 AOGCMs (under the A1B SRES) (Gutzler and Robbins 2011), resolved temperatures in 2076–2100 will exceed temperatures in 1976–2000 by >2 °C over the entire western US, and by > 3 °C over the majority of the region. The magnitude of warming is consistent with results from application of ESD to stations across California to derive a number of thermal metrics and precipitation variables based on output from HadCM3 and National Center for Atmospheric Research/Department of Energy Parallel Climate Model (PCM) AOGCMs for the B1 and A1FI emissions scenarios (Hayhoe et al. 2004). The results of that study indicated spatially averaged increases in summertime temperatures of 2.2–8.3 °C in 2070–2099 relative to 1961–1990, where the majority of the uncertainty was due to differences in the two emission scenarios. When projections of air temperatures for California were based on bias correction and spatial mapping applied to CMIP3 AOGCMs for three emissions scenarios and linked to electricity demand, the changes in thermal regimes increased annual electricity demand in 2077–2099 relative to 1961–1990 by 2.9–17.8% (depending on the AOGCM and SRES scenario used) and increased peak demand by 4.2–19.8% (Franco and Sanstad 2008).

ESD for summertime air temperature projections over the eastern US based on regression techniques combined with empirical orthogonal functions applied to output from the NASA Goddard Institute of Space Studies (GISS) AOGCM run with the A2 SRES emissions scenario indicated warming of approximately 2 °C by the 2080s relative to the 1990s, which is considerably less than implied by direct output from the

AOGCM (Spak et al. 2007). Indeed, downscaling of summertime temperatures using the NASA GISS AOGCM via both an RCM and statistical approaches indicated that the two methods projected similar regional mean warming over the period 2000–2087, but developed different spatial patterns of temperature across the region. For the 2050s the RCM MM5 showed higher temperatures, but in the 2080s the statistical approach resolved regions of higher magnitude warming (Spak et al. 2007). A hybrid ESD approach in which seasonal variations of the mean and standard deviation of daily maximum and minimum temperatures are derived using transfer functions applied to output from HadCM3 and the Canadian Centre for Climate Modelling and Analysis Coupled Global Climate Model (CGCM2) for the A2 emission scenario, which are then used as inputs to a stochastic weather generator (SWG), was used to produce time series of daily maximum (Tmax) and minimum (Tmin) temperatures at stations across the Midwest. Downscaled temperature projections for 2020–2029 indicate increases that range (across stations) up to 1.7 °C in Tmax and up to 1.5 °C in Tmin relative to 1990–2001. Comparable scenarios for 2050–2059 indicate increases in these two parameters of 1.4–2.4 °C and 0.8–2.2 °C, respectively. The major source of uncertainty in this analysis was traced to differences in the predictors from the two AOGCMs, which led to higher variability in downscaled Tmax from the HadCM3 output (Schoof et al. 2007). That study also demonstrated the superior skill in downscaling of Tmax and Tmin using SWG relative to multiple linear regression. Projections for 30-year moving windows of 10th to 90th percentile winter and summer temperatures from a statistically derived large ensemble suggest even greater amplification of the upper quartile of the temperature distribution. Under the A2 high emission scenario, the wintertime 90th percentile temperatures over the North Great Plains and upper Midwest are projected to exceed those in 1971–2000 by >5 °C in 2041–2070, and that summertime 90th percentile temperatures will be higher by 6 °C over most of the continental US (Li et al. 2012). Further discussion of downscaled extreme temperatures is given below in Sect. 8.3.3.3.

One of the clearest signals of climate trends in the historical record is the expansion of the growing season across much of the contiguous US (Kunkel et al. 2004). One ESD downscaling study of frost-free season

using transfer functions developed using output of 700 hPa temperature and specific humidity from eight of the CMIP-3 AOGCMs and Tmin and Tmax at 53 stations across the Midwest found evidence for continuation of the historical tendencies. The ESD scenarios indicated increases of approximately two weeks (15.8 days) in the duration of the frost-free period by 2046–2065 and by almost one month by 2081–2100 (both relative to 1961–1990) (Schoof 2009). This is consistent with the average increase in duration of the growing season for the Midwest by 2041–2062 of approximately three weeks derived based on an ensemble of the North American Regional Climate Change Assessment Program (NARCCAP) models (Pryor et al. 2013).

ESD-derived climate change projections over Mexico were derived by Montero-Martínez and Pérez-Lopez (2008) and Magaña et al. (2012). Around 20 CMIP-3 AOGCMs and four emission scenarios were downscaled, and indicated an ensemble mean warming of around 3.5 °C in northwest Mexico and about 3 °C in northeast Mexico by the end of the twenty-first century under the A2 emission scenario.

Precipitation

Downscaling precipitation regimes is considerably more challenging than thermal regimes due to the greater spatial heterogeneity in precipitation, and the need to accurately simulate two key components—the probability of any precipitation and the amount of precipitation on a "wet" day. For impact studies (and particularly water availability in some Western watersheds), an additional key consideration is the phase of the hydrometeors (Hay et al. 2011; Shepherd et al. 2010). Despite these challenges, most studies indicate a high degree of value-added in both ESD and dynamical downscaling of precipitation variables compared to output from parent AOGCMs (Maraun et al. 2010).

Schoof et al. (2010) used ESD to analyze possible changes in the frequency and intensity of precipitation at 963 stations across the contiguous US based on predictors derived from output of 10 CMIP3 AOGCMs driven by the A2 SRES emission scenario. The ESD method used first-order Markov chains to simulate precipitation occurrence, the gamma

probability distribution to quantify wet-day amount, and regionally specific large-scale predictors drawn from a suite that included: specific humidity, temperature and flow components at 700 and 500 hPa, and sea-level pressure. The results indicate that stations that are characterized by projected increases in seasonal total precipitation typically exhibit increased precipitation intensities. Conversely, those stations for which the future scenarios indicate negative changes in precipitation totals typically have projections characterized by large changes in small precipitation intensities with relatively little change in large events. This suggests that intense precipitation events are likely to either maintain their current frequency or increase in frequency regardless of the sign of changes in total precipitation. This tendency towards increased magnitude of high intensity events even in regions with declining overall precipitation receipt is consistent with historical tendencies in precipitation regimes (Groisman et al. 1999; Pryor et al. 2009). The projections developed by Schoof et al. (2010) from each individual AOGCM and each station exhibit a high degree of variability, but the ensemble average projections synthesized across all AOGCMs and all stations within six regions (Figs. 8.2 and 8.3) indicate:

(a) The largest total precipitation increases during the cold season (defined as NDJFM) are projected to occur in the Northwest and Northeast regions. These increases derive largely from projected increases in precipitation intensity, although the Northeast region is also projected to experience moderate increases in cold season precipitation occurrence. Large decreases (with an area average magnitude of – 15% for mid-twenty-first century) in cold season precipitation are projected for the Southwest, due to a large decrease in precipitation occurrence, which more than offsets projected moderate increases in wet-day precipitation intensity. Cold season projections for the Northern Plains indicate moderate precipitation decreases due to reductions in precipitation frequency.
(b) For the majority of the contiguous US, drier warm season (defined as MJJAS) conditions are projected, due largely to decreases in precipitation frequency (of up to 30% by mid-century in the Northwest, Southern Plains, and Southeast). Warm season total precipitation is

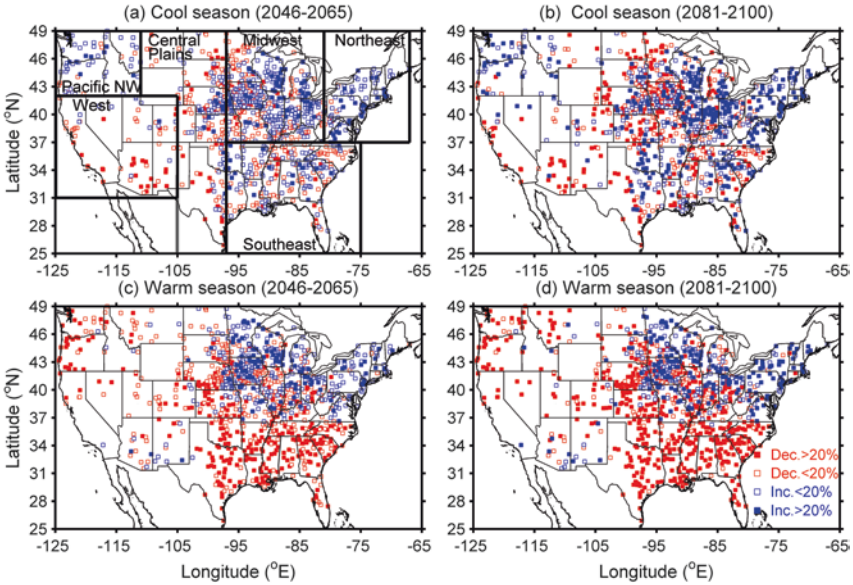

Fig. 8.2 Change in total precipitation (expressed in %) at 936 stations in (**a** and **b**) cold season (NDJFM) and (**c** and **d**) warm season (MJJAS) and for 2046–2065 or 2081–2100 relative to 1961–2000 derived from statistical downscaling of 10 AOGCMs (BCCR-BCM2, CCCMA-CGCM3, CNRM-CM3, CSIRO-MK3, GFDL-CM2, GISS-Model E-R, IPSL-CM4, MIUB-ECHO, MPI-ECHAM5, and MRI-CGCM2) (Schoof et al. 2010)

projected to decline by up to 40% by 2081–2100 in the Southeast and southern Plain states. Only the Northeast and Midwest are projected to experience area averaged increases in total warm season precipitation. In the Midwest, this is principally due to an increase in the magnitude of intense events. In the Northeast region, increases in large precipitation events are coupled with increases in precipitation occurrence.

While the finding of increased cold season precipitation over the Northeast is in accord with a prior regional analysis, the changes in the warm season are in contrast to an earlier analysis by (Hayhoe et al. 2007). In the work by Hayhoe et al. (2007), the ESD applied involved mapping of probability density functions for monthly and daily precipitation and

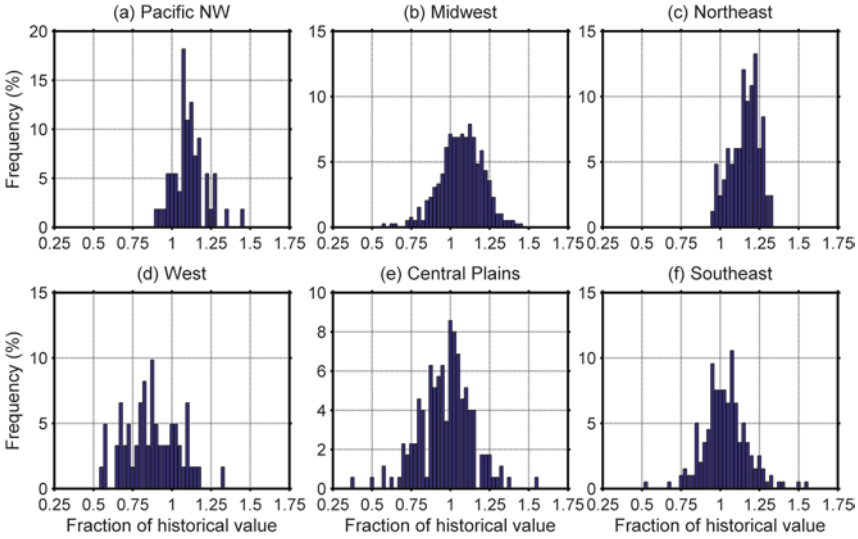

Fig. 8.3 Regional histograms for the ensemble mean difference in seasonal precipitation 2046–2065 v 1961–2000 at each station based on downscaling of 10 AOGCMs (BCCR-BCM2, CCCMA-CGCM3, CNRM-CM3, CSIRO-MK3, GFDL-CM2, GISS-Model E-R, IPSL-CM4, MIUB-ECHO, MPI-ECHAM5, and MRI-CGCM2) (Schoof et al. 2010). The *upper panels* show the results for the warm season (MJJAS), and the *lower panel* shows results for the cool season (NDJAM). The frequency denotes the percentage of stations in a given region that show a ratio of a given magnitude. If the Fraction of the historical value is 1 the historical and future periods have equal precipitation totals

temperature onto gridded historical observations. The results indicated winter (DJF) precipitation increases of 6–16% (where the range represents the variation between three different SRES emissions scenarios) by 2035–2064 relative to 1961–1990, and little or no change in summer (JJA) precipitation totals. In a separate ESD analysis for the Great Lakes region (specifically Michigan and Illinois) which used the same approach as that by Hayhoe et al. (2007) and an asynchronous quantile regression methodology applied to three AOGCMs from the CMIP3 archive, Hayhoe et al. (2010) found that annual precipitation was within a few percent of historical values, but was generally higher at the end of the twenty-first century (by up to 20%) relative to the end of the twentieth century. The cold season results mostly indicated increased precipitation

consistent with the findings of Hayhoe et al. (2007) and Schoof et al. (2010), while projections for the summer typically indicated zero or small magnitude declines in precipitation accumulation. The high spatial variability in the response in warm season precipitation evident for the Great Plains and Midwest as shown in Figs. 8.2 and 8.3 is consistent, at least in part, with other downscaling analyses that have indicated enhanced precipitation during the spring transition months, coupled with drying of the summer proper (Patricola and Cook 2013; Pryor et al. 2013) (see further discussion below).

The projected increases in precipitation in the Pacific Northwest are also consistent with simple downscaling applied to 10 CMIP3 AOGCMs (Salathé 2006). Salathé (2006) suggested that the projected increase in precipitation might be causally linked to simulated changes in the large-scale storm track increasing orographic enhancement of precipitation. Gutzler and Robbins (2011) used a simple ESD based on projected linear trends in temperature or precipitation from 18 CMIP3 AOGCMs (A1B SRES) superimposed onto the interannual variability as observed during the twentieth century to examine scenarios of possible drought statistics in the western US. The results indicated declines in precipitation totals (2076–2100 relative to 1976–2000) over much of California, Arizona, southern Nevada, and Texas, and increased precipitation projections north of those states. These findings are consistent with the scenarios developed by Schoof et al. (2010) in terms of sign of change but are of lesser magnitude. While the changes in precipitation receipt derived by Gutzler and Robbins (2011) are relatively modest, when the temperature and precipitation projections are used to derive estimates of future Palmer drought severity index scenarios they found a marked increase "in the severity and duration of twenty-first century drought (defined in terms of a twentieth century baseline), and the spatial scale of future droughts expands to cover much of the West" (Gutzler and Robbins 2011).

Increased drought probability and intensity was also projected for Mexico in Montero-Martínez and Pérez-Lopez (2008) and Magaña et al. (2012), where annual rainfall is projected to decrease by 10–20% by 2040–2069 under the A1B and A2 emission scenarios. Most of the downscaled models agree on the sign of change, but uncertainty is high because of the strong dependence on tropical cyclones as a water source

for much of the region. However, the changes in temperature, precipitation, and the variability of precipitation in these projections place northern Mexico in a state of semi-permanent moderate meteorological drought after the 2050s given the A2 scenario.

Comparatively few analyses have focused on development of precipitation scenarios for the Hawaiian Islands. One circulation-based ESD applied to six AOGCMs drawn from the CMIP3 archive indicated considerable divergence in projections based in part on the simulation of the trade winds by the parent AOGCM. Nevertheless, the study concluded "the most likely scenario for Hawaii is a 5%–10% reduction of the wet season precipitation and a 5% increase during the dry season" by the end of the twenty-first century (Timm and Diaz 2009).

Extreme Events

The economies and ecosystems of North America tend to be much more sensitive to extremes than to average conditions, and thus the impacts of climate change are likely to be disproportionately dictated by changes in the magnitude, frequency or characteristics of rare (but high magnitude) events. Accordingly, several ESD techniques (e.g., SWG and probabilistic approaches) have been applied to analysis of possible changes in extreme conditions (Pryor and Barthelmie 2010; Qian et al. 2008). In one example, projections of annual and growing season climate extremes were derived using an SWG and output from four CMIP3 generation AOGCMs forced with the A2 SRES for sites across Canada (Qian et al. 2010). All AOGCMs indicated a warmer future in both direct output and SWG derived local scenarios, and downscaled 50-year return period temperatures increased by up to 4 °C in 2041–2070 relative to 1961–1990. Consistent with other research on possible changes in precipitation regimes over Canada (Choi et al. 2009), potential changes in the 50-year return period daily precipitation for the mid-twenty-first century downscaled from each AOGCM were almost uniformly positive and increased by up to 25% from values during 1961–1990. Analysis of the downscaling results versus use of direct AOGCM output indicated (i) application of the SWG reduced bias in extreme metrics during the baseline period,

[*Harmon R. Holcomb*] climate change signals in the SWG localized projections differed markedly from the direct AOGCM output, and [*Harmon R. Holcomb*] uncertainty in future climate projected from the four different AOGCMs is a major contributor to overall analysis uncertainty (Qian et al. 2010).

Relatively few downscaling analyses have explicitly addressed heat waves, but one case study for Chicago found seven-day periods with temperatures in excess of 32.2 °C had a return period of two years in the historical period but an occurrence rate of over 1.8 in any year by 2070–2099 (Hayhoe et al. 2010). One ESD study based on output from HadCM3 examined the occurrence of heat waves in Mexicali, Mexico, using a temperature threshold of 44 °C and the statistical downscaling model (SDSM, which is a hybrid of regression analysis and SWG). The results indicate that the frequency of heat days, which has increased by over a factor of two during the last four decades, is projected to increase by 2.1–2.4, 3.4–3.6, and 4.0–5.1 times by 2020s, 2050s, and 2080s, respectively relative to the average for 1961–1990 based on the B2 and A2 SRES (Garcia Cueto et al. 2010). Other ESD analyses of heat-wave intensity have included the additional influence on apparent temperature of humidity, and have indicated apparent temperature and hence heat stress in the Midwest increased across all SRES scenarios considered, with the 90th percentile apparent temperature increasing by between 3 °C and 6 °C between 1961–1990 and 2081–2100 (Schoof 2012). This tendency toward intensification of thermal extremes is also manifest in the NARCCAP RCM simulations. For example, the number of days in the Chicago region each year with temperatures in excess of 32.2 °C (90 °F) is doubled by the mid-century based on an eight-member ensemble mean (Pryor et al. 2013).

A comprehensive ESD analysis of wintertime (DJFM) 20-year return period precipitation amounts has been performed from 4128 stations across North America and output from the Canadian Centre for Climate Modeling and Analysis version 3.1 AOGCM in combination with a downscaling method that employs circulation-based analysis and application of Generalized Extreme Value distributions. The results indicate that the current 20-year return period daily precipitation amount will be

observed with higher frequency in the future (2050–2099) across almost the entire study domain (with the exception of northern Alberta and southern Mexico), with largest magnitude increases in southern and central US (Wang and Zhang 2008). This is again consistent with other scenarios of intense precipitation that are suggestive of a continuation of tendencies toward intensification of extreme events that have been found in the historical record (see Chap. 2).

8.4 Dynamical Downscaling

Methods

Dynamical downscaling refers to the production of high-resolution climate information using models that are dynamically and physically based. In this way, they are similar to coupled atmosphere-ocean general circulation models (AOGCMs) and limited area models used for weather research and forecasting. Unlike AOGCMs, which have, in recent past, produced simulations with a spatial resolution of 100 km or more, methods for dynamical downscaling often produce simulations at 10–50 km or less. Differences like these in resolution are illustrated through terrain height fields in the western US in Figs. 8.4 and 8.5.

Four methods of dynamical downscaling exist: nested regional climate modeling, stretched grid global modeling, high-resolution simulation using atmosphere-only GCMs (AGCMs), and coarse resolution AOGCM modeling with high-resolution orography. Given its current popularity and establishment as a useful tool, our technique overview and climate-change results summary will focus mainly on nested regional climate modeling. Also, because a full methodological description is outside the scope of this book, only a brief overview of the modeling techniques will be given. The reader is referred to Rummukainen (2010) for a more in-depth, but general, overview and to Giorgi et al. (2001), Laprise (2008), and Warner (2011) for more technical descriptions, overviews, and reviews of common dynamical downscaling sensitivities.

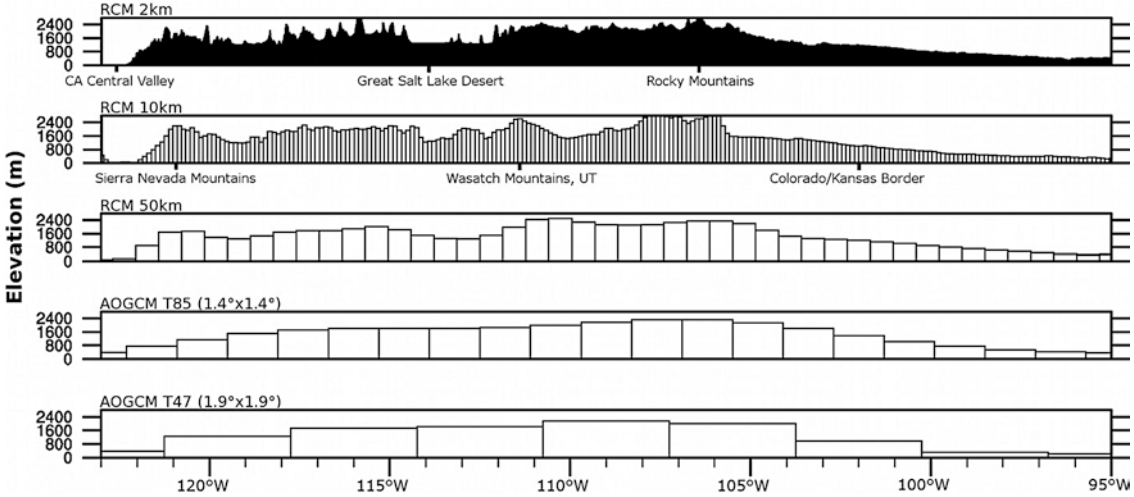

Fig. 8.4 Transect of terrain height (m) along, approximately, 40 °N from 95 °W westward to the California Central Valley in the regional climate models (RCMs), at five different resolutions. A few geographic landmarks are labeled for reference. Longitude labels at the *bottom* are valid for the AOGCMs only, as the transect paths in the RCMs vary from those in the AOGCMs due to differences in model map projections and model grid cell sizes. Paths of the transects from the west coast to about 100 °W are given in the *lower right panel* of Fig. 8.5

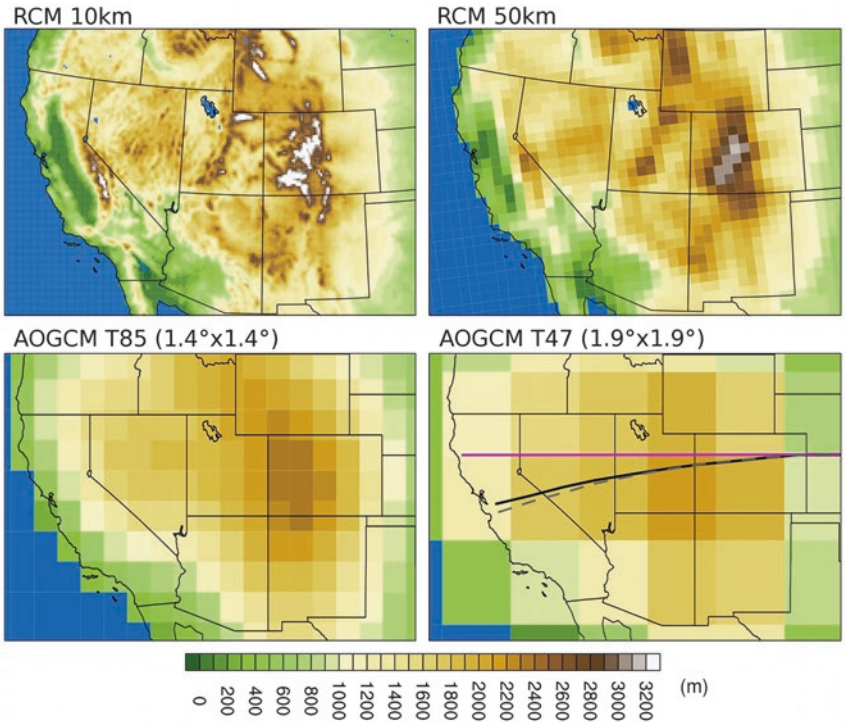

Fig. 8.5 Terrain height (m) for model grid cells at four different horizontal resolutions. Paths for the transects shown in Fig. 8.4 are given in the *lower right panel*. AOGCM transect paths are represented by the *pink line*, while the 2-km and 10-km RCM paths are given by the *solid black line*, and the 50-km RCM path is represented by the *dashed gray line*. Differences in the paths are a result of differences in map projections and grid cell sizes

Nested Regional Climate Modeling

Nested regional climate modeling refers to the production of high-resolution climate information using limited area models (LAMs)/ RCMs over any given location of interest. RCMs require lateral and lower boundary condition forcing from some source at a time frequency of around 3–6 hours. The source is often referred to as the driver, parent, or forcing model or data set. Variables used from the parent include temperature, moisture, winds, pressure/geopotential height, sea-surface temperature (SST), sea ice, and soil moisture and temperature.

Simulations are typically started in advance of the period when information is desired. This is to allow for spin-up of the variables inside of the RCM domain, that is, to allow them to obtain equilibrium after initialization. While atmospheric fields spin-up within a day or two, fields such as deep soil moisture may require a year or more to reach equilibrium (Christensen et al. 2001; Cosgrove et al. 2003; de Elía et al 2002). Spin-up periods should not be used in climate analysis.

To start, "perfect" boundary conditions are often used to drive an RCM. They are derived from an observational analysis or reanalysis and allow for the determination of any systematic biases in an RCM. This step also allows an RCM to be compared directly to observations during the simulation period, as opposed to only comparing it against long-term climate statistics, as any given day in an RCM should match a given day in reality with this experimental setup.

For climate change projections, boundary conditions are often derived from AOGCM simulations of baseline/historical climate and future climate. The climate change projections are taken as the difference between these two simulations. AOGCM-forced RCM simulations are subject to inheriting biases present in the AOGCMs (e.g., Noguer et al. 1998); this bias adds to an RCM's systematic bias, and users of RCM projections should remain conscious of this. In some cases, reanalyses have been modified to reflect potential future conditions (e.g., Patricola and Cook 2010; Rasmussen et al. 2011). This does help reduce inherited GCM bias, assuming bias is linear and constant current-to-future, but there is no best practice method established for this technique yet, and it is not the norm—most RCM studies use AOGCM output for boundary conditions directly.

Multiple nesting may be used in an RCM (i.e., a nest with an even finer resolution may occur within a limited area domain). This may be used to avoid large jumps in resolution between the parent and the RCM, to create a larger, main RCM domain to avoid placing boundaries of the desired nest in problematic locations with little extra computational cost, and/or to simply obtain higher resolution information over a specific region. This technique is used, for example, in Hall et al. (2012), where an 18-km parent domain and then a 6-km nest with an interior 2-km nest are used to better resolve the region encompassing Los Angeles County.

Generally, information calculated in an RCM is not passed back to the parent. This is referred to as one-way nesting. Two-way nesting, where there is feedback between an RCM and its driver is not common yet and is more complicated. Examples of two-way nesting are available in Lorenz and Jacob (2005), Inatsu and Kimoto (2009), and Chen et al. (2010).

In some studies, nudging techniques are applied in an RCM to keep the large-scale flow inside the RCM domain from diverging from the solution of its parent by "nudging" it back every few hours. Divergence in the large-scale fields is not always desirable, but it is not uncommon in large domains. While nudging can reduce drift and produce better simulations in some cases (especially when driven by a reanalysis, for example, Lo et al. 2008), it can cause damping of precipitation extremes, other small-scale features, and upscale feedbacks to the larger scale (Radu et al. 2008; Alexandru et al. 2009; Rummukainen 2010; Cha et al. 2011). Damping of extremes may not always be detrimental, however, and may produce a more realistic outcome (e.g., Otte et al. 2012).

It behooves the users of RCM information to know which physical processes are included in the models they choose. For instance, not all regional models contain lake models, or in some, using one is optional. It is currently common for RCMs to set surface temperatures over resolved lakes using interpolated values from the nearest ocean points. This practice allows for lake surface temperatures that are more realistic than if the temperature was set using surrounding land points, but this practice may negatively impact the simulated climate near large lakes. Likewise, not all RCMs include the formation of lake and sea ice, or they contain crude representations for it. Prudence is necessary, therefore, when using information from a given RCM in a specific locality.

Some of the limited area atmospheric modeling systems that have been adapted or developed for regional climate use and have been used over North America are listed below. Extensive references exist for each, but only a few relevant references are provided.

- The Canadian Regional Climate Model (CRCM, Caya and Laprise 1999; Laprise et al. 2003; Laprise 2008)
- The NCEP Eta Model (Janic 1994; Xue et al. 2007)
- The fifth-generation Pennsylvania State University-NCAR Mesoscale Model (MM5, Grell et al. 1993)

- Providing Regional Climates for Impacts Studies (PRECIS)/HadRM, the Met Office Hadley Centre Regional Climate Model (Jones et al. 2004)
- The International Centre for Theoretical Physics (ICTP) Regional Climate Model (RegCM, Giorgi et al. 1993a, b; Pal et al. 2007)
- The Regional Spectral Model (RSM), originally developed at the National Centers for Environmental Prediction (Juang et al. 1997)
- The Regional Atmospheric Modeling System (RAMS, Pielke et al. 1992; Cotton et al. 2003; Miguez-Macho et al. 2005)
- The Weather Research and Forecasting model (WRF, Skamarock et al. 2005)

High-Resolution AGCMs

In high-resolution AGCMs (HR-AGCM), simulations are completed globally, not over a limited domain, with the atmospheric model component of an AOGCM (e.g., Cubasch et al. 1995; May and Roeckner 2001; Duffy et al. 2003; Govindasamy et al. 2003; Déqué et al. 2005; Wehner et al. 2010). As global models, lateral boundary conditions (LBCs) are not necessary; these models are forced by surface boundary conditions, that is, SSTs and sea ice from reanalysis or AOGCMs.

HR-AGCM simulations are often referred to as "time-slice" simulations since only part of the coarser resolution fully transient (i.e., continuous long-term) simulation from the parent AOGCM is often downscaled. Completing time-slice simulations, as opposed to full transient simulations, is often necessary as this type of dynamical downscaling is more computationally expensive given the global domain. HR-AGCMs do have an advantage over RCMs in not having LBC issues. Plus, HR-AGCMs allow for feedback of resolved smaller-scale atmospheric processes from one region of the world to another.

Stretched Grid AGCMs

Stretched grid AGCMs (SG-AGCM), or variable resolution AGCMs, are similar to HR-AGCMs in that they both run globally, but SG-AGCM

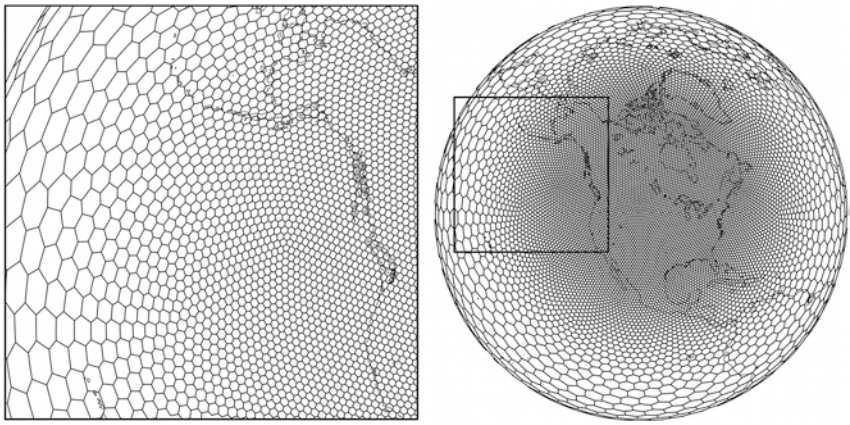

Fig. 8.6 An MPAS Voronoi hexagonal mesh centered over North America, configured with 10,242 grid cells with an 85-km horizontal resolution in the fine-mesh region and a 650-km resolution in the coarsest region. (Fig. 10 from Skamarock et al. 2012)

use high-resolution over chosen areas of interest only, and then transition to a coarser grid over the rest of the globe (e.g., Côté et al. 1998; Fox-Rabinovitz et al. 2006; McGregor and Dix 2008; Skamarock et al. 2010, 2011). In this way, they are less computationally expensive than HR-AGCMs. They have the same benefits as HR-AGCMs in terms of the lack of LBC issues and allowance of global feedback, but problems could develop in the high-to-coarse resolution transition areas, particularly if chosen parameterizations do not work well across multiple resolutions. An example of a stretched grid, illustrating one configuration of the Voronoi hexagonal mesh used by the Model for Predication Across Scales (MPAS; Skamarock et al. 2012), is given in Fig. 8.6.

High-Resolution Orography Within a Coarse Resolution AOGCM

Parameterizing high-resolution orographic forcing within a coarse resolution model is another method used to obtain improved regional detail in climate simulations. In this method, the impacts of sub-grid scale oro-

graphic precipitation, vegetation, and lakes, for instance, can be calculated on sub-grid scale elevation bands to provide enhanced regional detail in chosen variables (Leung and Ghan 1998; Ghan and Shippert 2006, Ghan et al. 2006). This methodology is not just applied in coarse AOGCMs, but may also be used to provide further detail in RCMs (Leung and Ghan 1999a, b; Lei and Yaocun 2007).

Skill and Uncertainties

Skill in reproducing historical climate is often used to infer which models might perform best in simulating future climate. However, while accurately portraying historical climate might give one more confidence in the model, it does not necessarily mean that the model will have skill in projecting future climate. How to differentiate models by skill to combine their projections is a current topic of debate (Knutti et al. 2010). However, one is likely to give little credit to a model that cannot produce a realistic simulation of observed climate.

As a result, there are many examples demonstrating the skill of dynamically downscaled simulations in reproducing historical climate in the published literature, particularly skill over that of GCMs, and mainly focused on increased skill in simulating precipitation and temperature and their extremes. Over North America, the demonstration of skill in dynamical downscaling started with the RCM study of Dickinson et al. (1989). Recent studies include, but are not limited to: Caldwell et al. (2009; for California), Castro et al. (2007; for the US and Mexico with a focus on the North American Monsoon System), Cocke et al. (2007; for the Southeast US), Evans et al. (2005; a multiple model example for Kansas), Fox-Rabinovitz et al. (2008; on the stretched grid model intercomparison project), Jiao and Caya (2006; for North American summer precipitation), Lucas-Picher et al. (2013; in North American Coordinated Regional Climate Downscaling Experiment (CORDEX) simulations), Martínez-Castro et al. (2006; an RCM sensitivity study for the Caribbean), Martynov et al. (2013; in North American CORDEX simulations), Rauscher et al. (2008; for Meso-American Drought), and Rupp et al. (2007; an impact study example of an RCM used with an ecosystem model in the Yukon River Basin).

Quantifying skill, however, is not one and the same with quantifying uncertainty. There are several basic sources of uncertainty (e.g., Yohe and Oppenheimer 2011) in the projection of climate change, such as uncertainty in emission scenarios. These are reviewed in a regional modeling context in Foley (2010). Some important sources of uncertainty that are specific to dynamical downscaling include: the LBCs, model formulation, effects of natural variability when simulations are short, regional feedbacks, and validation when simulations are at a higher temporal or spatial resolution than available observations and/or where observations are sparse. Uncertainties specific to AOGCMs are also often relevant in AGCMs. The remainder of this section will focus on uncertainties that are specific to dynamical downscaling, focusing on RCMs.

RCM skill in reproducing historical climate when driven by reanalysis is usually higher than when driven by a GCM. This uncertainty from lateral and lower boundary conditions can be summarized as the "garbage in/garbage out problem." With regional climate models, errors in a driving GCM are inherited by an RCM and combine with the systematic errors in the RCM, increasing the uncertainty in the simulation output. Downscaling a GCM that has an inadequate representation of large-scale flow, for example, may well be a futile effort. The use of observed conditions or reanalyses as a driver can also be a source of uncertainty for similar reasons—slight differences in the driving conditions can change the solution (e.g., de Elía et al. (2008) examine the impact of using two different reanalyses as part of their uncertainty analysis). One methodology that has been developed to test for RCM response to LBC errors and verify that RCMs can well reproduce small-scale climate statistics is known as a "Big-Brother/Little-Brother Experiment." Results from this type of experiment can be found in Denis et al. (2002), Antic et al. (2004), Dimitrijevic and Laprise (2005), and Diaconescu et al. (2007).

Lateral boundary conditions, in terms of their placement and treatment, can also cause uncertainty. Domain size and boundary placement can impact simulation outcome (Vannitsem and Chomé 2005; Rauscher et al. 2006; Leduc and Laprise 2009; Separovic et al. 2011). Similarly, RCM formulation can be a source of uncertainty. Certain types of regional analysis may be more uncertain than others if processes that are important to a specific region's climate are not included. For example, if

the use of irrigation over time in a small region has not remained constant, that region's climate has likely been impacted by that change, and this process is likely not included in an RCM unless it has been run specifically with that in mind. Other similar examples can be found in Pitman et al. (2010).

Neglecting feedbacks from an RCM to its parent also provides a source of uncertainty. Smaller-scale regional processes that impact other regions of the globe or grow upscale to impact large-scale circulation will not be allowed to feed back to the global scale when they are resolved with a finer grid in an RCM. In variable resolution global models, this could also be problematic, if two regions are important to one another's climate, but only one is benefiting from the finer mesh.

Uncertainty due to internal model variability is also present in RCMs, but is different than that in AOGCMs, HR-AGCMs, SG-AGCMs, or GCMs with high-resolution orography. While in any variety of GCM, two runs started with slightly different initial conditions (ICs) are bound to diverge after about two weeks, two RCMs started with perturbed ICs, but with the same driving LBCs, will remain correlated throughout their simulation, because of the shared LBCs (e.g., de ElÍa et al. 2002). However, because we are modeling a chaotic system, a single RCM ensemble with perturbed ICs and one driver will still provide an array of solutions, the degree of divergence of which can vary as a function of season, domain size, field of interest, and geographical location (e.g., Giorgi and Bi 2000; Christensen et al. 2001; Caya and Biner 2004; Alexandru et al. 2007; de Elía et al. 2008; Lucas-Picher et al. 2008). Similarly, RCMs can produce projections that are unlike or even opposite in sign to their driving GCMs due to a combination of factors, including differing parameterizations and resolutions (e.g., Han and Roads 2004; Pan et al. 2004; Liang et al. 2006; Bukovsky and Karoly 2011).

Ensembles using different RCMs driven by different GCMs, or ensembles using other dynamical downscaling methods allow for the assessment of some of the uncertainty in the projections referred to here. Multi-model ensembles give a sense of the uncertainty due to model formulation. One advantage to using an ensemble of relatively independent/different models is that they allow for different representations of

feedbacks within the climate system, something that is not possible in statistical downscaling. Feedbacks in clouds, for example, are a large uncertainty in climate change projections. Having different microphysical parameterizations combined with other differences in model formulation will allow for differing magnitudes of cloud feedbacks to temperature and other variables given the different treatment between the models. This allows for some of the uncertainty surrounding this aspect of climate change projections to be better encompassed as well.

Furthermore, an RCM ensemble that includes different drivers allows for a better estimate of the uncertainty due to the LBCs, and multiple realizations of a model facilitate the estimation of the internal variability. One large dynamical downscaling project that aims to aid in the characterization uncertainty in projections of future climate over North America is discussed below.

Review of Results over North America

Independent dynamical downscaling studies over North America have heavily focused on the western US to date. This is due to the desire for improved model performance over this topographically complex and water stressed region. Current published studies do not cover all regions of North America. However, we will review select regions, chosen based on the availability of existing publications, focusing on those since the IPCC AR4 (i.e., those based on the CMIP3 generation of AOGCMs).

More studies with greater regional breadth are expected as a result of the NARCCAP (Mearns et al. 2009), and we will start our review by providing a condensed version of some basic NARCCAP climate change results for the continent.

The results from this entire section are heavily based on RCMs. Other methods of dynamical downscaling used for climate change projection are not as common over North America. Since SG-AGCM are still in the development and testing phase, climate change simulations are not yet the norm. While HR-AGCMs and high-resolution orography AOGCMs are more established methodologically, they are still not as widely used.

North American Regional Climate Change Assessment Program (NARCCAP) Results

The NARCCAP is providing an ensemble of 50-km resolution RCM simulations covering most of North America to facilitate climate change projections for use in impacts research and the investigation of uncertainties in regional scale projections of future climate. Six different RCMs are being used to dynamically downscale four different CMIP3-era AOGCMs and one reanalysis. Twelve combinations/projections are being provided out of the possible 24. Two 50-km HR-AGCM time-slice simulations are also being provided by NARCCAP, the AGCMs representing the atmospheric component of two of the AOGCMs being downscaled in the program. Projections for the future are made using the SRES A2 emission scenario, and simulations cover the period from 2041 to 2070. Historical period simulations cover 1971–2000 (for the AOGCM-driven simulations) and 1981–2004 (for the reanalysis-driven simulations). More detailed information on NARCCAP may be found in Mearns et al. (2009, 2012), or at www.narccap.ucar.edu. Simulations with a North American domain (and also Arctic and Meso-American domains) dynamically downscaling CMIP5 AOGCMs are also being produced as a part of CORDEX (Coordinated Regional Climate Downscaling Experiment, Giorgi et al. 2009), and will be useful for additional analyses over North America in the future.

We provide here an overview of seasonal-mean changes derived from the ensemble of NARCCAP simulations. Numerous, independent publications have documented other aspects of North American climate change, as projected by this ensemble. A few are listed in Mearns et al. (2013a) and in the next section, but it is outside the scope of this section to summarize all NARCCAP-related publications.

At the time of this writing, 11 AOGCM-forced RCM simulations and two HR-AGCM simulations were available for analysis (Mearns et al. 2012). The ensemble means discussed below are composed of the CRCM driven by the Canadian Global Climate Model3.1 (CGCM3.1) and the Community Climate System Model version 3 (CCSM3), the Experimental Climate Prediction Center (ECPC) Regional Spectral

Model (RSM) driven by the GFDL CM 2.1, the HadRM3 driven by the UKMO HadCM3 and the GFDL CM 2.1, the MM5 driven by the CCSM3 and HadCM3, the Regional Climate Model Version 3(RegCM3) driven by the CGCM3.1 and by the GFDL CM 2.1, the WRF driven by the CCSM3 and the CGCM3, and the two AGCM time slices from the GFDL (AM2.1) and CCSM 3.0 (CAM3) atmospheric model components.

The ensemble mean 2-m temperature change projected for mid-century from 13 NARCCAP simulations indicates that the greatest magnitude changes will take place in winter over Canada (Fig. 8.7). Projections for a 2 °C or greater temperature increase in winter by mid-century (2041–2069) cover most of the continent. Overall, projected temperature changes for spring are lowest, but larger changes are found over the Rocky Mountains, especially from the Four Corners region southward into Mexico. Projected changes in summer are largest over the US, but with an ensemble mean change of over 3 °C in many places.

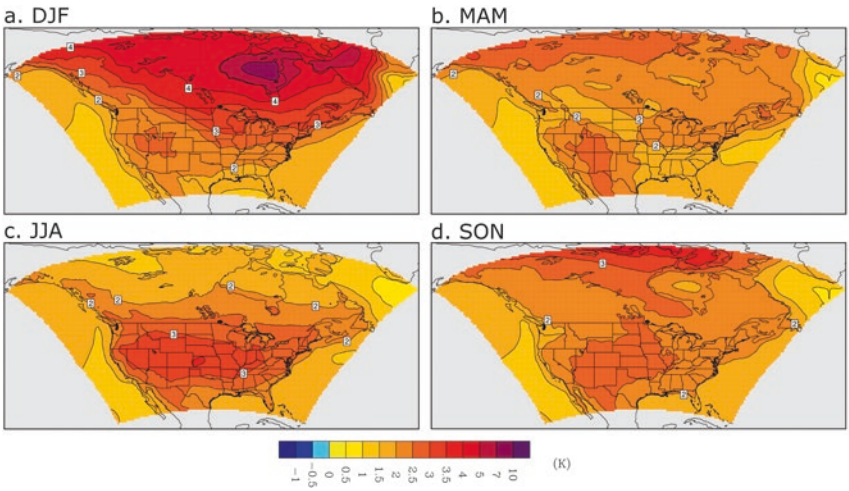

Fig. 8.7 11 RCM + 2 HR-AGCM ensemble mean 2-m temperature change from 1971–1999 to 2041–2069 for December–January (DJF), March–May (MAM), June–August (JJA), and September–November (SON)

Figure 8.8 shows the ensemble mean precipitation change and the number of models that project an increase in precipitation by mid-century. Overall, the simulations indicate more precipitation in the north, and less in the south, with the dividing line shifting by season throughout the US. There is perfect to near-perfect model agreement on an increase in precipitation in Canada in fall and winter of above 10% on

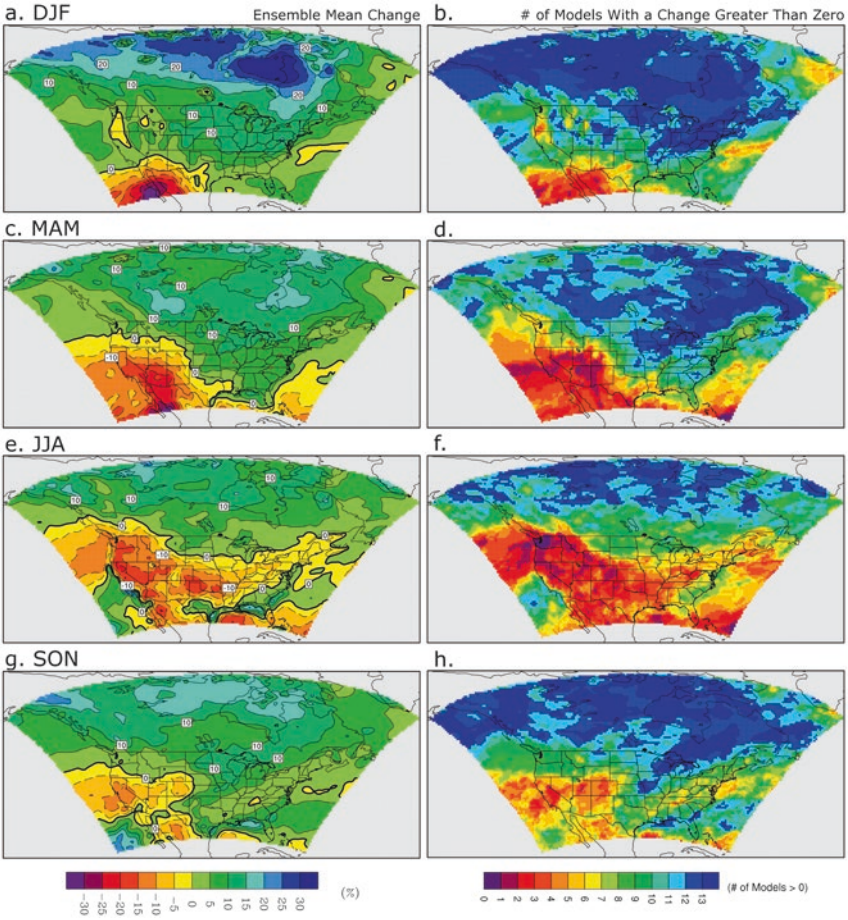

Fig. 8.8 Left column: 11 RCM + 2 HR-AGCM ensemble mean precipitation change from 1971–1999 to 2041–2069. Right column: The number of simulations (out of 13) that project an increase in precipitation

average (Fig. 8.8), with larger increases in the northern territories. The same is true for winter in the Northeast US and Great Lakes region. In the Southeast and Central US there is less agreement on the projection of precipitation, particularly in Fall–Spring, where the model mean comes out with an increase in precipitation, but with less agreement on the overall direction of the change compared to other regions. These regions are near the switch in the direction of the projection in the shoulder seasons, particularly, so less agreement is expected.

The multi-model mean projects drying for southwestern North America in all seasons, though for the southwest US in winter and the region in general in fall, this is not clearly agreed upon. A decrease in precipitation of at least 10% is projected with strong agreement for southwestern North America in spring.

The least agreement overall is found in summer through the center of the continent. This is no surprise, as precipitation in summer is not as dynamically forced as in other seasons. However, the simulations do project a decrease in precipitation over most of southern North America. Where the ensemble mean projects a decrease of 10% or more, for example, the central/southern Plains, northwestern Mexico, and most of the US west, there is often strong model agreement on the drying.

These results are similar to those found in the CMIP3 suite of simulations for winter (as well as to the four GCMs that drove the NARCCAP RCMs (Mearns et al. 2013b)), but in summer, the NARCCAP RCMs altogether indicate a greater decrease in precipitation across the US from the Northwest through the central and southern Plains, and eastward toward the Appalachians, as discussed in Mearns et al. (2013b). The causal processes behind this deeper drying are not yet known.

Summary of Other Projections in the Published Literature

As a point of departure for the discussion in this section, it is worth mentioning that projections derived from dynamical downscaling were not heavily relied upon in the IPCC AR4, perhaps because of their relative scarcity at the time. RCM results in Chap. 13 of the IPCC AR4 WG1 report ("Regional Climate Projections," Christensen et al. 2007) are

generally referred to in the context of evaluating RCM skill, and the CMIP3 model suite was relied upon for projections of mean precipitation and temperature. RCM simulations play the greatest role in the discussion of temperature and precipitation extremes in Christensen et al. (2007), but are mostly limited to extremes in the western US, where the bulk of North American RCM studies have been carried out to date. Since the IPCC AR4 was prepared, many more studies applying dynamical downscaling over North America have been published and are summarized below. Given the distinct national and/or regional isolation of most dynamical downscaling studies, this section has been organized by nation and region as well.

United States

Using the WRF model with a 30-km horizontal resolution, driven by the CCSM 3.0, Bukovsky and Karoly (2011) projected an approximately 18% decrease in average May–August precipitation for the central US for the end of the century (e.g., the 2090s) using the A2 emission scenario, a change opposite in sign to that given by the CCSM 3.0, but in better agreement with other CMIP3 AOGCM projections. This was accompanied by an increase in the number of consecutive dry six-hour periods (and days), but an increase in the intensity of precipitation, and an increase in the magnitude and frequency of extreme precipitation events. Gutowski et al. (2008) also projected an increase in the magnitude of extreme precipitation in this region, specifically the Upper Mississippi River Basin (UMRB), but for the cold season. Gutowski et al. 2008 used the RegCM2 at 50-km driven by the HadCM2 using an equivalent CO2 increase of 1% per year after 1990 (the IS92a scenario). The results of Bukovsky and Karoly (2011) for average precipitation are not consistent with the mid-century results of Jha et al. (2004) for the UMRB. Jha et al. used the same RCM/GCM/emission scenario/resolution combination as Gutowski et al. (2008). The projections from this realization indicated an increase in annual mean precipitation of about 21%, with an increase in every month but November for this basin. Combined with the Soil and Water Assessment Tool (SWAT) hydrologic

model, these results project a 50% increase in annual mean UMRB streamflow. Using the same emission scenario (IS92a) with two versions of the CRCM at 45-km forced by the CGCM2; however, Sushama et al. (2006) indicated a decrease in annual average precipitation over the Mississippi River basin by mid-century. Compounded with a significant decrease in snow cover and an increase in evaporation, this yielded a decrease in flow in the Mississippi River.

Hayhoe et al. (2008) produced dynamical and statistical projections for the Northeast US. They present RCM results from the MM5 driven by the PCM given the A1FI and B1 SRES scenarios for early, mid, and late century. This model study indicates increases in daily maximum temperature of over 3 °C by late century in the northern part of this region, and a doubling of the number of days per year above the 1990 90th percentile temperature. Hayhoe et al. (2008) also projected mixed changes in average precipitation and precipitation intensity, depending upon the location within the Northeast US. For example, in Maine, they found that statistical downscaling resulted in precipitation decreases of about 100 mm, whereas the regional model projected increases of about 100 mm. Decreases in intensity and mean rainfall were found for the northern part, with increases in both in the southern portion. Rawlins et al. (2012), using NARCCAP simulations, show a significant increase in temperature in the Northeast (2–3 °C), a significant increase in winter precipitation, and changes generally still within natural variability by mid-century in other seasons, though a significant increase in precipitation is projected in spring in the central Northeast, and a significant decrease in the southern part of this region in summer.

Studies in the western US are often concerned with changes in snowfall, snowpack, and snowmelt. All RCM climate change studies predict warming for the West in all seasons. Warming is strongest at high elevations due to a snow-albedo feedback, particularly in regions where snow is transitioning to rain more frequently and/or the average melting level is moving upwards (e.g., Snyder and Sloan 2005, for the Sierra Nevada Mountains; Duffy et al. 2006, in one of four simulations for the western US; Salathé et al. 2008, forthe US Pacific Northwest; Wi et al. 2012, for the Colorado River Basin (CRB)). Gao et al. (2011), using six NARCCAP

RCMs, show that the magnitude of the warming in the RCMs is less than in their driving GCMs for the CRB headwaters, indicating that, in some cases at high elevation, rivers may be less susceptible to a warming climate in the RCMs than the GCMs. Dynamically downscaled projections in Hall et al. (2012) for Los Angeles County illustrate the detail one may gain in projections at very high-resolution. Their 2-km nested WRF simulations exhibit clear differences between coastal and inland warming, with larger warming by 1–2 °F inland and at higher elevations, as shown in Fig. 8.9. The strongest warming in the Los Angeles region is found over inland desert in summer and fall. Temperature

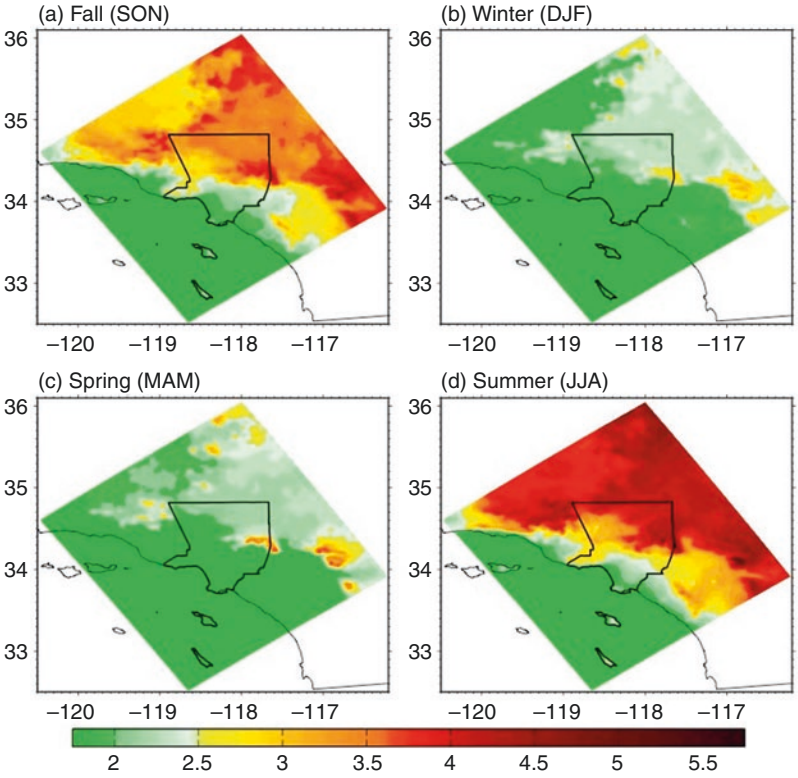

Fig. 8.9 Dynamically downscaled seasonal-mean surface air temperature change (2041–2060 minus 1981–2000) from the CCSM4 downscaled by WRF to 2-km in °F (Fig. 7 from Hall et al. 2012)

extremes are not as extensively covered in the literature; however, in summer, more hot days and extreme hot days are also expected in some regions (e.g., Sacramento Valley, southern CA, and Nevada: Pan et al. 2010).

In the West, there is also an overall consensus that there will be an increasing fraction of precipitation falling as rain instead of snow in the winter (e.g., Leung et al. 2004). The direction of change for precipitation varies, however, based on sub-region, season, RCM, and study. For example, Leung et al. (2004) project no significant change in precipitation except for a drying trend in summer for the West; little consistency is indicated in the direction of change in the multi-model ensembles of Duffy et al. (2006) and Dominguez et al. (2012); Pan et al. (2010) projects an increase in winter precipitation, but a decrease in summer for California and Nevada; Wi et al. (2012) show an insignificant increase in winter precipitation over the CRB, but Rasmussen et al. (2011) have a cool season (November–May) increase of 26% over their full CRB domain; Gao et al. (2011) project a decrease in summer precipitation over the CRB from their six-member ensemble, however; Snyder and Sloan (2005) show little change in total precipitation in the Sierra Nevada; and inconsistency is present between RCMs in Salathé et al. (2010) in Washington State, but more fall precipitation is possible in Washington and the Pacific Northwest, particularly along the windward side of the mountains (Salathé et al. 2008, 2010).

Regardless of the direction or magnitude of change in average precipitation, increase in the intensity of future extreme cool season precipitation for the West is more uniformly projected (e.g., Leung et al. 2004; Rosenberg et al. 2010; Salathé et al. 2010 (including Vancouver Island and the British Columbia coastal range), and Dominguez et al. 2012).

Canada

Sushama et al. (2007), using the CRCM at 45-km driven by the CGCM3 with an A2 SRES scenario, investigated the impacts of climate change on North American permafrost zones. The majority of Canada is covered by some fraction of Tundra, categorized by Sushama et al. as isolated,

sporadic, discontinuous, or continuous tundra, moving Northward with category, generally. Significant increases in near-surface soil temperature were indicated in all four zones, with a 4–6 °C increase in the continuous permafrost by mid-century (2041–2070). An increase in precipitation was also projected in all zones in all months, with a 15–20% increase in annual average precipitation by mid-century. However, in all but the continuous permafrost zone, a decrease in snow-water-equivalent was projected. A decrease in frozen soil content from the warming combined with the increase in precipitation could lead to intensification of the hydrological cycle, and this was indicated for the isolated permafrost zone.

With the same model combination as above, but using the IS92a emission scenario, Sushama et al. (2006) projected a 2–4% increase in annual average precipitation and an increase in annual average runoff in the Mackenzie, Yukon, and Fraser River basins by mid-century. These basins were also projected to see a significant decrease in snow cover, a related attenuated and earlier snowmelt peak, but increased fall, winter, and spring streamflow.

The increase in precipitation seen in the above two studies was also projected in Bresson and Laprise (2011).

Mailhot et al. (2007) and Mladjic et al. (2011) examined changes in extreme precipitation, and both project an increase in the magnitude of extreme events. In Mailhot et al. (2007), with a model set up similar to the Sushama et al. studies above, May–October return periods for annual maximum rainfall depths were halved for two-hour and six-hour events and decreased by 1/3 for 12–24-hour duration events by mid-century for southern Quebec. The extreme events in Mailhot et al. (2007) were more likely to come from more localized convective weather systems than previously. Using the NARCCAP suite of model simulations, Mailhot et al. (2012) projected large increases in annual maximum rainfall depth in the mid-latitude, inland, and Great Lakes regions in all examined return periods and durations by mid-century. Similarly, Mladjic et al. (2011) used a 10-member CRCM ensemble driven by the CGCM3 with the A2 SRES scenario to show an increase in the magnitude of extreme precipitation events of varying duration for Canada.

Mexico and Island Nations of the Caribbean Sea

Fewer dynamically downscaled projections exist for Mexico and the Caribbean than other parts of North America. A PRECIS simulation by Karmalkar et al. (2011) at an approximately 25-km resolution projects that Mexico will experience a warming of 3–4 °C by the end of the twenty-first century, with warming greater than or near 4 °C in the wet season and over the Yucatan Peninsula, and an amplified warming at elevation as well. This result is similar to that made in Pérez-Pérez et al. (2007), where a 22-km AGCM simulation projected a 3–4 °C increase in temperature by the end of the twenty-first century, mainly over northwestern Mexico. Over the Caribbean, a 50-km PRECIS simulation by Campbell et al. (2011) projects warming of 1–5 °C, with the greatest warming over land, particularly over the largest islands. Precipitation projections from these two studies generally agree with the CMIP3 AOGCM ensemble average presented in Christensen et al. (2007). For Mexico and the Caribbean, significant drying, outside of natural variability, is projected for the wet season. Precipitation decreases of around 30–40% are projected for eastern and southern Mexico and the Yucatan Peninsula, and decreases of 25–50% around the Caribbean basin. Rauscher et al. (2011) present potential explanations for this drying. A split pattern of precipitation change is indicated in Campbell et al. (2011) for the dry season, with the northern Caribbean (above 22 °N) seeing up to a 75% increase in precipitation through an increase in the intensity of precipitation and a decreased number of dry days and the opposite (around 50% less precipitation) in the southern Caribbean.

Hurricanes constitute one of the most important meteorological hazards and sources of moisture for the coastal region of Mexico, Central America, the islands of the Caribbean, and even the southwestern US (Englehart and Douglas 2001; Larson et al. 2005; Ritchie et al. 2011). Therefore, changes in tropical cyclone trajectory, frequency, and intensity may have a large influence on projections of precipitation over Mexico and the Caribbean Islands. However, dynamical downscaling is not always successful in reproducing precipitation associated with tropical cyclone activity over the tropical Americas because of fatal biases in driving AOGCM boundary conditions and regional model resolution

and configuration. Common problems include tropical cyclones that are too weak and/or too few (e.g., Karmalkar et al. 2011; Knutson et al. 2008).

Knutson et al. (2008), using an 18-km resolution RCM framework developed for downscaling Atlantic hurricane activity, suggest that tropical cyclone frequency over the tropical Atlantic will decrease, while rainfall rates increase. However, Bender et al. (2010), using the GFDL hurricane model with a grid spacing of 8 km, project a decrease in the overall frequency of Atlantic hurricanes, but a near doubling of Category 4 and 5 storms, categories not well captured in Knutson et al. (2008). Although it is likely that precipitation rates associated with tropical cyclones will increase (Knutson et al. 2010), it is not clear to what extent this, combined with a change in the frequency of tropical cyclones, is responsible for projected changes in wet season precipitation over Mexico and the Caribbean. It is clear that more work needs to be done in this region, accounting for changes in tropical cyclone activity and their interannual and multidecadal variability (Goldenberg et al. 2001; Pérez-Pérez et al. 2007).

Discussion of Dynamical Downscaling Results

Despite the plethora of regional modeling studies completed over North America, it remains difficult to make definitive statements about climate change over North America from dynamically downscaled simulations, outside of perhaps the US West and Great Plains regions. In the West, for example, enhanced warming at higher elevations due to the snow-albedo feedback, changes in the timing of snow melt, and the increase in the contribution of rain instead of snow to seasonal precipitation totals are consistently reproduced projections from RCMs in most of this region. This difficulty is due to the lack of overlap of the studies and absence of a broad range of modeling uncertainties. Many studies are local-to-regional in scale, and do not use a variety of emission scenarios, AOGCMs, or RCMs. It is infrequent that the climate changes from dynamically downscaled simulations are compared to those from their parent GCMs. Similarly, when analysis focuses on temperature and/or precipitation, too

seldom are attempts made to more completely explain differences in verification and differences in climate changes seen between RCM and parent AOGCM through a thorough analysis of causal atmospheric processes. It is also rare to find studies that try to go beyond verifying the performance of their dynamical downscaling approach, to showing that they do or do not add value to the projections from their coarser resolution parents, a subject that will be discussed more in Sect. 8.6.

8.5 Comparison Among Methods and Shared Uncertainties

Numerous comparisons of downscaling methods have been performed in the past couple of decades. The early studies focused on comparison of regional climate model simulations and regression-based statistical downscaling approaches (see, e.g., Giorgi et al. 2001). The overarching conclusions were that for the present climate both techniques had similar skill; for future climate projections, the two techniques had important differences. No conclusions regarding the relative credibility of the different techniques were established in these early works. However, it was pointed out that statistical techniques can "go wrong" based on the choice of predictors, since those predictors with high explanatory power for the present climate could exclude predictors important for conditions under the changed climate (Giorgi et al. 2001). The reader is directed to Giorgi et al. (2001) and Christensen et al. (2007) for more details on these earlier comparisons. In this section, we concern ourselves with work published primarily since the 2007 IPCC Reports.

The relative skill of ESD versus dynamical downscaling approaches and hybrids thereof is highly dependent on the specific variable under consideration, the location, and the specific ESD and RCM applied. Examples of comparative analyses include: Haylock et al. (2006); Landman et al. (2009); Schoof et al. (2009); and Wood et al. (2004). In general, intensive studies of the performance of ESD relative to dynamical approaches reveal a similar level of skill (e.g., Lim et al. 2007). See the recent review of Maraun et al. (2010) for a summary of evaluation methods and metrics.

In a comparison of cold season precipitation with a five-month lead time, output from a seven-member ensemble RCM suite exhibited similar performance in terms of correlation with observations in the topographically complex western US with values derived from interpolation downscaling using Bias Correction and Spatial Disaggregation (BCSD) but generally outperformed statistical downscaling using a Bayesian merging technique (Yoon et al. 2012). The overall hindcast skill (1982–2003) is relatively poor and spatially variable. But the skill of the downscaling methods is generally greater than that of the of the driving global forecast model, as measured by precipitation anomaly correlation coefficients relative to independent observations (Yoon et al. 2012).

A statistical downscaling model (Statistical DownScaling Model 6 Version 4.2) was better able to capture the observed climatology of extreme precipitation at 15 stations distributed across the Northeastern US than output from the Hadley Centre regional climate model when each was driven with output from the HadCM3 Global Climate Model (Tryhorn and DeGaetano 2011). When applied in a climate projection mode (2041–2060 relative to 1981–2000), HadRM3 indicated much larger magnitude increases in extreme precipitation (return periods of 2, 50, and 100 years) than were derived using SDSM (Tryhorn and DeGaetano 2011). For example, in one location, the RCM projected much higher changes in 100-year events (29%) compared to that of the SDSM statistical technique (7%). However, there was considerable site-to-site variability in the degree of agreement with extreme values derived from observed data, and this result is to some extent specific to the particular RCM used.

Another study comparing the CLIGEN stochastic weather generator (which produces daily estimates of precipitation and other weather variables for a single geographic point, using monthly parameters (means, SDs, skewness, etc.) derived from the historic measurements), SDSM (with bias correction applied), and the Canadian RCM (both with and without bias correction) driven by reanalysis data found that the seasonality of precipitation over Quebec was better simulated by CRCM and SDSM than by the stochastic weather generator. The uncertainty in the climate change signal resulting from the application of the various downscaling techniques to the GCM output contained almost equal

contributions from that deriving from the downscaling method and that deriving from the GCM (Chen et al. 2011b).

The various uncertainties relevant to statistical and dynamical downscaling are discussed above in Sects. 8.3.2 and 8.4.2. Several of these are directly inherited from the uncertainties surrounding projections of climate change via global climate models (see Chap. 6) and thus are common to the different downscaling approaches. But, as discussed, additional uncertainties arise specific to the particular downscaling context.

Recent attempts have been made to attribute the relative uncertainty across the different sources of uncertainty (for global models, see Hawkins and Sutton 2011) including the downscaling method. A good example of such work is Li et al. (2012), in which the projection uncertainty in high-resolution temperatures was decomposed into that deriving from (i) statistical downscaling, [*Harmon R. Holcomb*] choice of regional climate model, [*Harmon R. Holcomb*] choice of AOGCM, (iv) internal climate variability, and (v) linear and non-linear interactions (Fig. 8.10). Hence, they examine both statistical and dynamical downscaling. Their results indicate that downscaling the AOGCM using the RCM dominates uncertainty in high-resolution temperature projections for short lead times, but by the end of the century (i)-[*Harmon R. Holcomb*] have comparable contributions to the total uncertainty.

8.6 Research Issues and Future Needs

Added Value

One of the most important issues is that of added value, which is the additional knowledge about the climate (current and future) gained from applying an RCM or any other downscaling method. Many articles refer to the added value of RCMs and claim to have demonstrated this, but in reviewing the many articles about RCMs (over North America) it is difficult to determine if added value has actually been established (Feser et al. 2011). Of course, this is partially a function of what metric is used to define added value. For example, Kanamitsu and de Haan (2011) developed an added value index (AVI) based on a characteristic spatial

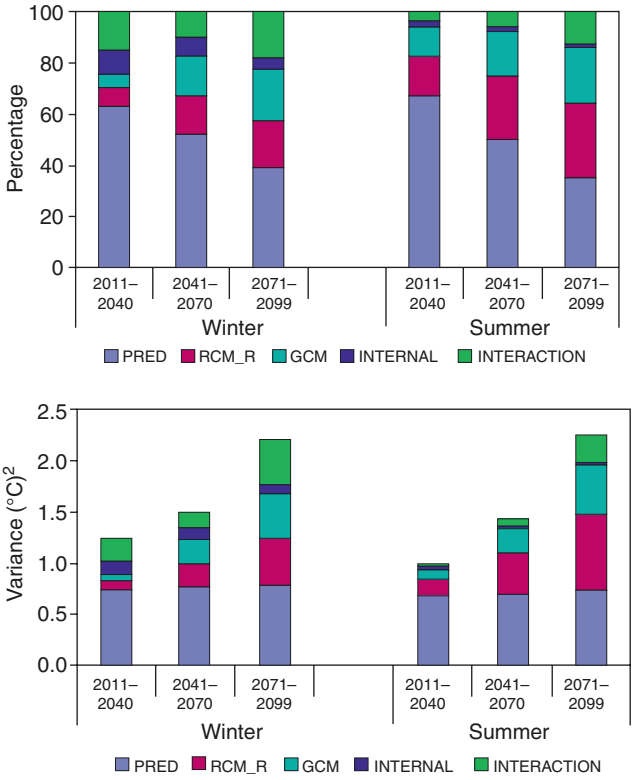

Fig. 8.10 The percentage (*right*) and variance (*left*) of different factors contributing to the total uncertainty under a given emissions scenario averaged across the domain of North America. Terms PRED, RCM_R, GCM, Internal, and Interaction represent contributions from statistical downscaling, choice of RCM, choice of AOGCM, internal variability simulated by the AOGCM, and interactions terms combined, respectively (From Li et al. 2012, Fig. 6)

distribution of skill rather than average values for regional models and applied this to downscaled seasonal forecasts. Di Luca et al. (2012) used variance decomposition techniques to develop a potential added value index (POV). Many papers rely mainly on decreased biases in mean seasonal temperature or precipitation and improved spatial and temporal correlations of same with observations to measure added value (e.g., Prömmel et al. 2010; Racherla et al. 2012).

Castro et al. (2005) noted that RCMs add value by resolving small-scale features but not on the larger scale. Di Luca et al. (2013) demonstrate, using a subset of the NARCCAP temperature results, that the RCMs have low potential to add value over coarser models, but that one area where value is added is along coastlines. However, for precipitation, the POV is more distinct, particularly over fine time scales (e.g., three hourly), during the warm season months, and over complex topography in all seasons (Di Luca et al. 2012). Racherla et al. (2012) demonstrated some added value in simulations with very near-term climatic change (based on the GISS Class E global model, and the WRF RCM) comparing different current period decades (1970s vs. 1990s–2000s) with and without analysis-nudging in the RCM, and examined seasonal temperature and precipitation over North America. They found that only with nudging did the downscaled simulations improve the reproduction of the near-term climate change, and then only slightly. Feser et al. (2011) demonstrated that RCMs add value for some variables in some locations (e.g., temperature along coasts) but not others (e.g., sea-level pressure over the oceans). (See Prömmel et al. (2010) and Feser et al. (2011) for reviews of other efforts).

There has been relatively little research on the ability of RCMs to capture near-surface wind climates. However, analyses of wind speeds from the NARCCAP suite provide clear indication of added value in applying RCMs relative to the driving AOGCM (Pryor et al. 2012a), and there is some indication that adopting a non-hydrostatic formulation even at these spatial scales does "improve" model simulations of extreme wind speeds (Pryor et al. 2012b).

Ultimately the evidence for added value of RCMs is mixed at this point, and seems to vary based on variable investigated, metrics used, the temporal scale, season, and region**.** There is, however, mounting evidence of added value in topographically complex regions and coastlines as well as for certain types of extremes (e.g., daily precipitation).

Added value is also discussed somewhat in ESD, but much less frequently than in the context of RCM evaluation. Multivariate Adapted Constructed Analogs [*MacArthur and Wilson*] were demonstrated to exhibit skill above direct interpolation for temperature, humidity, wind speed, and precipitation over the western US (Abatzoglou and Brown

2012). An analysis of statistical downscaling for stream-flow in Quebec found added value (relative to use of direct model output) in precipitation occurrence and amount by statistical downscaling propagated through to enhanced flow forecasts relative to results generated by a hydrologic model conditioned on the raw output from a numerical weather prediction model (Muluye 2011).

In all of these cases, however, model (statistical or dynamical) performance vis á vis observations has been the key component of establishing added value. While it is obvious that this is a necessary condition for establishing added value, it may not be a sufficient condition. Racherla et al. (2012) performed an interesting experiment that broke out of this mode by viewing models' performance regarding observed climate change. Their conclusions, however, regarding limited added value are hampered by, among other things, lack of statistical tests to determine if these very near-term changes rise above the noise of natural variability.

We suggest that more process-based analyses of the effect of biases/errors in the current period on how the model responds under changed forcing (e.g., increased GHGs) are also necessary. One can have poor validation results in some aspects (e.g., mean temperature bias) of an RCM simulation but still find its current and future climate simulations credible based on careful process level analyses. The scale of the evaluation is an important factor in this context: at what scale should there be added value? There has been no full exploration of this condition, and we consider this a very important research need (see Bukovsky et al. 2013 for an effort in the right direction). Moreover, a thorough review of what sensibly constitutes "added value" needs to be performed.

High-Resolution AOGCMs Versus Downscaling

As the spatial resolution of fully coupled AOGCMs and more advanced Earth System Models (ESMs) continues to increase, another important issue is the future for various downscaling techniques. It has been predicted that ESMs will be running for at least 100-year transients at 10 km within 10 years (NAS 2012). This condition will certainly affect how and in what contexts downscaling will be used. However, given that RCM

simulations at 2 km are currently being produced (e.g., Rasmussen et al. 2011; Hall et al. 2012), certainly for the foreseeable future various downscaling techniques will remain useful tools. It is abundantly clear that certain important phenomena (e.g., tropical cyclones) will require very high resolutions, and it will certainly be to scientists' and society's benefit to understand these phenomena better under conditions of climate change. Statistical downscaling to point locations obviously would also remain relevant.

Whether we consider very high-resolution AOGCM simulations or downscaling approaches, selection of appropriate techniques to verify simulations at such high resolution becomes more and more problematic. New statistical and data mining efforts are needed to produce data sets that are up to the task of high-resolution validation.

More Complete and Balanced Exploration of Uncertainty

In various sections of this chapter, certain types of uncertainty relevant to downscaling are focused on as opposed to others. For example, while there are comparisons of statistical and dynamical downscaling (see Sect. 8.4.1), these have not been very systematic, and any conclusions drawn from them have been very limited. Much progress could be made in this arena by increasing coordination among projects, and/or producing integrated programs to begin with. The NOAA Climate Prediction and Projection Platform (NCPP) is on target to more systematically explore multiple uncertainties including different downscaling methods (Barsugli et al. 2013).

The uncertainties that have been most typically explored are those concerned with emissions/concentration trajectories, different AOGCMs, and different downscaling models (e.g., Schoof et al. 2010; Mearns et al. 2013b; Hall et al. 2012). While these have been useful explorations, most remain incomplete, and the effects of these uncertainties on downstream impacts models remain even more unclear. We need to more systematically examine these uncertainties and evaluate possible shortcuts for complete exploration (e.g., Hall et al. 2012).

The uncertainty of internal variability in simulations with GCMs and how this would translate into uncertainties in statistical and dynamical downscaling has been woefully neglected. Evidence from Deser et al. (2010, 2012) indicates that the uncertainty in GCMs due to internal variability has been underestimated in transient simulations of current/future projections of climate change. While uncertainty in realizations of RCMs is limited due to the commonality of the LBCs, the uncertainty from different realizations of global models providing different LBCs and what that effect would be has not been investigated. Statistical downscaling using large-scale variables from multiple realizations of global models has also not been fully explored.

8.7 Key Findings

We have presented an overview of the common methods of downscaling global climate models to simulate long-term current climate and future climate change. We have reviewed: the major features of these different methods, the information about future climate change over North America based on the application of these methods, and the literature comparing these different methods over North America. Finally, we have presented key issues and suggestions for future research.

The key findings from this chapter are the following:

- On a large region level, the climate changes projected by downscaling techniques are not dissimilar from those produced by global models. However, there is mounting evidence that downscaling does provide additional information—added value—beyond that of the driving large-scale models in topographically complex regions and coastal areas as well as for certain types of extremes (e.g., daily precipitation).
- Comparisons of the methods (e.g., statistical downscaling vs. dynamical downscaling) indicate that they often result in different climate changes. For example, in the northeast US one study found that statistical downscaling resulted in precipitation decreases of about 100 mm, whereas a regional model projected increases of about 100 mm. Similarly, in a study of precipitation extremes in the Northeast, an RCM projected much higher changes in 100-year events (29%) com-

pared to that of the SDSM statistical technique (7%). However, most of the literature does not make evaluative comparisons of the methods. Most comparisons do not embrace a wide variety of methods or the whole domain of North America. Lack of uniformity of experiments makes intercomparisons difficult.

- Critical research needs to include much more attention to where, why, and when different methods would be most useful, which methods add value, which do not, and the future use of downscaling versus high-resolution AOGCMs.
- Commonalities in downscaled projections of temperature include: clear tendencies towards increased temperatures particularly in winter, increased duration of the growing (or frost-free) season and an increase in the frequency with which extreme temperatures will be observed. It should be noted, however, that these broad-scale general tendencies are the same as found in AOGCM results (see Chap. 6), but the magnitudes of change vary across different downscaling techniques and different sub-regions.
- An increasing consensus is also appearing with respect to precipitation regimes. For example, recent dynamical and statistical downscaling over North America indicates evidence for "wetting" of the northern Pacific Northwest in winter (e.g., about 10% increase, although somewhat less in the NARCCAP suite) and drying of large areas within the continental interior in summer. Again, however, these broad-scale results are also seen in global models (Chap. 6), but the magnitudes of change vary across different downscaling techniques. For example, in the most complete program using different AOGCMs to drive different RCMs over most of North America (NARCCAP), it was found that changes in precipitation were more extreme (decreases in summer, increases in winter) compared to the GCMs that drove the RCMs. In the Central Plains, for example, mean decreases from the CMIP3 models were in the range of 5–10% but 10–20% from the NARCCAP RCMs.
- Much of the downscaling analyses conducted to date have focused on changes in temperature and precipitation regimes. However, there is a need to expand the suite of variables to include others, such as extreme wind speeds, to meet the requirements of climate change adaptation researchers.

- The uncertainties explored with respect to future climate and various downscaling methods have not been well balanced. Typically, different AOGCMs are downscaled, but less attention is paid to different downscaling techniques, emissions/concentrations scenarios, and internal variability. Broader exploration of the various uncertainties is warranted.

References

Abatzoglou, John T., and Timothy J. Brown. 2012. A Comparison of Statistical Downscaling Methods Suited for Wildfire Applications. *International Journal of Climatology* 32 (5): 772–780.

Alexandru, Adelina, Ramon de Elia, and René Laprise. 2007. Internal Variability in Regional Climate Downscaling at the Seasonal Scale. *Monthly Weather Review* 135 (9): 3221–3238.

Alexandru, Adelina, Ramon De Elia, René Laprise, Leo Separovic, and Sébastien Biner. 2009. Sensitivity Study of Regional Climate Model Simulations to Large-Scale Nudging Parameters. *Monthly Weather Review* 137 (5): 1666–1686.

Antic, S., R. Laprise, B. Denis, and R. De Elía. 2004. Testing the Downscaling Ability of a One-Way Nested Regional Climate Model in Regions of Complex Topography. *Climate Dynamics* 23: 473–493.

Barsugli, Joseph J., Galina Guentchev, Radley M. Horton, Andrew Wood, Linda O. Mearns, Xin-Zhong Liang, Julie A. Winkler, et al. 2013. The Practitioner's Dilemma: How to Assess the Credibility of Downscaled Climate Projections. *Eos, Transactions American Geophysical Union*. https://doi.org/10.1002/2013EO460005/full.

Bender, Morris A., Thomas R. Knutson, Robert E. Tuleya, Joseph J. Sirutis, Gabriel A. Vecchi, Stephen T. Garner, and Isaac M. Held. 2010. Modeled Impact of Anthropogenic Warming on the Frequency of Intense Atlantic Hurricanes. *Science* 327 (5964): 454–458.

Benestad, Rasmus E., Inger Hanssen-Bauer, and Deliang Chen. 2008. *Empirical-Statistical Downscaling*. Hackensack: World Scientific Publishing Co Inc.

Bresson, Raphaël, and René Laprise. 2011. Scale-Decomposed Atmospheric Water Budget over North America as Simulated by the Canadian Regional Climate Model for Current and Future Climates. *Climate Dynamics* 36 (1–2): 365–384.

Bukovsky, Melissa S., and David J. Karoly. 2011. A Regional Modeling Study of Climate Change Impacts on Warm-Season Precipitation in the Central United States. *Journal of Climate* 24 (7): 1985–2002.

Bukovsky, Melissa S., David J. Gochis, and Linda O. Mearns. 2013. Towards Assessing NARCCAP Regional Climate Model Credibility for the North American Monsoon: Current Climate Simulations. *Journal of Climate* 26 (22): 8802–8826. https://doi.org/10.1175/JCLI-D-12-00538.1.

Caldwell, Peter, Hung-Neng S. Chin, David C. Bader, and Govindasamy Bala. 2009. Evaluation of a WRF Dynamical Downscaling Simulation over California. *Climatic Change* 95 (3): 499–521.

Campbell, Jayaka D., Michael A. Taylor, Tannecia S. Stephenson, Rhodene A. Watson, and Felicia S. Whyte. 2011. Future Climate of the Caribbean from a Regional Climate Model. *International Journal of Climatology* 31 (12): 1866–1878.

Cañón, Julio, Francina Domínguez, and Juan B. Valdés. 2011. Downscaling Climate Variability Associated with Quasi-Periodic Climate Signals: A New Statistical Approach Using MSSA. *Journal of Hydrology* 398 (1): 65–75.

Carter, T. R., M. L. Parry, and H. Harasawa. 1994. *IPCC Technical Guidelines for Assessing Climate Change Impacts and Adaptations*. London: Department of Geography, University College London.

Castro, Christopher L., Roger A. Pielke, and Giovanni Leoncini. 2005. Dynamical Downscaling: Assessment of Value Retained and Added Using the Regional Atmospheric Modeling System (RAMS). *Journal of Geophysical Research: Atmospheres* 110: 1–21.

Castro, Christopher L., Roger A. Pielke Sr., and Jimmy O. Adegoke. 2007. Investigation of the Summer Climate of the Contiguous United States and Mexico Using the Regional Atmospheric Modeling System (RAMS). Part I: Model Climatology (1950–2002). *Journal of Climate* 20 (15): 3844–3865.

Caya, Daniel, and Sébastien Biner. 2004. Internal Variability of RCM Simulations over an Annual Cycle. *Climate Dynamics* 22 (1): 33–46.

Caya, Daniel, and Rene Laprise. 1999. A Semi-Implicit Semi-Lagrangian Regional Climate Model: The Canadian RCM. *Monthly Weather Review* 127 (3): 341–362.

Cha, Dong-Hyun, Chun-Sil Jin, Dong-Kyou Lee, and Ying-Hwa Kuo. 2011. Impact of Intermittent Spectral Nudging on Regional Climate Simulation Using Weather Research and Forecasting Model. *Journal of Geophysical Research: Atmospheres* 116: 1–11.

Chen, W., Z. Jiang, L. Li, and P. Yiou. 2010. Simulation of Regional Climate Change Under the IPCC A2 Scenario in Southeast China. *Climate Dynamics* 36: 491–507.

Chen, Jie, François P. Brissette, and Robert Leconte. 2011a. Uncertainty of Downscaling Method in Quantifying the Impact of Climate Change on Hydrology. *Journal of Hydrology* 401 (3): 190–202.

Chen, Weilin, Zhihong Jiang, Laurent Li, and Pascal Yiou. 2011b. Simulation of Regional Climate Change under the IPCC A2 Scenario in Southeast China. *Climate Dynamics* 36 (3–4): 491–507.

Choi, Woonsup, Peter F. Rasmussen, Adam R. Moore, and Sung Joon Kim. 2009. Simulating Streamflow Response to Climate Scenarios in Central Canada Using a Simple Statistical Downscaling Method. *Climate Research* 40 (1): 89–102.

Christensen, O.B., M.A. Gaertner, J.A. Prego, and J. Polcher. 2001. Internal Variability of Regional Climate Models. *Climate Dynamics* 17 (11): 875–887.

Christensen, Jens Hesselbjerg, Bruce Hewitson, Aristita Busuioc, Anthony Chen, Xuejie Gao, R. Held, et al. 2007. Chapter 11: Regional Climate Projections. In *Climate Change, 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, ed. S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor, and H.L. Miller. Cambridge/New York: Cambridge University Press.

Cocke, Steven, T.E. LaRow, and D.W. Shin. 2007. Seasonal Rainfall Predictions over the Southeast United States Using the Florida State University Nested Regional Spectral Model. *Journal of Geophysical Research: Atmospheres* 112: 1–14.

Cosgrove, B.A., et al. 2003. Land Surface Model Spin-Up Behavior in the North American Land Data Assimilation System (NLDAS): GEWEX Continental-Scale International Project, Part 3 (GCIP3). *Journal of Geophysical Research* 108 (D22). https://doi.org/10.1029/2002JD003316.

Côté, Jean, Sylvie Gravel, André Méthot, Alain Patoine, Michel Roch, and Andrew Staniforth. 1998. The Operational CMC–MRB Global Environmental Multiscale (GEM) Model. Part I: Design Considerations and Formulation. *Monthly Weather Review* 126 (6): 1373–1395.

Cotton, William R., R.A. Pielke Sr., R.L. Walko, G.E. Liston, C.J. Tremback, H. Jiang, R.L. McAnelly, et al. 2003. RAMS 2001: Current Status and Future Directions. *Meteorology and Atmospheric Physics* 82 (1): 5–29.

Cubasch, U., J. Waszkewitz, G. Hegerl, and J. Perlwitz. 1995. Regional Climate Changes as Simulated in Time-Slice Experiments. *Climatic Change* 31 (2–4): 273–304.

Cueto, Rafael O. García, Adalberto Tejeda Martínez, and Ernesto Jáuregui Ostos. 2010. Heat Waves and Heat Days in an Arid City in the Northwest of Mexico: Current Trends and in Climate Change Scenarios. *International Journal of Biometeorology* 54 (4): 335–345.

de Elia, Ramón, René Laprise, and Bertrand Denis. 2002. Forecasting Skill Limits of Nested, Limited-Area Models: A Perfect-Model Approach. *Monthly Weather Review* 130 (8): 2006–2023.

de Elía, Ramón, Daniel Caya, Hélène Côté, Anne Frigon, Sébastien Biner, Michel Giguère, Dominique Paquin, Richard Harvey, and David Plummer. 2008. Evaluation of Uncertainties in the CRCM-Simulated North American Climate. *Climate Dynamics* 30 (2–3): 113–132.

Denis, Bertrand, René Laprise, Daniel Caya, and J. Côté. 2002. Downscaling Ability of One-Way Nested Regional Climate Models: The Big-Brother Experiment. *Climate Dynamics* 18 (8): 627–646.

Déqué, M., R.G. Jones, M. Wild, F. Giorgi, J.H. Christensen, D.C. Hassell, P.L. Vidale, et al. 2005. Global High Resolution vs. Regional Climate Model Climate Change Scenarios over Europe: Quantifying Confidence Level from PRUDENCE Results. *Climate Dynamics* 25: 653–670.

Deser, Clara, Adam Phillips, Vincent Bourdette, and Haiyan Teng. 2010. Uncertainty in Climate Change Projections: The Role of Internal Variability. *Climate Dynamics*. https://doi.org/10.1007/s00382-010-0977-x.

Deser, Clara, Reto Knutti, Susan Solomon, and Adam S. Phillips. 2012. Communication of the Role of Natural Variability in Future North American Climate. *Nature Climate Change* 2 (11): 775–779.

Di Luca, Alejandro, Ramón de Elía, and René Laprise. 2012. Potential for Added Value in Precipitation Simulated by High-Resolution Nested Regional Climate Models and Observations. *Climate Dynamics* 38 (5–6): 1229–1247.

———. 2013. Potential for Added Value in Temperature Simulated by High-Resolution Nested RCMs in Present Climate and in the Climate Change Signal. *Climate Dynamics* 40 (1–2): 443–464.

Diaconescu, Emilia Paula, René Laprise, and Laxmi Sushama. 2007. The Impact of Lateral Boundary Data Errors on the Simulated Climate of a Nested Regional Climate Model. *Climate Dynamics* 28 (4): 333–350.

Dibike, Y.B., P. Gachon, A. St-Hilaire, T.B.M.J. Ouarda, and Van T.-V. Nguyen. 2008. Uncertainty Analysis of Statistically Downscaled Temperature and Precipitation Regimes in Northern Canada. *Theoretical and Applied Climatology* 91 (1): 149–170.

Dickinson, Robert E., Ronald M. Errico, Filippo Giorgi, and Gary T. Bates. 1989. A Regional Climate Model for the Western United States. *Climatic Change* 15 (3): 383–422.

Dimitrijevic, Milena, and René Laprise. 2005. Validation of the Nesting Technique in a Regional Climate Model and Sensitivity Tests to the Resolution of the Lateral Boundary Conditions during Summer. *Climate Dynamics* 25 (6): 555–580.

Dominguez, Francina, E. Rivera, D.P. Lettenmaier, and C.L. Castro. 2012. Changes in Winter Precipitation Extremes for the Western United States Under a Warmer Climate as Simulated by Regional Climate Models. *Geophysical Research Letters* 39 (5): 1–7.

Duffy, P.B., B. Govindasamy, J.P. Iorio, J. Milovich, K.R. Sperber, K.E. Taylor, M.F. Wehner, and S.L. Thompson. 2003. High-Resolution Simulations of Global Climate, Part 1: Present Climate. *Climate Dynamics* 21 (5–6): 371–390.

Duffy, P.B., R.W. Arritt, J. Coquard, William Gutowski, J. Han, J. Iorio, Jongil Kim, L.-R. Leung, J. Roads, and E. Zeledon. 2006. Simulations of Present and Future Climates in the Western United States with Four Nested Regional Climate Models. *Journal of Climate* 19 (6): 873–895.

Englehart, Phil J., and Arthur V. Douglas. 2001. The Role of Eastern North Pacific Tropical Storms in the Rainfall Climatology of Western Mexico. *International Journal of Climatology* 21 (11): 1357–1370.

Evans, Jason P., Robert J. Oglesby, and William M. Lapenta. 2005. Time Series Analysis of Regional Climate Model Performance. *Journal of Geophysical Research: Atmospheres* 110: 1–23.

Feser, Frauke, Burkhardt Rockel, Hans von Storch, Jörg Winterfeldt, and Matthias Zahn. 2011. Regional Climate Models Add Value to Global Model Data: A Review and Selected Examples. *Bulletin of the American Meteorological Society* 92 (9): 1181–1192. https://doi.org/10.1175/2011BAMS3061.1.

Foley, A.M. 2010. Uncertainty in Regional Climate Modeling: A Review. *Progress in Physical Geography* 34: 647–670.

Fowler, H.J., S. Blenkinsop, and C. Tebaldi. 2007. Linking Climate Change Modelling to Impacts Studies: Recent Advances in Downscaling Techniques for Hydrological Modelling. *International Journal of Climatology* 27 (12): 1547–1578. https://doi.org/10.1002/joc.1556.

Fox-Rabinovitz, Michael, Jean Côté, Bernard Dugas, Michel Déqué, and John L. McGregor. 2006. Variable Resolution General Circulation Models: Stretched-Grid Model Intercomparison Project (SGMIP). *Journal of Geophysical Research: Atmospheres* 111 (D16). https://doi.org/10.1029/2005JD006520.

Fox-Rabinovitz, Michael, Jean Cote, Bernard Dugas, Michel Deque, John L. McGregor, and A. Belochitski. 2008. Stretched-Grid Model Intercomparison Project: Decadal Regional Climate Simulations with Enhanced Variable and Uniform-Resolution GCMs. *Meteorology and Atmospheric Physics* 100 (1): 159–177.

Franco, Guido, and Alan H. Sanstad. 2008. Climate Change and Electricity Demand in California. *Climatic Change* 87: 139–151.

Gao, Yanhong, Julie A. Vano, Chunmei Zhu, and Dennis P. Lettenmaier. 2011. Evaluating Climate Change over the Colorado River Basin Using Regional Climate Models. *Journal of Geophysical Research: Atmospheres* 116: 1–20.

Gates, W. Lawrence. 1985. The Use of General Circulation Models in the Analysis of the Ecosystem Impacts of Climatic Change. *Climatic Change* 7 (3): 267–284.

Ghan, Steven J., and Timothy Shippert. 2006. Physically Based Global Downscaling: Climate Change Projections for a Full Century. *Journal of Climate* 19 (9): 1589–1604.

Ghan, Steven J., Timothy Shippert, and Jared Fox. 2006. Physically Based Global Downscaling: Regional Evaluation. *Journal of Climate* 19 (3): 429–445.

Giorgi, F., and L.O. Mearns. 1991. Approaches to the Simulation of Regional Climate Change: A Review. *Reviews of Geophysics* 29: 191–216.

Giorgi, F., and L.O. Mearns. 1999. Regional Climate Modeling Revisited: An Introduction to the Special Issue. *Journal of Geophysical Research* 104 (D6): 6335–6352.

Giorgi, Filippo, Maria Rosaria Marinucci, and Gary T. Bates. 1993a. Development of a Second-Generation Regional Climate Model (RegCM2). Part I: Boundary-Layer and Radiative Transfer Processes. *Monthly Weather Review* 121 (10): 2794–2813.

Giorgi, Filippo, Maria Rosaria Marinucci, Gary T. Bates, and Gerardo De Canio. 1993b. Development of a Second-Generation Regional Climate Model (RegCM2). Part II: Convective Processes and Assimilation of Lateral Boundary Conditions. *Monthly Weather Review* 121 (10): 2814–2832.

Giorgi, Filippo, Bruce Hewitson, J. Christensen, Michael Hulme, Hans Von Storch, Penny Whetton, R. Jones, et al. 2001. Regional Climate Information—Evaluation and Projections. *The Scientific Basis, Contribution of Working Group I to the Third Assessment Report of the IPCC* 1: 739–768.

Giorgi, Filippo, Colin Jones, Ghassem R. Asrar, et al. 2009. Addressing Climate Information Needs at the Regional Level: The CORDEX Framework. *World Meteorological Organization (WMO) Bulletin* 58 (3): 175–183.

Giorgi, F., and X. Bi. 2000. A Study of Internal Variability of a Regional Climate Model. *Journal of Geophysical Research* 105: 503–521.

Goldenberg, Stanley B., Christopher W. Landsea, Alberto M. Mestas-Nuñez, and William M. Gray. 2001. The Recent Increase in Atlantic Hurricane Activity: Causes and Implications. *Science* 293 (5529): 474–479.

Govindasamy, Balasubramanian, Philip B. Duffy, and Jeremy Coquard. 2003. High-Resolution Simulations of Global Climate, Part 2: Effects of Increased Greenhouse Cases. *Climate Dynamics* 21 (5–6): 391–404.

Grell, Georg A., Jimy Dudhia, and David R. Stauffer. 1993. A Description of the Fifth-Generation Penn State/NCAR Mesoscale Model (MM5). *NCAR Tech NCAR/TN-398* (1A): 107.

Groisman, Pavel Ya, Thomas R. Karl, David R. Easterling, Richard W. Knight, Paul F. Jamason, Kevin J. Hennessy, et al. 1999. Changes in the Probability of Heavy Precipitation: Important Indicators of Climatic Change. *Climatic Change* 42 (1): 243–283.

Gutmann, Ethan, Tom Pruitt, Martyn P. Clark, Levi Brekke, Jeffrey R. Arnold, David A. Raff, and Roy M. Rasmussen. 2013. Submitted 2014. An Intercomparison of Statistical Downscaling Methods Used for Water Resource Assessments in the United States. *Water Resources Research* 50 (9): 7167–7186.

Gutowski, William J., Stephanie S. Willis, Jason C. Patton, Benjamin R.J. Schwedler, Raymond W. Arritt, and Eugene S. Takle. 2008. Changes in Extreme, Cold-Season Synoptic Precipitation Events Under Global Warming. *Geophysical Research Letters* 35 (20). https://doi.org/10.1029/2008GL035516/full.

Gutzler, David S., and Tessia O. Robbins. 2011. Climate Variability and Projected Change in the Western United States: Regional Downscaling and Drought Statistics. *Climate Dynamics* 37 (5–6): 835–849.

Hall, Alex, Fengpeng Sun, Daniel Walton, Scott Capps, Qu Xin, Hsin-Yuan Huang, Neil Berg, et al. 2012. *Mid-Century Warming in the Los Angeles Region-Part I of the 'Climate Change in the Los Angeles Region' Projec*, 47. Los Angeles: UCLA and LARC.

Han, Jongil, and John O. Roads. 2004. US Climate Sensitivity Simulated with the NCEP Regional Spectral Model. *Climatic Change* 62 (1): 115–154.

Harpham, Colin, and Robert L. Wilby. 2005. Multi-Site Downscaling of Heavy Daily Precipitation Occurrence and Amounts. *Journal of Hydrology* 312 (1): 235–255.

Hawkins, Ed, and Rowan Sutton. 2011. The Potential to Narrow Uncertainty in Projections of Regional Precipitation Change. *Climate Dynamics* 37 (1–2): 407–418. https://doi.org/10.1007/s00382-010-0810-6.

Hay, Lauren E., Steven L. Markstrom, and Christian Ward-Garrison. 2011. Watershed-Scale Response to Climate Change Through the Twenty-First Century for Selected Basins Across the United States. *Earth Interactions* 15 (17): 1–37. https://doi.org/10.1175/2010ei370.1.

Hayhoe, Katharine, Daniel Cayan, Christopher B. Field, Peter C. Frumhoff, Edwin P. Maurer, Norman L. Miller, Susanne C. Moser, et al. 2004. Emissions Pathways, Climate Change, and Impacts on California. *Proceedings of the National Academy of Sciences of the United States of America* 101 (34): 12422–12427. https://doi.org/10.1073/pnas.0404500101.

Hayhoe, Katharine, Cameron P. Wake, Thomas G. Huntington, Lifeng Luo, Mark D. Schwartz, Justin Sheffield, Eric Wood, et al. 2007. Past and Future Changes in Climate and Hydrological Indicators in the US Northeast. *Climate Dynamics* 28 (4): 381–407.

Hayhoe, Katharine, Cameron Wake, Bruce Anderson, Xin-Zhong Liang, Edwin Maurer, Jinhong Zhu, James Bradbury, Art DeGaetano, Anne Marie Stoner, and Donald Wuebbles. 2008. Regional Climate Change Projections for the Northeast USA. *Mitigation and Adaptation Strategies for Global Change* 13 (5–6): 425–436.

Hayhoe, Katharine, Jeff VanDorn, I.I. Thomas Croley, Nicole Schlegal, and Donald Wuebbles. 2010. Regional Climate Change Projections for Chicago and the US Great Lakes. *RES* 36 (2 SI): 7–21. https://doi.org/10.1016/j.jglr.2010.03.012.

Haylock, Malcolm R., Gavin C. Cawley, Colin Harpham, Rob L. Wilby, and Clare M. Goodess. 2006. Downscaling Heavy Precipitation over the United Kingdom: A Comparison of Dynamical and Statistical Methods and Their Future Scenarios. *International Journal of Climatology* 26 (10): 1397–1415. https://doi.org/10.1002/joc.1318.

Horton, Radley, Vivien Gornitz, Malcolm Bowman, and Reginald Blake. 2010. Climate Observations and Projections. *Annals of the New York Academy of Sciences* 1196 (1): 41–62.

Inatsu, Masaru, and Masahide Kimoto. 2009. A Scale Interaction Study on East Asian Cyclogenesis Using a General Circulation Model Coupled with an Interactively Nested Regional Model. *Monthly Weather Review* 137 (9): 2851–2868.

Janjić, Zaviša I. 1994. The Step-Mountain Eta Coordinate Model: Further Developments of the Convection, Viscous Sublayer, and Turbulence Closure Schemes. *Monthly Weather Review* 122 (5): 927–945.

Jha, Manoj, Zaitao Pan, Eugene S. Takle, and Gu Roy. 2004. Impacts of Climate Change on Streamflow in the Upper Mississippi River Basin: A Regional Climate Model Perspective. *Journal of Geophysical Research: Atmospheres* 109 (D9). https://doi.org/10.1029/2003JD003686.

Jiao, Yanjun, and Daniel Caya. 2006. An Investigation of Summer Precipitation Simulated by the Canadian Regional Climate Model. *Monthly Weather Review* 134 (3): 919–932.

Johnson, T.E., J.B. Butcher, A. Parker, and C.P. Weaver. 2012. Investigating the Sensitivity of US Streamflow and Water Quality to Climate Change: US EPA Global Change Research Program's 20 Watersheds Project. *Journal of Water Resources Planning and Management* 138 (5): 453–464.

Jones, R.G., D.C. Hassell, D. Hudson, S.S. Wilson, G.J. Jenkins, and J.F.B. Mitchell. 2004. *Workbook on Generating High Resolution Climate Change Scenarios Using PRECIS*. New York: UNDP.

Juang, Hann-Ming Henry, Song-You Hong, and Masao Kanamitsu. 1997. The NCEP Regional Spectral Model: An Update. *Bulletin of the American Meteorological Society* 78 (10): 2125–2143.

Kanamitsu, Masao, and Laurel DeHaan. 2011. The Added Value Index: A New Metric to Quantify the Added Value of Regional Models. *Journal of Geophysical Research: Atmospheres* 116 (D11). https://doi.org/10.1029/2011JD015597.

Karmalkar, Ambarish V., Raymond S. Bradley, and Henry F. Diaz. 2011. Climate Change in Central America and Mexico: Regional Climate Model Validation and Climate Change Projections. *Climate Dynamics* 37 (3–4): 605–629.

Katz, Richard W., Marc B. Parlange, and Claudia Tebaldi. 2003. Stochastic Modeling of the Effects of Large-Scale Circulation on Daily Weather in the Southeastern US. In *Issues in the Impacts of Climate Variability and Change on Agriculture*, 189–216. Dordrecht: Springer.

Khan, Mohammad Sajjad, Paulin Coulibaly, and Yonas Dibike. 2006. Uncertainty Analysis of Statistical Downscaling Methods. *Journal of Hydrology* 319 (1): 357–382. https://doi.org/10.1016/j.jhydrol.2005.06.035.

Knutson, Thomas R., Joseph J. Sirutis, Stephen T. Garner, Gabriel A. Vecchi, and Isaac M. Held. 2008. Simulated Reduction in Atlantic Hurricane Frequency under Twenty-First-Century Warming Conditions. *Nature Geoscience* 1 (6): 359–364.

Knutson, Thomas R., John L. McBride, Johnny Chan, Kerry Emanuel, Greg Holland, Chris Landsea, Isaac Held, James P. Kossin, A.K. Srivastava, and Masato Sugi. 2010. Tropical Cyclones and Climate Change. *Nature Geoscience* 3 (3): 157–163.

Knutti, R., G. Abramowitz, M. Collins, V. Eyring, P.J. Gleckler, B. Hewitson, et al. 2010. Good Practice Guidance Paper on Assessing and Combining Multi Model Climate Projections. In *Meeting Report of the Intergovernmental Panel on Climate Change Expert Meeting on Assessing and Combining Multi Model Climate Projections*, ed. T.F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, and P.M. Midgley. Bern: IPCC Working Group I Technical Support Unit.

Kunkel, Kenneth E., David R. Easterling, Kenneth Hubbard, and Kelly Redmond. 2004. Temporal Variations in Frost-Free Season in the United States: 1895–2000. *Geophysical Research Letters* 31 (3.) https://doi:03210.01029/02003GL018624.

Landman, Willem A., Mary-Jane Kgatuke, Maluta Mbedzi, Asmerom Beraki, Anna Bartman, and Annelise du Piesanie. 2009. Performance Comparison of Some Dynamical and Empirical Downscaling Methods for South Africa from a Seasonal Climate Modelling Perspective. *International Journal of Climatology* 29 (11): 1535–1549. https://doi.org/10.1002/joc.1766.

Laprise, René. 2008. Regional Climate Modelling. *Journal of Computational Physics* 227 (7): 3641–3666.

Laprise, René, Daniel Caya, Anne Frigon, and Dominique Paquin. 2003. Current and Perturbed Climate as Simulated by the Second-Generation Canadian Regional Climate Model (CRCM-II) over Northwestern North America. *Climate Dynamics* 21 (5–6): 405–421.

Larson, Joshua, Yaping Zhou, and R. Wayne Higgins. 2005. Characteristics of Landfalling Tropical Cyclones in the United States and Mexico: Climatology and Interannual Variability. *Journal of Climate* 18 (8): 1247–1262.

Lawler, Joshua J., Sarah L. Shafer, Denis White, Peter Kareiva, Edwin P. Maurer, Andrew R. Blaustein, and Patrick J. Bartlein. 2009. Projected Climate-Induced Faunal Change in the Western Hemisphere. *Ecology* 90 (3): 588–597.

Leduc, Martin, and René Laprise. 2009. Regional Climate Model Sensitivity to Domain Size. *Climate Dynamics* 32 (6): 833–854.

Lei, F., and Z. Yaocun. 2007. Impacts of the Thermal Effects of Sub-Grid Orography on the Heavy Rainfall Events Along the Yangtze River Valley in 1991. *Advances in Atmospheric Sciences* 24: 881–892.

Leung, Lai R., and Steven J. Ghan. 1998. Parameterizing Subgrid Orographic Precipitation and Surface Cover in Climate Models. *Monthly Weather Review* 126 (12): 3271–3291.

Leung, L.R., and S.J. Ghan. 1999a. Pacific Northwest Climate Sensitivity Simulated by a Regional Climate Model Driven by a GCM. Part II: 2$\times$ CO2 Simulations. *Journal of Climate* 12 (7): 2031–2053.

Leung, Lai R., and Steven J. Ghan. 1999b. Pacific Northwest Climate Sensitivity Simulated by a Regional Climate Model Driven by a GCM. Part I: Control Simulations. *Journal of Climate* 12 (7): 2010–2030.

Leung, L. Ruby, Yun Qian, Xindi Bian, Warren M. Washington, Jongil Han, and John O. Roads. 2004. Mid-Century Ensemble Regional Climate Change Scenarios for the Western United States. *Climatic Change* 62 (1): 75–113.

Li, Xiangshang, and David J. Sailor. 2000. Application of Tree-Structured Regression for Regional Precipitation Prediction Using General Circulation Model Output. *Climate Research* 16: 17–30.

Li, Guilong, Xuebin Zhang, Francis Zwiers, and Qiuzi H. Wen. 2012. Quantification of Uncertainty in High-Resolution Temperature Scenarios for North America. *Journal of Climate* 25 (9): 3373–3389.

Liang, Xin-Zhong, Jianping Pan, Jinhong Zhu, Kenneth E. Kunkel, Julian X.L. Wang, and Aiguo Dai. 2006. Regional Climate Model Downscaling of the US Summer Climate and Future Change. *Journal of Geophysical Research: Atmospheres* 111 (D10). https://doi.org/10.1029/2005JD006685.

Lim, Young-Kwon, D.W. Shin, Steven Cocke, T.E. LaRow, Justin T. Schoof, James J. O'Brien, and Eric P. Chassignet. 2007. Dynamically and Statistically Downscaled Seasonal Simulations of Maximum Surface Air Temperature over the Southeastern United States. *Journal of Geophysical Research: Atmospheres* 112 (D24). https://doi.org/10.1029/2007jd008764.

Liverman, D.M., W.H. Terjung, J.T. Hayes, and L.O. Mearns. 1986. Climatic Change and Grain Corn Yields in the North American Great Plains. *Climatic Change* 9 (3): 327–347.

Lo, J.C.F., Z.L. Yang, and R.A. Pielke. 2008. Assessment of Three Dynamical Climate Downscaling Methods Using the Weather Research and Forecasting (WRF) Model. *Journal of Geophysical Research: Atmospheres* 113 (D9).

Lorenz, Philip, and Daniela Jacob. 2005. Influence of Regional Scale Information on the Global Circulation: A Two-Way Nesting Climate Simulation. *Geophysical Research Letters* 32 (18). https://doi.org/10.1029/2005GL02335.

Lucas-Picher, Philippe, Daniel Caya, Sebastien Biner, and Rene Laprise. 2008. Quantification of the Lateral Boundary Forcing of a Regional Climate Model Using an Aging Tracer. *Monthly Weather Review* 136 (12): 4980–4996.

Lucas-Picher, Philippe, Samuel Somot, Michel Déqué, Bertrand Decharme, and Antoinette Alias. 2013. Evaluation of the Regional Climate Model ALADIN to Simulate the Climate over North America in the CORDEX Framework. *Climate Dynamics* 41 (5–6): 1117–1137.

Magana, Victor, David Zermeño, and Carolina Neri. 2012. Climate Change Scenarios and Potential Impacts on Water Availability in Northern Mexico. *Climate Research* 51 (2): 171–184.

Mailhot, Alain, Sophie Duchesne, Daniel Caya, and Guillaume Talbot. 2007. Assessment of Future Change in Intensity–Duration–Frequency (IDF) Curves for Southern Quebec Using the Canadian Regional Climate Model (CRCM). *Journal of Hydrology* 347 (1): 197–210.

Mailhot, Alain, Ian Beauregard, Guillaume Talbot, Daniel Caya, and Sébastien Biner. 2012. Future Changes in Intense Precipitation over Canada Assessed from Multi-Model NARCCAP Ensemble Simulations. *International Journal of Climatology* 32 (8): 1151–1163.

Manning, L.J., J.W. Hall, H.J. Fowler, C.G. Kilsby, and C. Tebaldi. 2009. Using Probabilistic Climate Change Information from a Multimodel Ensemble for Water Resources Assessment. *Water Resources Research* 45 (11). https://doi.org/10.1029/2007wr006674.

Maraun, Douglas, F. Wetterhall, A.M. Ireson, R.E. Chandler, E.J. Kendon, M. Widmann, S. Brienen, et al. 2010. Precipitation Downscaling under Climate Change: Recent Developments to Bridge the Gap Between Dynamical Models and the End User. *Reviews of Geophysics* 48 (3). https://doi.org/10.1029/2009rg000314.

Martínez-Castro, D., R. Porfirio da Rocha, A. Bezanilla-Morlot, L. Alvarez-Escudero, J.P. Reyes-Fernández, Y. Silva-Vidal, and R.W. Arritt. 2006. Sensitivity Studies of the RegCM3 Simulation of Summer Precipitation, Temperature and Local Wind Field in the Caribbean Region. *Theoretical and Applied Climatology* 86 (1): 5–22.

Martynov, Andrey, René Laprise, Laxmi Sushama, Katja Winger, L. Šeparović, and B. Dugas. 2013. Reanalysis-Driven Climate Simulation over CORDEX North America Domain Using the Canadian Regional Climate Model, Version 5: Model Performance Evaluation. *Climate Dynamics* 41 (11–12): 2973–3005.

Maurer, E.P., and H.G. Hidalgo. 2008. Utility of Daily vs. Monthly Large-Scale Climate Data: An Intercomparison of Two Statistical Downscaling Methods. *Hydrology and Earth System Sciences* 12 (2): 551–563.

Maurer, Edwin P., Levi Brekke, Tom Pruitt, and Philip P. Duffy. 2007. Fine-Resolution Climate Projections Enhance Regional Climate Change Impact Studies. *Eos* 88 (47): 504.

May, W., and E. Roeckner. 2001. A Time Slice-Experiment with the EC HAM4 AGCM at High Resolution: The Impact of Horizontal Resolution on Annual Mean Climate Change. *Climate Dynamics* 17: 407–420.

McGregor, J.L., and M.R. Dix. 2008. An Updated Description of the Conformal-Cubic Atmospheric Model. In *High Resolution Simulation of the Atmosphere and Ocean,* eds. K Hamilton, W Ohfuchi, 51–76. Springer, New York.

Mearns, L.O., M. Hulme, T.R. Carter, R. Leemans, M. Lal, and P. Whetton. 2001. Climate Scenario Development (Chapter 13). In *Climate Change 2001: The Scientific Basis, Contribution of Working Group I to the Third Assessment Report of the IPCC*, ed. J.T. Houghton et al., 583–638. Cambridge: Cambridge University Press.

Mearns, L.O., G. Carbone, E. Tsvetsinskaya, R. Adams, B. McCarl, and R. Doherty. 2003. The Uncertainty of Spatial Scale of Climate Scenarios in Integrated Assessments: An Example from Agriculture. *Integrated Assessment* 4 (4): 225–235.

Mearns, L.O., W.J. Gutowski, R. Jones, L.-Y. Leung, S. McGinnis, A.M.B. Nunes, et al. 2009. A Regional Climate Change Assessment Program for North America. *Eos* 90: 311–312.

Mearns, L.O., R. Arritt, S. Biner, M.S. Bukovsky, S. McGinnis, S. Sain, et al. 2012. The North American Regional Climate Change Assessment Program: Overview of Phase I Results. *Bulletin of the American Meteorological Society* 93: 1337–1362.

Mearns, L.O., R. Leung, R. Arritt, S. Biner, M. Bukovsky, D. Caya, J. Correia, W. Gutowski, R. Jones, Y. Qian, L. Sloan, and G. Takle. 2013a. Response to R. Pielke, Sr. Commentary on Mearns et al. 2012. *Bulletin of the American Meteorological Society* 94: 1077–1078.

Mearns, L.O., S. Sain, R. Leung, M. Bukovsky, et al. 2013b. Climate Change Projections of the North American Regional Climate Change Assessment Program (NARCCAP). *Climatic Change Letters*. https://doi.org/10.1007/s10584-013-0831-3.

Miguez-Macho, G., G.L. Stenchikov, and A. Robock. 2005. Regional Climate Simulations Over North America: Interaction of Local Processes with Improved Large-Scale Flow. *Journal of Climate* 18: 1227–1246.

Mladjic, B., L. Sushama, M.N. Khaliq, R. Laprise, D. Caya, and R. Roy. 2011. Canadian RCM Projected Changes to Extreme Precipitation Characteristics Over Canada. *Journal of Climate* 24: 2565–2584.

Montero-Martínez, M., and J.L. Pérez-López. 2008. Regionalización de proyecciones climáticas en México de precipitación y temperatura en superficie usando el método REA para el siglo XXI. In *Efectos del Cambio Climático en los Recursos Hídricos de México*, ed. P. Martínez and A. Aguilar, vol. 2, 11–21. Jiutepec: Instituto Mexicano de Tecnología del Agua.

Muluye, G.Y. 2011. Implications of Medium-Range Numerical Weather Model Output in Hydrologic Applications: Assessment of Skill and Economic Value. *Journal of Hydrology* 400 (3–4): 448–464.

NAS. 2012. *A National Strategy for Advancing Climate Modeling*. Washington, DC: National Academies Press.

Noguer, M., R.G. Jones, and J.M. Murphy. 1998. Sources of Systematic Errors in the Climatology of a Nested Regional Climate Model (RCM) Over Europe. *Climate Dynamics* 14: 691–712.

Otte, T.L., C.G. Nolte, M.J. Otte, and J.H. Bowden. 2012. Does Nudging Squelch the Extremes in Regional Climate Modeling? *Journal of Climate* 25: 7046–7066.

Pal, J.S., F. Giorgi, X. Bi, N. Elguindi, F. Solmon, S. Rauscher, X. Gao, R. Francisco, A. Zakey, J. Winter, M. Ashfaq, et al. 2007. Regional Climate Modeling for the Developing World—The ICTP RegCM3 and RegCNET. *Bulletin of the American Meteorological Society* 88: 1395–1409.

Patricola, C.M., and K.H. Cook. 2010. Northern African Climate at the End of the Twentyfirst Century: Integrated Application of Regional and Global Climate Models. *Climate Dynamics* 35: 193–212.

Patricola, C.M., and K.H. Cook. 2013. Mid-Twenty-First Century Warm Season Climate Change in the Central United States. Part 1: Regional and Global Model Predictions. *Climate Dynamics* 40: 551–568.

Pan, Z., R.W. Arritt, E.S. Takle, W.J. Gutkowski Jr., C.J. Anderson, and M. Segal. 2004. Altered Hydrologic Feedback in a Warming Climate Introduces a "Warming Hole". *Geophysical Research Letters* 31: L17109. https://doi.org/10.1029/2004GL020528.

Pan, L.-L., S.-H. Chen, D. Cayan, M.-Y. Lin, Q. Hart, M.-H. Zhang, et al. 2010. Influences of Climate Change on California and Nevada Regions Revealed by a High-Resolution Dynamical Downscaling Study. *Climate Dynamics*. https://doi.org/10.1007/s00382-010-0961-5.

Payne, J.T., A.W. Wood, A.F. Hamlet, R.N. Palmer, and D.P. Lettenmaier. 2004. Mitigating the Effects of Climate Change on the Water Resources of the Columbia River Basin. *Climatic Change* 62: 233–256.

Pérez-Pérez, E., M. Méndez, and V. Magaña. 2007. High Spatial Resolution Climate Change Scenarios for Mexico Based on Experiments Conducted with the Earth Simulator. In *Visualizing Future Climate in Latin America: Results from the Application of the Earth Simulator,* W. Vergara (Coordinator). Latin America and Caribbean Region Sustainable Development Working Paper 30, the World Bank, p. 90.

Pielke, R.A., W.R. Cotton, R.L. Walko, C.J. Trembeck, W.A. Lyons, L.D. Grasso, M.E. Nicholls, M.D. Moran, D.A. Wesley, T.J. Lee, and J.H. Copeland. 1992. A Comprehensive Meteorological Modeling System? RAMS. *Meteorology and Atmospheric Physics* 49: 65–78.

Pitman, A.J., A. Arneth, and L. Ganzeveld. 2010. Regionalizing Global Climate Models. *International Journal of Climatology*. https://doi.org/10.1002/joc.2279.

Prömmel, K., B. Geyer, J.M. Jones, and M. Widmann. 2010. Evaluation of the Skill and Added Value of a Reanalysis-Driven Regional Simulation for Alpine Temperature. *International Journal of Climatology* 30: 760–773.

Pryor, S.C., and J.T. Schoof. 2010. Importance of the SRES in Projections of Climate Change Impacts on Near-Surface Wind Regimes. *Meteorologische Zeitschrift* 19 (3): 267–274.

Pryor, S.C., J.T. Schoof, and R.J. Barthelmie. 2006. Winds of Change: Projections of Near-Surface Winds Under Climate Change Scenarios. *Geophysical Research Letters* 33: L11702. https://doi.org/10.1029/2006GL026000.

Pryor, S.C., J.A. Howe, and K.E. Kunkel. 2009. How Spatially Coherent and Statistically Robust Are Temporal Changes in Extreme Precipitation in the Contiguous USA? *International Journal of Climatology* 29 (1): 31–45. https://doi.org/10.1002/joc.1696.

Pryor, S.C., and R.J. Barthelmie. 2010. Climate Change Impacts on Wind Energy: A Review. *Renewable and Sustainable Energy Reviews* 14: 430–437.

Pryor, S.C., R.J. Barthelmie, and J.T. Schoof. 2012a. Past and Future Wind Climates Over the Contiguous USA Based on the NARCCAP Model Suite. *Journal of Geophysical Research* 117: D19119. https://doi.org/10.1029/2012JD017449.

Pryor, S.C., G. Nikulin, and C. Jones. 2012b. Influence of Spatial Resolution on Regional Climate Model Derived Wind Climates. *Journal of Geophysical Research* 117: D03117. https://doi.org/10.1029/2011JD016822.

Pryor, S.C., R.J. Barthelmie, and J.T. Schoof. 2013. High-Resolution Projections of Climate- Related Risks for the Midwestern USA. *Climate Research* 56: 61–79.

Qian, B., S. Gameda, and H. Hayhoe. 2008. Performance of Stochastic Weather Generators LARS-WG and AAFC-WG for Reproducing Daily Extremes of Diverse Canadian Climates. *Climate Research* 37 (1): 17–33. https://doi.org/10.3354/cr00755.

Qian, B.D., S. Gameda, R. de Jong, P. Falloon, and J. Gornall. 2010. Comparing Scenarios of Canadian Daily Climate Extremes Derived Using a Weather Generator. *Climate Research* 41 (2): 131–149. https://doi.org/10.3354/cr00845.

Racherla, P.N., D.T. Shindell, and G.S. Faluvegi. 2012. The Added Value to Global Model Projections of Climate Change by Dynamical Downscaling: A Case Study Over the Continental US Using the GISS-ModelE2 and WRF Model. *Journal of Geophysical Research* 117:D20118, pp. 3015–3048.

Radu, R., M. Déqué, and S. Somot. 2008. Spectral Nudging in a Spectral Regional Climate Model. *Tellus* 60: 898–910.

Rasmussen, R., C. Liu, K. Ikeda, D. Gochis, D. Yates, F. Chen, M. Tewari, M. Barlage, J. Dudhia, W. Yu, K. Miller, et al. 2011. High Resolution Coupled Climate-Runoff Simulations of Seasonal Snowfall Over Colorado: A Process Study of Current and Warmer Climate. *Journal of Climate* 24: 3015–3048.

Rauscher, S.A., A. Seth, J.-H. Qian, and S.J. Camargo. 2006. Domain Choice in an Experimental Nested Modeling Prediction System for South America. *Theoretical and Applied Climatology* 86: 229–246.

Rauscher, S.A., F. Giorgi, N.S. Diffenbaugh, and A. Seth. 2008. Extension and Intensification of the Meso-American Mid-Summer Drought in the Twenty-First Century. *Climate Dynamics*. https://doi.org/10.1007/s0038-007-0359-1.

Rauscher, S.A., F. Kucharski, and D.B. Enfield. 2011. The Role of Regional SST Warming Variations in the Drying of Meso-America in Future Climate Projections. *Journal of Climate* 24: 2003–2016.

Rawlins, M.A., R.S. Bradley, and H.F. Diaz. 2012. Assessment of Regional Climate Model Simulation Estimates Over the Northeast United States. *Journal of Geophysical Research* 117: D23112. https://doi.org/10.1029/2012JD018137.

Ritchie, E.A., K.M. Wood, D.S. Gutzler, and S.R. White. 2011. The Influence of Eastern Pacific Tropical Cyclone Remnants on the Southwestern United States. *Monthly Weather Review* 139: 192–210. https://doi.org/10.1175/2010MWR3389.1.

Rosenberg, E.A., P.W. Keys, D.B. Booth, D. Hartley, J. Burkey, A.C. Steinemann, et al. 2010. Precipitation Extremes and the Impacts of Climate Change on Stormwater Infrastructure in Washington State. *Climatic Change* 102: 319–349.

Rosenzweig, C. 1985. Potential CO2-Induced Climate Effects on North American Wheatproducing Regions. *Climatic Change* 7: 367–389. https://doi.org/10.1007/BF00139053.

Rummukainen, M. 2010. State-of-the-Art with Regional Climate Models. *WIRE Advanced Review* 1 (1): 82–96.

Rupp, T.S., X. Chen, M. Olson, and A.D. McGuire. 2007. Sensitivity of Simulated Boreal Fire Dynamics to Uncertainties in Climate Drivers. *Earth Interactions* 11 (3): 1–21.

Salathé, E.P. 2006. Influences of a Shift in North Pacific Storm Tracks on Western North American Precipitation Under Global Warming. *Geophysical Research Letters* 33 (19). https://doi.org/10.1029/2006gl026882.

Salathé, E.P., Jr., R. Steed, C.F. Mass, and P.H. Zahn. 2008. A High-Resolution Climate Model for the US Pacific Northwest: Mesoscale Feedbacks and Local Responses to Climate Change. *Journal of Climate* 21: 5708–5726.

Salathé, E.P., Jr., L.R. Leung, Y. Qian, and Y. Zhang. 2010. Regional Climate Model Projections for the State of Washington. *Climatic Change* 102: 51–75.

Schoof, J.T. 2009. Historical and Projected Changes in the Length of the Frost-Free Season. In *Understanding Climate Change: Climate Variability, Predictability and Change in the Midwestern United States*, ed. S.C. Pryor, 42–54. Bloomington: Indiana University Press.

Schoof, J.T. 2012. Historical and Projected Changes in Human Heat Stress in the Midwestern USA. In *Understanding Climate Change: Climate Change Impacts, Risks, Vulnerability and Adaptation in the Midwestern United States*, ed. S.C. Pryor, 146–156. Bloomington: Indiana University Press.

Schoof, J.T., and S.C. Pryor. 2001. Downscaling Temperature and Precipitation: A Comparison of Regression-Based Methods and Artificial Neural Networks. *International Journal of Climatology* 21: 773–790.

Schoof, J.T., and S.C. Pryor. 2008. On the Proper Order of Markov Chain Model for Daily Precipitation Occurrence in the Contiguous United States. *Journal of Applied Meteorology and Climatology* 47: 2477–2486.

Schoof, J.T., S.C. Pryor, and S.M. Robeson. 2007. Downscaling Daily Maximum and Minimum Temperatures in the Midwestern USA: A Hybrid Empirical Approach. *International Journal of Climatology* 27 (4): 439–454. https://doi.org/10.1002/joc.1412.

Schoof, J.T., S.C. Pryor, and J. Suprenant. 2010. Development of Daily Precipitation Projections for the United States Based on Probabilistic Downscaling. *Journal of Geophysical Research* 115 (D13). https://doi.org/10.1029/2009JD013030.

Schoof, J.T., D.W. Shin, S. Cocke, T.E. LaRow, Y.K. Lim, and J.J. O'Brien. 2009. Dynamically and Statistically Downscaled Seasonal Temperature and Precipitation Hindcast Ensembles for the Southeastern USA. *International Journal of Climatology* 29 (2): 243–257. https://doi.org/10.1002/joc.1717.

Semenov, M.A., R.J. Brooks, E.M. Barrow, and C.W. Richardson. 1998. Comparison of the WGEN and LARS-WG Stochastic Weather Generators for Diverse Climates. *Climate Research* 10 (2): 95–107.

Separovic, L., R. de Elía, and R. Laprise. 2011. Impact of Spectral Nudging and Domain Size in Studies of RCM Response to Parameter Modification. *Climate Dynamics*. https://doi.org/10.1007/s00382-011-1072-7.

Shepherd, A., K.M. Gill, and S.B. Rood. 2010. Climate Change and Future Flows of Rocky Mountain Rivers: Converging Forecasts from Empirical

Trend Projection and Down-Scaled Global Circulation Modelling. *Hydrological Processes* 24 (26): 3864–3877. https://doi.org/10.1002/hyp.7818.

Sheridan, S.C., and C.C. Lee. 2010. Synoptic Climatology and the General Circulation Model. *Progress in Physical Geography* 34 (1): 101–109. https://doi.org/10.1177/0309133309357012.

Skamarock, W.C., J.B. Klemp, J. Dudhia, D.O. Gill, D.M. Barker, W. Wang, and et al. 2005. A Description of the Advanced Research WRF Version 2. In NCAR Technical Note NCAR/TN-468 + STR.

Skamarock, W.C., J.B. Klemp, M. Duda, L. Fowler, and S-H. Park. 2010. Global Nonhydrostatic Modeling Using Voronoi Meshes: The MPAS Model. In Proceedings of the ECMWF Workshop on Nonhydrostatic Modelling, European Center for Medium Range Forecasting, Reading, 8–10 Nov 2010. http://www.ecmwf.int/newsevents/meetings/workshops/2010/Non_hydrostatic_Modelling/presentations/Skamarock.pdf

Skamarock, W.C., J.B. Klemp, M.G. Duda, L.D. Fowler, and S.-H. Park. 2012. A Multiscale Nonhydrostatic Atmospheric Model Using Centroidal Voronoi Tessellations and C-Grid Staggering. *Monthly Weather Review* 140: 3090–3105.

Snyder, M.A., and L.C. Sloan. 2005. Transient Future Climate Over the Western United States Using a Regional Climate Model. *Earth Interactions* 9: 1–21.

Solomon, S., ed. 2007. *Climate Change 2007-the Physical Science Basis: Working Group I Contribution to the Fourth Assessment Report of the IPCC*. Vol. 4. Cambridge: Cambridge University Press.

Spak, S., T. Holloway, B. Lynn, and R. Goldberg. 2007. A Comparison of Statistical and Dynamical Downscaling for Surface Temperature in North America. *Journal of Geophysical Research-Atmospheres* 112 (D8). https://doi.org/10.1029/2005jd006712.

Stoll, S., H.J.H. Franssen, M. Butts, and W. Kinzelbach. 2011. Analysis of the Impact of Climate Change on Groundwater Related Hydrological Fluxes: A Multi-Model Approach Including Different Downscaling Methods. *Hydrology and Earth System Sciences* 15 (1): 21–38. https://doi.org/10.5194/hess-15-21-2011.

Sushama, L., R. Laprise, D. Caya, A. Frigon, and M. Slivitzky. 2006. Canadian RCM Projected Climate Change Signal and Its Sensitivity to Model Errors. *International Journal of Climatology* 26: 2141–2159.

Sushama, L., R. Laprise, D. Caya, D. Versghy, and M. Allard. 2007. An RCM Projection of Soil Thermal and Moisture Regimes for North American

Permafrost Zones. *Geophysical Research Letters* 34: L20711. https://doi.org/10.1029/2007GL031385.

Timm, O., and H.F. Diaz. 2009. Synoptic-Statistical Approach to Regional Downscaling of IPCC Twentyfirst-Century Climate Projections: Seasonal Rainfall Over the Hawaiian Islands. *Journal of Climate* 22 (16): 4261–4280. https://doi.org/10.1175/2009jcli2833.1.

Tryhorn, L., and A. DeGaetano. 2011. A Comparison of Techniques for Downscaling Extreme Precipitation Over the Northeastern United States. *International Journal of Climatology* 31 (13): 1975–1989.

Vannitsem, S., and F. Chomé. 2005. One-Way Nested Regional Climate Simulations and Domain Size. *Journal of Climate* 18: 229–233.

Vrac, M., M. Stein, and K. Hayhoe. 2007. Statistical Downscaling of Precipitation Through Nonhomogeneous Stochastic Weather Typing. *Climate Research* 34: 169–184. https://doi.org/10.3354/cr00696.

Wang, J.F., and X.B. Zhang. 2008. Downscaling and Projection of Winter Extreme Daily Precipitation Over North America. *Journal of Climate* 21 (5): 923–937. https://doi.org/10.1175/2007jcli1671.1.

Warner, T.T. 2011. *Numerical Weather and Climate Prediction*. Cambridge, UK: Cambridge University Press.

Wehner, M.F., G. Bala, P. Duffy, A.A. Mirin, and R. Romano. 2010. Towards Direct Simulation of Future Tropical Cyclone Statistics in a High-Resolution Global Atmospheric Model. *Advances in Meteorology*. https://doi.org/10.1155/2010/915303.

Wetterhall, F., F. Pappenberger, Y. He, J. Freer, and H.L. Cloke. 2012. Conditioning Model Output Statistics of Regional Climate Model Precipitation on Circulation Patterns. *Nonlinear Processes in Geophysics* 19 (6): 623–633.

White, M.R. 1985. *Characterization of Information Requirements for Studies of CO2 Effects: Water Resources, Agriculture, Fisheris, Forests, and Human Health*. Washington, DC: US Department of Energy, Office of Energy Research. DOE/ER-0236.

Wi, Sungwook, Francina Dominguez, Matej Durcik, Juan Valdes, Henry F. Diaz, and Christopher L. Castro. 2012. Climate Change Projection of Snowfall in the Colorado River Basin Using Dynamical Downscaling. *Water Resources Research* 48 (5): W05504. https://doi.org/10.1029/2011WR010674.

Wilby, R.L., and H.J. Fowler. 2012. Regional Climate Downscaling. In *Modelling the Impact of Climate Change on Water Resources*, ed. C.F. Fung, A. Lopez, and M. New, 34–85. Chichester: Wiley-Blackwell.

Wilby, R.L., and T.M.L. Wigley. 1997. Downscaling General Circulation Model Output: A Review of Methods and Limitations. *Progress in Physical Geography* 21: 530–548.

Wilby, R.L., and T.M.L. Wigley. 2000. Precipitation Predictors for Downscaling: Observed and General Circulation Model Relationships. *International Journal of Climatology* 20: 641–661. https://doi.org/10.1002/(SICI)1097-0088(200005)20:6<641::AID-JOC501>3.3.CO;2-T.

Wilby, R.L., J. Troni, Y. Biot, L. Tedd, B.C. Hewitson, D.M. Smith, and R.T. Sutton. 2009. A Review of Climate Risk Information for Adaptation and Development Planning. *International Journal of Climatology* 29: 1193–1215. https://doi.org/10.1002/joc.1839.

Wilks, D.S. 2012. Stochastic Weather Generators for Climate-Change Downscaling, Part II: Multivariable and Spatially Coherent Multisite Downscaling. *Wiley Interdisciplinary Reviews: Climate Change* 3 (3): 267–278.

Wood, A.W., E.P. Maurer, A. Kumar, and D.P. Lettenmaier. 2002. Long Range Experimental Hydrologic Forecasting for the Eastern United States. *Journal of Geophysical Research* 107 (D20): ACL6.1–ACL6.15.

Wood, A.W., L.R. Leung, V. Sridhar, and D.P. Lettenmaier. 2004. Hydrologic Implications of Dynamical and Statistical Approaches to Downscaling Climate Model Outputs. *Climatic Change* 62 (1–3): 189–216.

Xue, Y., R. Vasic, Z. Janjic, F. Mesinger, and K.E. Mitchell. 2007. Assessment of Dynamic Downscaling of the Continental US Regional Climate Using the Eta/SSiB Regional Climate Model. *Journal of Climate* 20: 4172–4193.

Yohe, G., and M. Oppenheimer. 2011. Evaluation, Characterization, and Communication of Uncertainty by the Intergovernmental Panel on Climate Change—An Introductory Essay. *Climatic Change 108* (4): 629.

Yoon, J.-H., L.R. Leung, and J. Correia Jr. 2012. Comparison of Dynamically and Statistically Downscaled Seasonal Climate Forecasts for the Cold Season Over the United States. *Journal of Geophysical Research-Atmospheres* 117 (D21). https://doi.org/10.1029/2012JD017650.

Part II

Uncertainties and Robustness

9

The Significance of Robust Climate Projections

Wendy S. Parker

9.1 Introduction

There is now a broad scientific consensus—underwritten by a substantial and growing body of evidence—that Earth's climate has warmed significantly over the last century, that increased atmospheric concentrations of greenhouse gases due to human activities are a major cause of this warming, and that Earth's climate will be still warmer by the end of the twenty-first century (Solomon et al. 2007; IPCC 2013). Less clear are the quantitative details, especially regarding future climate change. How much will Earth's average surface temperature increase by the end of the twenty-first century if greenhouse gas concentrations continue rising as they have in recent decades? Under that scenario, will the central United States experience much drier summers as the century unfolds? What will climatic conditions in various locales be like late in the twenty-first century if instead greenhouse gas concentrations are stabilized at 450 ppm?

W.S. Parker (✉)
Department of Philosophy, Durham University, Durham, UK

E.A. Lloyd, E. Winsberg (eds.), *Climate Modelling*,
https://doi.org/10.1007/978-3-319-65058-6_9

Current scientific understanding suggests that answers to questions like these, about long-term changes in global and regional climate, may depend on the details of complex interactions among many climate system processes—details that cannot be tracked without the help of computer simulation models. Numerous simulation models have been developed, differing in their spatiotemporal resolution, the range of climate system processes that they take into account, and the ways in which they represent those processes. When collections—or *ensembles*—of these models are used to simulate future climate, it sometimes happens that the models all (or nearly all) agree regarding some interesting predictive hypothesis.[1] For instance, two dozen state-of-the-art climate models might agree that, under a particular greenhouse gas emission scenario, Earth's average surface temperature in the 2090s would be more than 2 °C warmer than it was in the 1890s.[2] These agreed-upon or *robust* findings are sometimes highlighted in papers and reports on climate change, but what exactly is their significance?[3] For instance, are they likely to be true?

Such questions have sparked debate before, outside of the context of climate prediction. Orzack and Sober (1993) argued that the derivation of a result from multiple models (i.e. robustness) does not on its own constitute evidence that the agreed-upon result is true. They attributed the opposite view to Levins (1966), who, in his discussion of robustness in modeling in biology, had concluded that "our truth is the intersection of independent lies" (p. 20). In response, Levins (1993) and Weisberg (2006) did not dispute that robustness alone does not warrant inferences about real-world systems but emphasized that in practice such inferences are made in light of robustness *plus* empirical considerations. Weisberg (2006) also offered a more detailed account of robustness analysis—the procedure (or family of procedures) that investigates whether a result is derivable from each of a set of carefully chosen models. His account makes clear that the goal of robustness analysis is sometimes the identification of causes of phenomena, rather than the advance prediction of their occurrence. Recently, Lloyd (2010, 2015) has applied and expanded Weisberg's account in the context of climate modeling; she argues that robustness analysis here supports the conclusion that twentieth-century global warming was caused at least in part by increased greenhouse gas concentrations.

The present chapter returns the focus to the predictive use of models; it is concerned not with the causes of already observed phenomena but with the future occurrence (or not) of phenomena/events predicted by a set of models. More specifically, it is concerned with the conditions under which robustness should influence our expectations about the occurrence of such phenomena/events. The aim is to identify some of these conditions and to consider whether they are met in the context of ensemble climate prediction today. In doing so, the chapter further articulates the sort of empirical background knowledge that (as the aforementioned authors have in different ways suggested) is needed for robust predictive modeling results to take on special epistemic significance.

Section 9.2 gives a brief introduction to ensemble climate prediction, explaining how and why multiple models are used to investigate future climate change. The next three sections investigate the prospects for inferring from robust modeling results, and from robust climate modeling results in particular, that:

- an agreed-upon predictive hypothesis *H* is likely to be true (Sect. 9.3);
- significantly increased confidence in *H* is warranted (Sect. 9.4);
- the security of a claim to have evidence for *H* is enhanced (Sect. 9.5).

The findings for climate modeling are disappointing. When today's climate models agree that an interesting hypothesis about future climate change is true, it cannot be inferred—via the arguments considered here anyway—that the hypothesis is likely to be true, nor that scientists' confidence in the hypothesis should be significantly increased, nor that a claim to have evidence for the hypothesis is now more secure. In closing, Sect. 9.6 reflects upon these findings and suggests that the prospects may be brighter in some other modeling contexts.

9.2 Ensemble Climate Prediction

A computer simulation model is a computer-implemented set of instructions for repeatedly solving a set of equations in order to produce a representation of the temporal evolution of selected properties of a target

system. In the case of global climate modeling, the target system is Earth's climate system—encompassing the atmosphere, oceans, sea ice, and land surface—and the equations are ones that describe in an approximate way the local rate of change of temperature, wind speed, humidity, and other quantities of interest in response to myriad processes at work in the system. When it comes to formulating such equations, considerable uncertainty remains for several reasons. Although a theory of large-scale atmospheric dynamics (grounded in fluid dynamics) has long been in place and provides the foundation for some parts of today's climate models, some other important climate system processes are less well understood. In addition, for processes that are believed to influence climate in important ways but that occur on scales finer than those resolved in today's models (i.e. on spatial scales smaller than ~100 km in the horizontal and/or on time scales shorter than ~½ hour), rough representations in terms of larger-scale variables must be developed, and it is rarely obvious how this can best be done. The upshot is that multiple climate models, which differ in various ways in their equations and in the methods used to estimate solutions to them, are nevertheless judged to have approximately equal prima facie plausibility as tools for projecting future climate change (Parker 2006).[4] Indeed, even after examining how well these different models simulate past and present climate, it is often unclear which would be best for a given predictive task.[5]

Given this uncertainty, how should climate scientists proceed? If it is unclear which of several models will turn out to give the best projection in a particular case, then it would be unwise to select just one of the models and rely on its projection, unless all of the models are expected to be so accurate that any would be good enough. Since the latter cannot be expected of today's climate models, ensemble studies present a better option. These studies involve running each of several climate models (or model versions) with the same (or similar) initial conditions and under the same (or similar) future emission scenarios (see e.g. Stainforth et al. 2005; Tebaldi et al. 2005; Murphy et al. 2007). Ensemble studies acknowledge that there is uncertainty about how to represent the climate system; they explore how much this uncertainty matters when it comes to predictions of interest.

There are two main types of ensemble climate prediction study today. *Multi-model ensemble studies* produce simulations of future climate using

models that differ in a number of ways—in the form of some of their equations, in some of their parameter values, and often in their spatio-temporal resolution, their solution algorithms, and their computing platforms as well. A typical multi-model study requires the participation of research groups at various modeling centers around the world, each running their in-house models on local supercomputers, and delivers a total of a few dozen simulations of future climate under a given emission scenario (see e.g. Meehl et al. 2007; Collins et al. 2013). *Perturbed-physics ensemble studies* employ multiple versions of a single climate model whose best parameter values remain uncertain. The model is run repeatedly, leaving the structure of its equations unchanged, but allowing its uncertain parameters to take different values on each run. The selection of these parameter values can be made using formal sampling methods or in more informal ways; usually values are chosen from a range identified by expert judgment. A single perturbed physics study may produce a large number of simulations of future climate, depending on how computationally intensive it is to run a single simulation. Studies carried out by the climateprediction.net project, for example, relied on donated idle processing time on ordinary home computers to produce thousands of simulations using different versions of a relatively complex climate model (Stainforth et al. 2005; BBC 2010).

The discussion that follows will focus on results from multi-model ensemble studies. This is because perturbed-physics studies typically explore such a broad range of parameter values that they deliver a very wide range of results—so wide that the results are not in unanimous (or even near unanimous) agreement regarding interesting predictive hypotheses. It tends to be multi-model ensemble studies, rather, in which such agreement occurs. For instance, the Fourth Assessment Report of the Intergovernmental Panel on Climate Change presented the result of a multi-model study that investigated a "high" emission scenario using 17 state-of-the-art climate models; each of the models indicated that, by 2050, global mean surface temperature would be between 1 °C and 2 °C warmer than during the period 1980–1999 (see Meehl et al. 2007, 763).[6] Likewise, virtually all of the models agreed that, under a "medium" emission scenario, summer rainfall in east Africa would be greater in the late twenty-first century than it was in the late

twentieth century (Christensen et al. 2007, 869). The question is whether agreed-upon multi-model results like these have special epistemic significance and, if so, what that significance is.

9.3 Robustness and Truth

Are robust projections from today's multi-model ensembles likely to be true? More generally, under what conditions can an inference from robustness to likely truth be justified? Consider the following argument, inspired by the discussions of Orzack and Sober (1993) and Woodward (2006):

(1a) It is likely that one model in this collection is true.
(1b) Each of the models in this collection entails hypothesis *H*.
∴ It is likely that *H*.

While its logic is unobjectionable, this argument seems mostly inapplicable in science. Insofar as a scientific model can be identified with a complex hypothesis about the workings of a target system, there is usually good reason to believe that such a hypothesis is (strictly) false, since most scientific models are known from the outset to involve idealizations, simplifications, and/or outright fictions. So (1a) will rarely hold.[7]

But models that incorporate false assumptions might nevertheless be expected to produce simulations that indicate correctly regarding various hypotheses of interest.[8] Thus a similar argument with greater potential for applicability is as follows:

(2a) It is likely that at least one simulation in this collection indicates correctly regarding *H*.
(2b) Each of the simulations in this collection indicates the truth of *H*.
∴ It is likely that *H*.

The question is then whether there is good reason to think that (2a) holds for interesting hypotheses about future climate change that today's multi-model ensembles indicate to be true. At least two approaches to justifying (2a) might be pursued.

A first approach would focus on the extent to which the collection of models samples current uncertainty about how to represent the climate system (for purposes of accurately predicting the quantity of interest). If it samples enough of this representational uncertainty, or samples it in the right way, this could justify (2a). But such a claim about sampling cannot be made for today's multi-model ensembles, which are "ensembles of opportunity"—assembled from existing climate models and "not designed to span an uncertainty range" (Knutti et al. 2008, 2653; see also Tebaldi and Knutti 2007; Meehl et al. 2007). Indeed, when it comes to discerning the truth/falsity of quantitative hypotheses about long-term climate change, climate scientists today are not in a position to specify a small set of models that can be expected to include at least one whose projections are highly accurate. In part, this is because it remains unclear whether processes and feedbacks that will significantly shape long-term climate change have been overlooked (so-called "unknown unknowns"). But it also reflects the challenge of anticipating how recognized simplifications, approximations, and omissions will impact the accuracy of predictions produced by complex, nonlinear models for forcing conditions unlike those previously experienced (see also Parker 2009).

A second approach to justifying (2a) would view an ensemble as a tool for indicating the truth/falsity of hypotheses of a particular sort, of which the predictive hypothesis *H* is an instance; the ensemble's past reliability with respect to *H*-type hypotheses would be cited as evidence that it is likely that at least one of its simulations is indicating correctly regarding this particular *H*.[9] Assuming that *H* concerns the value of a given variable, this is tantamount to arguing that it is likely that the range of values spanned by the ensemble's predictions will either include the true value of that variable or else come within some specified distance of that value. For instance, consider *H*: Under this emission scenario, global mean surface temperature (GMST) for the period 2080–2089 would be between 1.5 °C and 2.0 °C warmer than GMST for the period 1980–1989. Suppose that all of the climate models in an ensemble indicate the truth of this hypothesis and, specifically, that their predicted changes all fall between 1.6 °C and 1.9 °C. Then (2a) will hold if it is likely that the range of predictions delivered by the ensemble will either include the true temperature change or else come within 0.1 °C of doing so.

But there is no good evidence that today's ensembles reliably "capture truth" in this way—or come close enough to capturing it—for predictive variables that interest scientists and decision makers.[10] Today's climate models (and ensembles) have a minimal track record of performance; they are virtually untested when it comes to long-term prediction tasks (we have to wait a long time to see if even a single prediction is borne out), and the relevance of their performance on past data is difficult to determine, for at least three reasons. First, those data are for greenhouse gas levels quite different from the future levels of interest. Second, for some variables, the models have already been tuned to the available observational data, either directly or indirectly.[11] Third, for some variables, the available data come in the form of reanalysis datasets, which are produced by synthesizing traditional observations with results from weather forecasting models; the latter share various simplifying assumptions with climate models.[12]

Thus, unless some other justification for (2a) can be found, the argument presented above from robustness to likely truth remains out of reach for interesting hypotheses about future climate change.[13]

9.4 Robustness and Confidence

Even when an argument from robustness to the likely truth of an agreed-upon predictive hypothesis cannot be given, it might be possible to argue that robustness warrants significantly increased confidence in the hypothesis. Indeed, an analysis by Pirtle et al. (2010) suggests that climate scientists often do assume that agreement warrants this. In this section, three general approaches to arguing from robustness to significantly increased confidence are identified. Each runs into problems in the context of ensemble climate prediction.

A Bayesian Perspective

Within a standard Bayesian framework, one's confidence (or degree of belief) in a hypothesis H is the subjective probability that one assigns to H, and Bayes' Theorem provides a rule for updating that assignment in

light of new evidence *e*. According to the rule, one's new probability assignment, p(*H*|*e*), should be set as follows: p(*H*|*e*) = p(*H*)×p(*e*|*H*)/p(*e*), where p(*H*) is one's probability assignment for *H* prior to obtaining *e*, p(*e*|*H*) is the probability that one assigns to *e* under the assumption that *H* is true, and p(*e*) is the probability that one assigned to *e* before actually encountering *e*. Given this updating rule, confidence in *H* should increase in light of evidence *e* if and only if p(*e*|*H*) > p(*e*|~*H*).[14] That is, *e* will increase confidence in *H* if and only if the occurrence of *e* is more probable if *H* is true than if *H* is false. Similarly, *e* will significantly increase confidence in *H* if and only if the occurrence of *e* is substantially more probable if *H* is true than if *H* is false, i.e. if and only if p(*e*|*H*) >> p(*e*|~*H*), where what counts as "significant" and "substantial" is context-relative.

So a Bayesian argument from robustness to significantly increased confidence in an interesting predictive hypothesis *H* might go as follows:

(3a) *e* warrants significantly increased confidence in *H* if p(*e*|*H*) >> p(*e*|~*H*).
(3b) *e* obtains, where *e* = all of the models in this ensemble indicate *H* to be true.
(3c) p(*e*|*H*) >> p(*e*|~*H*).
∴ Significantly increased confidence in *H* is warranted.

The argument has a valid form. But are its premises true in the case of ensemble climate prediction?

(3a) is part and parcel of the Bayesian framework, as just discussed. (3b) is simply a statement of robustness/agreement. (3c) is where the real action of the argument will be in any particular case and also where the potential weakness of this Bayesian approach becomes clear. For (3c) concerns the subjective probability assignments of a particular epistemic agent, and if those assignments do not reflect substantial evidence, then the move from robustness to increased confidence in *H* could come very cheaply. If the argument above is to have much persuasive force, (3c) should have some substantive justification.

Once again, at least two justificatory approaches are possible, analogous to those discussed for (2a). First, it might be argued that, given the conditions under which the individual models in the ensemble can be

expected to err—inferred from information about how the models are constructed, such as the sorts of idealizations that they include—the models are substantially more likely to agree that *H* is true when it is true than when it is false. A performance-based justification, by contrast, might demonstrate that, in a large set of trials up to now, ensemble members agreed that *H*-type hypotheses were true much more often when those hypotheses were in fact true than when they were in fact false.

Unfortunately, neither sort of justification is readily supplied in ensemble climate prediction today, for reasons already mentioned in Sect. 9.3. Current understanding of the climate system and of the limitations of today's models is not extensive enough to warrant precise-enough conclusions about the conditions under (and the extent to which) the models can be expected to err. And there is no large set of trials of the relevant sort to point to, no significant track record of performance in making long-term predictions. Moreover, there are reasons to worry that simulations from today's state-of-the-art climate models might not so infrequently agree that a predictive hypothesis of interest is true even though it is false.[15] There are climate system features and processes—some recognized and perhaps some not—that are not represented in any of today's models. In addition, when it comes to features and processes that are represented, different models sometimes make use of similar idealizations and simplifications. And errors in simulations of past climate produced by today's models have already been found to display some significant correlation (see e.g. Knutti et al. 2010; Pennell and Reichler 2011). Thus, in general, the possibility should be taken seriously that a given instance of robustness in ensemble climate prediction is, as Nancy Cartwright once put it, "an artifact of the kind of assumptions we are in the habit of employing" (1991, 154).[16]

Perhaps with additional reflection and analysis, persuasive arguments for $p(e|H) >> p(e|{\sim}H)$ can be developed in some cases, but at present such arguments are not readily available.

Condorcet's Jury Theorem

Another possible approach draws on Cordorcet's Jury Theorem. According to the traditional version of this Theorem, if each of $n > 1$ voters has the

same probability $p > 0.5$ of voting correctly regarding which of two options is "better" (on some criterion) and if the votes are statistically independent, then the probability that at least a majority of voters will choose the "better" option exceeds p and, moreover, exceeds p to a greater extent with increasing n (see e.g. Ladha 1995). Treating the indications of individual simulations regarding the truth of a predictive hypothesis as votes, an argument from robustness to increased confidence might be made as follows[17]:

(4a) The indications from these simulations are statistically independent, and each simulation has the same probability $p > 0.5$ of giving the correct indication regarding H.[18]
(4b) All of the simulations in this collection indicate that H is true.
(4c) If (4a) and (4b), then increased confidence in H (beyond the confidence had in light of just one simulation indicating H) is warranted.
∴ Increased confidence in H is warranted.

When it comes to ensemble climate prediction, the most obvious difficulties with this argument arise in connection with (4a). First, while including a model in an ensemble study aimed to discern the truth/falsity of a particular predictive hypothesis would presumably imply a belief that $p > 0.5$ for that model, in many cases (i.e. for many predictive hypotheses of interest) the basis for such a belief may not be very strong, for reasons already discussed. In addition, the assumption of independence is clearly questionable. In traditional applications of Cordorcet's Jury Theorem, independence is assumed to require that voters do not confer with one another, do not base their votes on shared information, do not have similar training and experience, and are not influenced by opinion leaders (see Ladha 1995, 354). How independence should be evaluated in the context of climate modeling is still a matter of some discussion (see e.g. Abramowitz 2010; Pirtle et al. 2010). But many modeling groups do have similar training and experience, and predictions from today's climate models clearly are based on substantial shared information, including but not limited to previously published predictions, which may influence modeling groups as they develop and fine-tune their models

(see also Tebaldi and Knutti 2007, 2067–8). Moreover, as noted above, recent investigations have found that errors in simulations of past and present climate produced by today's state-of-the-art climate models show significant correlation (see Knutti et al. 2010; Pennell and Reichler 2011).

There are generalizations of Condorcet's Jury Theorem that have more relaxed assumptions about the competence of voters (e.g. Owen et al. 1989) or that allow certain kinds of dependence among votes (e.g. Ladha 1992, 1995). For instance, while still assuming that voters have the same probability $p > 0.5$ of voting correctly, Ladha (1992) argues that the probability that the majority vote is correct exceeds p if the average correlation among the voters' choices remains small enough. Perhaps these generalized versions of the Theorem hold some promise when it comes to developing a sound argument from robust climate modeling results to significantly increased confidence in agreed-upon predictive hypotheses.[19] But once again, such arguments will require information that is not easy to come by, such as information about how reliably today's models indicate correctly the truth/falsity of hypotheses of a relevant class.

A Sampling-Based Perspective

Although it is commonly assumed that ensemble studies somehow involve sampling, it is not obvious how a sampling-based argument from robust model predictions to significantly increased confidence might best be constructed. What follows is one good-faith attempt.

Let q be a set of criteria that can be used to rate any given model's perceived quality as a tool for correctly indicating the truth/falsity of some particular predictive hypothesis H. Assume that today's scientists construct this quality metric q in light of current scientific understanding and computing power—it might take into account whether a model includes particular physical assumptions, how it performs in simulating the behavior of the target system up to now, its spatiotemporal resolution, and so on. Let M_B be the collection of all models, whether already constructed by scientists or not, whose score on q would exceed some chosen threshold; the models in M_B have features such that they are considered to be, at present, the best models for the predictive purposes at

hand. Then the following argument from robustness to increased confidence might be given[20]:

(5a) In the absence of other overriding evidence, confidence in predictive hypothesis H should

equal f, the fraction of models in M_B whose simulations indicate that H is true.

(5b) If all of the simulations produced by models in a random sample from M_B are found to agree in

indicating that H is true, then an increase in the current estimate of f—and correspondingly an increase in confidence in H—is warranted.

(5c) This collection of models is a random sample from M_B.
(5d) The simulations produced by models in this collection all indicate that H is true.
∴ Increased confidence in H is warranted.

Compared to previous arguments, the logic of this one is less tight. While a number of concerns about the argument might be raised, in the context of ensemble climate prediction the most obvious problem is (5c), which asserts that some particular ensemble of today's models is a random sample from M_B. This suggests that the scope of some M_B has been identified—that scientists have some sense of the space of models that it encompasses—and that a randomizing procedure was employed when selecting today's models from M_B. But this is not so.

As noted in Sect. 9.3, today's multi-model ensembles are widely acknowledged to be ensembles of opportunity; any "sampling" by which they are assembled "is neither systematic nor random" (Tebaldi and Knutti 2007, 2068). In fact, according to some climate scientists, "it is not clear how to define a space of possible model configurations of which [today's multi-model ensemble] members are a sample" (Murphy et al. 2007, 1995; see also Parker 2010). Given present uncertainty about how

to represent the climate system, any reasonable quality metric that today's climate scientists might specify would allow that many climate models that differ significantly (in their construction) from today's models would qualify for inclusion in M_B. Given the shared history of today's models, it may well be that they differ from one another much less than random samples from M_B typically would, which in turn might make them biased estimators of f.[21]

To sum up, various arguments from robustness to significantly increased confidence in an agreed-upon predictive hypothesis of interest are possible, but none of the arguments considered above is readily applicable in the context of ensemble climate prediction today. Arguments invoking a Bayesian perspective or a generalized version of the Cordorcet Jury Theorem show some promise, but further information is needed before these arguments can be advanced in a way that is persuasive.

9.5 Robustness and Security

A third view regarding the significance of robustness can be found in work by Kent Staley (2004). He sets aside the question of whether robustness can increase the strength of evidence for a hypothesis and instead focuses on the *security* of evidence claims—the degree to which an evidence claim is immune to defeat when there is a failure of one or more auxiliary assumptions relied upon in reaching it (ibid., 468). Staley argues that robust test results can increase the security of evidence claims in several ways, one of which will be developed in greater detail here.[22]

Suppose that in light of the results of some test procedure, such as a laboratory experiment or a computer simulation, scientists arrive at an evidence claim, *E:* "We have evidence of at least strength *S* for hypothesis *H*." The strength *S* might be expressed qualitatively (e.g. weak, strong, conclusive) or perhaps quantitatively.[23] In order to arrive at *E*, the scientists rely on a set of auxiliary assumptions, *A,* which includes assumptions about the test procedure (e.g. that the apparatus involved did not malfunction, that the test procedure is of a moderately reliable kind, etc.). These auxiliary assumptions are ones that the scientists believe to

be true.[24] If any one of the assumptions turns out to be mistaken, the inference from the results of the test procedure to *E* will need to be reconsidered. Now suppose the scientists conduct a second test of *H*, and the results of the second test, in conjunction with a set of auxiliary assumptions, *A',* lead the scientists to the same evidence claim *E*. That is, as with the first test results, the scientists consider the second test results to provide evidence of at least strength *S* for hypothesis *H*. Then as long as *A'* is at least *partially logically independent* of *A*—that is, as long as there is at least one assumption in *A* such that, even if that assumption is false, all assumptions in *A'* could still be true—then the security of the scientists' evidence claim *E* will be enhanced, since in effect they will have discovered that there is a "back-up route" to *E* that *might* remain intact even if their original inference to *E* turns out to involve a mistaken assumption (see also Staley 2004, 474–475).[25]

A version of this argument that might be applied to members of a set of modeling results (e.g. to each ensemble member in turn) is as follows:

(6a) A modeling result r_n enhances the security of an evidence claim *E* if:

(i). *E* is derivable from each of modeling results $r_1 \ldots r_{n-1}$, respectively, in conjunction with sets of auxiliary assumptions $A_1 \ldots A_{n-1}$, respectively;
(ii). *E* is derivable from r_n in conjunction with some set of auxiliary assumptions, A_n; and
(iii). A_n is partially logically independent of each of $A_1 \ldots A_{n-1}$.

(6b) (i)–(iii) are met in this case.
∴The security of *E* is enhanced by r_n.

If (6a) is accepted as an analysis of the minimal conditions for increasing security, then the question is whether (i)–(iii) are met in ensemble climate prediction today.[26]

Working backwards, it seems that (iii) often is met. In reaching an evidence claim *E* from any given simulation result, climate scientists will make use of a number of auxiliary assumptions. Assuming that these

concern the appropriateness of the model's physical assumptions and numerical solution techniques, the absence of significant programming errors, the reliability of the computing platform on which the model is run, etc., then the sets of auxiliary assumptions used in conjunction with different simulation results can be expected to differ from one another in various ways, since the models producing the simulations will not all reflect the same assumptions about the climate system, will not all be run on the same computing platform, and so on. It seems clear that each set of auxiliary assumptions will be at least partially logically independent of each of the other sets.

For (i) and (ii), the situation is less clear. In practice, it is often assumed that results from different state-of-the-art climate models each constitute weak (positive) evidence regarding the truth/falsity of interesting predictive hypotheses. (Only together might they provide strong evidence.) This suggests that, when results from these climate models agree that predictive hypothesis *H* is true, climate scientists might conclude on the basis of each result, in conjunction with various auxiliary assumptions, that *E*: There is weak evidence for *H*.

Unfortunately, it is not clear that the key underlying assumption—that each simulation result has positive evidential relevance—can be given solid justification.[27] The reasons are by now familiar: uncertainty about the importance of various climate system processes, constraints on model construction due to limited computing power, relatively few opportunities to test climate model performance, and difficulty in interpreting the significance of model-data fit in cases where comparisons can be made. While it is true that today's state-of-the-art climate models are constructed using an extensive body of knowledge about the climate system and that they generally deliver projections of future climate that are (from a subjective point of view) quite plausible in light of current scientific understanding, their individual reliability in indicating the truth/falsity of quantitative predictive hypotheses of the sort that interest today's scientists and decision makers remains significantly uncertain; indeed, it is in part because of this uncertainty that the move to ensembles is made in the first place (see Sect. 9.2).[28] Thus, it appears that even claims of enhanced security do not come easily in the context of ensemble climate prediction today.

9.6 Concluding Remarks

The foregoing analysis revealed that, while there are conditions under which robust predictive modeling results have special epistemic significance, scientists generally are not in a position to argue that those conditions are met in the context of present-day climate modeling. Typically, when today's climate models are in agreement that an interesting hypothesis about future climate is true, it cannot be inferred—via the arguments considered here anyway—that the hypothesis is likely to be true, nor that confidence in the hypothesis should be significantly increased, nor that a claim to have evidence for the hypothesis is now more secure. This is disappointing.

Nevertheless, the analysis did reveal goals for the construction and evaluation of ensembles—whether in the study of climate change or in any other context—such that robust results will have desired epistemic significance. One goal, for instance, is the identification of a collection or space of models that can be expected to include at least one model that is adequate for indicating the truth/falsity of the hypothesis of interest; sampling from this collection (in order to construct the ensemble) should then be exhaustive, if possible, or else aimed to produce maximally different results. In other cases, when ensembles are not carefully constructed in this way, the goal might be to obtain extensive error statistics regarding the past performance of the ensemble in indicating the truth/falsity of hypotheses of the relevant sort; this in turn will require careful consideration of which hypotheses are relevant.

When it comes to ensemble climate prediction, the prospects for reaching these goals in the near future seem dim. Certainly, the design of multimodel ensemble studies could be improved, aiming to better sample recognized uncertainty about how to adequately represent the climate system for a given predictive task, but the specification and deployment of ensembles that can (with justification) be expected to include adequate models—while still giving robust results—seems likely to remain beyond scientific understanding for some time. Likewise, in the near term it will be difficult to obtain desired error statistics for climate ensembles, given the long-term nature of the predictions of interest, the limited time span for

which reliable observational data are available, the lack of comprehensiveness of these data (leading to reanalysis), and the practice of tuning.[29]

That said, prospects seem substantially brighter in some other predictive modeling contexts. For instance, when it comes to hypotheses about the next opportunities to see solar eclipses from various locations on Earth, today's physicists might well have sufficient background knowledge to design ensembles that can be expected to include at least one model that is adequate for predicting the eclipse occurrence with desired accuracy. Likewise, today's weather forecasters might collect extensive error statistics on the performance of ensemble weather forecasting systems, providing good evidence that $p(e|H) >> p(e|\sim H)$ for quantitative hypotheses about next-day high temperatures in a given locale. In cases like these, robust model predictions may well have special epistemic significance.

Acknowledgments This is a revised version of Parker, W.S. 2011. "When climate models agree: The significance of robust model predictions," *Philosophy of Science* 78(4): 579–600. Thanks to University of Chicago Press for permission to republish substantial portions of that paper. I have benefitted from the suggestions and criticisms of Dan Steel, Reto Knutti, Kent Staley, Phil Ehrlich, Leonard Smith, Joel Katzav, Charlotte Werndl, and two anonymous referees for *Philosophy of Science*.

Notes

1. By an interesting predictive hypothesis, I mean a hypothesis about the future that scientists (i) do not already consider very likely to be true or very likely to be false and (ii) consider a priority for further investigation. In climate science today, these are typically, but not always, quantitative hypotheses about changes in global or regional climate on the timescale of several decades to centuries.
2. When does an ensemble agree that a hypothesis is true? Assume that the values of model variables can be translated into statements regarding target system properties. Then a simulation indicates the truth (falsity) of some hypothesis H about a target system if its statements about the target system entail that H is true (false). For example, if H says that

temperature will increase by between 1 °C and 1.5 °C, and each of the simulations in an ensemble indicates an increase between 1.2 °C and 1.4 °C, then each of those simulations indicates the truth of *H* and the ensemble is in agreement that *H* is true.

3. I take *agreement* among modeling results to be synonymous with *robustness*, as is common in the climate modeling literature. Some authors define robustness differently (see e.g. Pirtle et al. 2010).
4. Projections are predictions of what would happen under specified scenarios in which greenhouse gases and other climate forcing factors evolve in particular ways.
5. In part, this is because it is difficult to determine what a model's performance in simulating past and present climate indicates about its accuracy in predicting various quantities of interest (see Randall et al. 2007; Gleckler et al. 2008; Parker 2009).
6. More precisely, average results for individual models were in agreement regarding the hypothesis; some models were run more than once with different initial conditions, and only average results for each model were shown in the main body of the report.
7. Woodward (2006) notes the limited applicability of a related analysis.
8. A simulation indicates correctly regarding a hypothesis *H* if it indicates the correct truth value for *H*.
9. So, while we might not know *which member(s)* of the ensemble will indicate correctly regarding a given *H*, we have evidence that there is usually *at least one such member* in the ensemble.
10. The "capturing truth" terminology is taken from Judd et al. (2007), which includes a related technical definition of the "bounding box" of an ensemble.
11. Tuning a climate model involves making ad hoc changes to its parameter values or to the form of its equations in order to improve the fit between the model's output and observational data.
12. See Edwards (1999, 2010) and Parker (2016), for non-technical discussions of reanalysis datasets.
13. It is important to recall the definition of "interesting hypotheses" given in Fn. 1. The conclusion here is fully compatible with there being some hypotheses about future climate that scientists can, with justification, consider likely to be true. The expectation that global climate will continue to warm, for instance, is grounded not just in agreement among predictions from complex climate models, but also in basic understanding

of physical processes, theoretical analysis, observational data, and results from simpler models.

14. Here I assume that p(H) takes a value between zero and one, i.e. it is not known to be certainly true or certainly false. From the updating rule, we see that $p(H|e) > p(H)$ iff $p(e|H)/p(e) > 1$. But $p(e|H)/p(e) > 1$ iff $p(e) < p(e|H)$. When is $p(e) < p(e|H)$? By the law of total probability, $p(e) = p(e|H) \times p(H) + p(e|\sim H) \times p(\sim H)$. Since $p(H) + p(\sim H) = 1$, $p(e)$ is in effect a weighted average of $p(e|H)$ and $p(e|\sim H)$; it takes a value between $p(e|H)$ and $p(e|\sim H)$. So $p(e)$ will be smaller than $p(e|H)$ iff $p(e|H) > p(e|\sim H)$. So $p(H|e) > p(H)$ iff $p(e|H) > p(e|\sim H)$.
15. The reasons given here are also discussed by Tebaldi and Knutti (2007).
16. Wimsatt (2007) discusses a case in biology in which an apparently robust modeling result turned out to be grounded in erroneous assumptions shared by the models. See also Orzack and Sober (1993, 539).
17. For the sake of simplicity, the argument given here targets increased confidence, rather than significantly increased confidence. It is relatively easy to imagine how an analogous argument for significantly increased confidence might be given, once what counts as "significant" is defined in the case of interest.
18. Note that from this it follows that $p(e|H) > p(e|\sim H)$, so a Bayesian argument from robustness to increased confidence (similar to that of Sect. 9.4.1) could also be made.
19. Odenbaugh (2012) considers how a relaxed version of the Condorcet Jury Theorem might be used to analyze the significance of scientific consensus (among experts, rather than models) regarding the existence and causes of global climate change.
20. See Footnote 17.
21. This is assuming that *f* can be defined for M_B; this issue is not addressed here. If *f* cannot be defined, then (5a) is also problematic.
22. The present analysis expands upon the insightful but brief discussion given by Staley (2004, 474–475).
23. Important questions about how the strength of evidence is defined and determined remain to be addressed; for the sake of discussion, it is assumed here that some reasonable and coherent analysis can be given.
24. In fact, scientists may only believe that these assumptions are close enough to being true. For the sake of simplicity, this is ignored in the discussion above; including it would complicate but not undermine the argument.
25. The mathematical logician typically uses a somewhat different notion of logical independence.

26. Security can be enhanced more or less. Ceteris paribus, the closer the sets of auxiliary assumptions come to being fully logically independent of one another, the more security is enhanced. A set of assumptions *A'* is *fully logically independent* of another set *A* if every assumption in *A* is such that, if that assumption is false, all of the assumptions in *A'* could still be true. Security is also enhanced more, ceteris paribus, to the extent that it is not only possible that all of the assumptions in *A'* could be true even while some assumption in *A* is false, but likely that all of the assumptions in *A'* will be true if some assumption in *A* is false. For simplicity, the discussion above does not consider this quantitative aspect of enhanced security.
27. Note that even if results from each climate model in an ensemble do have positive evidential relevance, this is not necessarily enough for the argument of Sect. 4.1 to work. That argument also depends upon the correlations among erroneous indications from the models, and even models that individually are more reliable than chance may nevertheless be more likely to agree in indicating that *H* is true when in fact it is false than when in fact it is true. Thanks to Dan Steel for reminding me to attend to connections between the discussion here and in Sect. 4.1.
28. The claim here is not that individual modeling results have negative evidential relevance, but that their evidential status (with regard to interesting hypotheses about long-term climate change) is largely unknown.
29. Of course, it does not follow that climate policy decisions should be put on hold. Expectations of a warmer world are well founded; the challenge is rather to make sensible decisions despite remaining uncertainties about the details of future climate change.

References

Abramowitz, Gabriel. 2010. Model Independence in Multi-Model Ensemble Prediction. *Australian Meteorological and Oceanographic Journal* 59: 3–6.

British Broadcasting Corporation (BBC). 2010. *Climate Change Experiment Results*. http://www.bbc.co.uk/sn/climateexperiment

Cartwright, Nancy, et al. 1991. Replicability, Reproducibility, and Robustness: Comments on Harry Collins. *History of Political Economy* 23 (1): 143–155.

Christensen, Jens Hesselbjerg, Bruce Hewitson, Aristita Busuioc, Anthony Chen, Xuejie Gao, R. Held, Richard Jones, et al. 2007. Chapter 11: Regional

Climate Projections. In *Climate Change, 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, 847–940. Cambridge: University Press.

Collins, M., R. Knutti, J. Arblaster, J.L. Dufresne, T. Fichefet, P. Friedlingstein, X. Gao, W.J. Gutowski, T. Johns, G. Krinner, and M. Shongwe. 2013. Long-Term Climate Change: Projections, Commitments and Irreversibility. In *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, 1029–1136. Cambridge: Cambridge University Press.

Edwards, Paul N. 1999. Global Climate Science, Uncertainty and Politics: Data-Laden Models, Model-Filtered Data. *Science as Culture* 8 (4): 437–472.

———. 2010. *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. Cambridge, MA: MIT Press.

Gleckler, P.J., K.E. Taylor, and C. Doutriaux. 2008. Performance Metrics for Climate Models. *Journal of Geophysical Research: Atmospheres* 113 (D6): D06104. https://doi.org/10.1029/2007JD008972.

IPCC. 2013. In *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, ed. Thomas F. Stocker, Dahe Qin, Gian-Kasper Plattner, Melinda M.B. Tignor, Simon K. Allen, Judith Boschung, Alexander Nauels, Yu Xia, Vincent Bex, and Pauline M. Midgley. Cambridge: Cambridge University Press.

Judd, Kevin, Leonard A. Smith, and Antje Weisheimer. 2007. How Good Is an Ensemble at Capturing Truth? Using Bounding Boxes for Forecast Evaluation. *Quarterly Journal of the Royal Meteorological Society* 133 (626): 1309–1325.

Knutti, Reto, Myles R. Allen, Pierre Friedlingstein, Jonathan M. Gregory, Gabriele C. Hegerl, Gerald A. Meehl, Malte Meinshausen, et al. 2008. A Review of Uncertainties in Global Temperature Projections over the Twenty-First Century. *Journal of Climate* 21 (11): 2651–2663.

Knutti, Reto, et al. 2010. Challenges in Combining Projections from Multiple Climate Models. *Journal of Climate* 23: 2739–2758.

Ladha, Krishna K. 1992. The Condorcet Jury Theorem, Free Speech, and Correlated Votes. *American Journal of Political Science* 36 (3): 617–634.

———. 1995. Information Pooling through Majority-Rule Voting: Condorcet's Jury Theorem with Correlated Votes. *Journal of Economic Behavior & Organization* 26 (3): 353–372.

Levins, Richard. 1966. The Strategy of Model Building in Population Biology. *American Scientist* 54 (4): 421–431.

———. 1993. A Response to Orzack and Sober: Formal Analysis and the Fluidity of Science. *The Quarterly Review of Biology* 68 (4): 547–555.

Lloyd, Elisabeth A. 2010. *Confirmation and Robustness of Climate Models*. *Philosophy of Science*. 77 (5): 971–984.

———. 2015. Model Robustness as a Confirmatory Virtue: The Case of Climate Science. *Studies in History and Philosophy of Science Part A* 49: 58–68.

Meehl, Gerard A., Thomas F. Stocker, William D. Collins, A.T. Friedlingstein, Amadou T. Gaye, Jonathan M. Gregory, Akio Kitoh, et al. 2007. Global Climate Projections. In *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, 747–845. New York: Cambridge University Press

Murphy, James M., Ben B.B. Booth, Mat Collins, Glen R. Harris, David M.H. Sexton, and Mark J. Webb. 2007. A Methodology for Probabilistic Predictions of Regional Climate Change from Perturbed Physics Ensembles. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 365 (1857): 1993–2028.

Odenbaugh, Jay. 2012. *Climate, Consensus, and Contrarians. The Environment: Philosophy, Science, and Ethics*, 137–150. Cambridge: MIT Press.

Orzack, Steven Hecht, and Elliott Sober. 1993. A Critical Assessment of Levins's the Strategy of Model Building in Population Biology (1966). *The Quarterly Review of Biology* 68 (4): 533–546.

Owen, Guillermo, Bernard Grofman, and Scott L. Feld. 1989. Proving a Distribution-Free Generalization of the Condorcet Jury Theorem. *Mathematical Social Sciences* 17 (1): 1–16.

Parker, Wendy S. 2006. Understanding Pluralism in Climate Modeling. *Foundations of Science* 11 (4): 349–368.

———. 2009. II—Wendy S. Parker: Confirmation and Adequacy-for-Purpose in Climate Modelling. *Aristotelian Society Supplementary Volume*. 83: 233–249.

———. 2010. Whose Probabilities? Predicting Climate Change with Ensembles of Models. *Philosophy of Science* 77 (5): 985–997.

———. 2016. Reanalyses and Observations: What's the Difference? *Bulletin of the American Meteorological Society* 97 (9): 1565–1572.

Pennell, Christopher, and Thomas Reichler. 2011. On the Effective Number of Climate Models. *Journal of Climate* 24 (9): 2358–2367.

Pirtle, Zachary, Ryan Meyer, and Andrew Hamilton. 2010. What Does It Mean When Climate Models Agree? A Case for Assessing Independence Among General Circulation Models. *Environmental Science & Policy* 13 (5): 351–361.

Randall, David A., Richard A. Wood, Sandrine Bony, Robert Colman, Thierry Fichefet, John Fyfe, Vladimir Kattsov, et al. 2007. Climate Models and Their Evaluation. In *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the IPCC*, 589–662. Cambridge: Cambridge University Press.

Solomon, Susan, et al. 2007. *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change.* New York: Cambridge University Press.

Stainforth, David A., Tolu Aina, Carl Christensen, Mat Collins, Nick Faull, Dave J. Frame, Jamie A. Kettleborough, et al. 2005. Uncertainty in Predictions of the Climate Response to Rising Levels of Greenhouse Gases. *Nature* 433 (7024): 403–406.

Staley, Kent W. 2004. Robust Evidence and Secure Evidence Claims. *Philosophy of Science* 71 (4): 467–488.

Tebaldi, Claudia, and Reto Knutti. 2007. The Use of the Multi-Model Ensemble in Probabilistic Climate Projections. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 365 (1857): 2053–2075.

Tebaldi, Claudia, Richard L. Smith, Doug Nychka, and Linda O. Mearns. 2005. Quantifying Uncertainty in Projections of Regional Climate Change: A Bayesian Approach to the Analysis of Multimodel Ensembles. *Journal of Climate* 18 (10): 1524–1540.

Weisberg, Michael. 2006. Robustness Analysis. *Philosophy of Science* 73 (5): 730–742.

Wimsatt, William C. 2007. Robustness, Reliability, and Overdetermination. In *Re-engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*. Cambridge: Harvard University Press.

Woodward, Jim. 2006. Some Varieties of Robustness. *Journal of Economic Methodology* 13 (2): 219–240.

10

Building Trust, Removing Doubt? Robustness Analysis and Climate Modeling

Jay Odenbaugh

10.1 Introduction

In this chapter, I first provide a conceptual framework for thinking about model building and evaluation and apply this framework to climate modeling and in particular to global circulation models (GCM).[1] After considering in detail the question of what makes models independent, I turn to model robustness. Insofar as we are in doubt regarding our model's idealizations and thus in doubt regarding their predictions, robustness analysis can remove those doubts by showing when idealizations are irrelevant. Thirdly, I consider a dilemma for robustness analysis; namely, it leads to either an infinite regress of idealizations or a complete removal of idealizations. A response to the dilemma is given defending a form of epistemic contextualism and by drawing a distinction between relative and absolute robustness.

J. Odenbaugh (✉)
Department of Philosophy, Lewis and Clark College, Portland, OR, USA

E.A. Lloyd, E. Winsberg (eds.), *Climate Modelling*,
https://doi.org/10.1007/978-3-319-65058-6_10

10.2 A Philosophical Sketch of Climate Model Building and Evaluation

As we are often told, weather and climate are distinct. Weather, on the one hand, concerns the state of the atmosphere and ocean at a given moment in time. On the other, climate concerns statistical properties regarding these states such as average temperature, average precipitation, and average humidity along with other properties of weather variability. Climate is the causal product of several interacting systems, including the atmosphere, ocean, land surfaces, sea and land ice, and the biosphere.[2] In a Laplacean universe, climate scientists might hope for a climate theory in which all of the causal processes are truly represented and for any time *t*, an exactly correct prediction of the climate at *t* could be given. That is, we might strive for a true exact representation of atmospheric circulation, ocean circulation, heat balances, cloud cover, the uptake and release of CO_2 by biological systems, and so on. However, no such theory is forthcoming; we uncertainly and imprecisely study aspects of our climate. For example, we separately study the *physical climate system*, which includes weather, El Niño, North Atlantic Oscillation, monsoon variations, droughts, floods, ice ages, and so forth; *environmental chemistry*, including the ozone hole, air pollution, aerosol formation, and so on; *biosphere*, which includes the atmosphere's evolution, the production of oxygen, the carbon cycle, and so on. In addition, we do study some of the linkages between these different systems.

In sum, our climate theories are *models* (Odenbaugh 2005). The term "model" is typically used to denote abstract and idealized representations. Abstractions are representations in which only some of the properties of the phenomena are represented. Idealizations are representations that falsely or inaccurately represent the properties they include.[3] The representational vehicles may be mathematical, graphical, or even "physical." In this chapter, I suppose that a model consists of a set of propositions some of which are idealized. Moreover, I will talk of a model's assumptions consisting in scientists taking a certain attitude towards them; namely, they *assume* or *suppose* them for the purposes of explanation, prediction, and intervention (Callendar and Cohen 2006; Sorensen 2012).[4]

So, some climate scientists build models. What do these models consist in? To begin, climate modelers suppose we have a grid of points over the Earth and which are spaced at say 100 km in the horizontal and 1 km in the vertical (with 20 such vertical levels). At each point, a GCM specifies the values of a variety of variables, including pressure, temperature, humidity, wind velocity, and so on. The "fineness" of the spacing of the grid points is largely determined by computational power and available data. However, given some initial state of the climate, how do you forecast or project what the climate will be?

Numerical climate models include several basic equations.[5] First, we have the horizontal momentum equations, known as Newton's second law of motion or the Navier-Stokes equations of motion, representing the horizontal acceleration of a volume of air resulting from the Coriolis force balanced by the horizontal pressure gradient and friction. In addition, we have the vertical velocity equation that represents the balance between vertical pressure gradient and gravity. Second, we have an equation of state that for the atmosphere (there is a distinct equation for the ocean that depends on temperature, salinity, and pressure). This is known as the ideal gas law. Third, we have the thermodynamic energy equations that are for the ocean and air. Finally, we have the continuity equation (conservation of mass). Of course, in order to initialize a GCM we must input data that are collected from surface observations, from ships and buoys, from radiosonde balloons and satellites.

Lastly, there are assumptions that vary across models, including what forcing agents are present (e.g. GHG, aerosols), and "parameterizations." Parameterizations involve providing representations of sub-grid processes such as moist processes (e.g. evaporation, condensation, formation and dispersal of clouds), absorption, emission, reflection of solar and thermal radiation, convective processes, and friction, heat and water vapor at the surface. Much of climate uncertainty concerns these sub-grid processes. GCM are very complex since we must solve these equations at every grid point for each time step.

So far, we have considered model building but now we consider model evaluation. Following the work of philosopher Elisabeth Lloyd, model evaluation involves several different components: (a) there is

goodness-of-fit, (b) testing independent assumptions of models, (c) variety of evidence, and (d) robustness analysis (Lloyd 1994, 2010, 2015).[6]

Considering (a), Lloyd writes,

> The most obvious way to support a claim of the form 'this natural system is described by the model' is to demonstrate the simple matching of some part of the model with some part of the natural system being described (1994, 146)

As one example of goodness-of-fit, climate scientists have evaluated their GCM by examining predictions and data including the volcanic activity of Mt. Pinatubo (see IPCC AR4 2007 for evaluation of many other GCM predictions). Here we have the predicted and observed changes in global land and ocean sea surface air temperature after Mt. Pinatubo erupted (Fig. 10.1).

As we can see, the GCM reproduces the cooling effect of the volcanic eruption. For GCMs, goodness-of-fit is complicated for at least three reasons. First, not every fit between data and model is confirmatory. For example, if we test a model against data used to initialize the model—to "tune" it—then it is not surprising that the model fits the data. Hence,

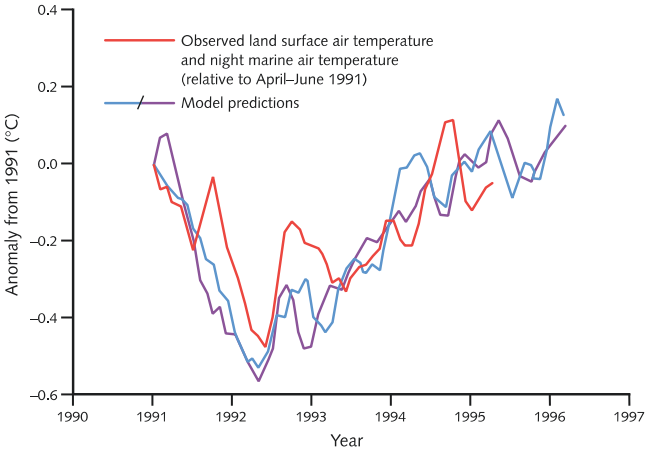

Fig. 10.1 A comparison of GCM and Mt. Pinatubo (From Houghton 2009, 123)

we must test our models against distinct data sets, which raises important questions about when data sets are distinct.[7] Second, it can be very difficult to determine what statistical measure of fit to use since there are many and they often involve substantive empirical and philosophical assumptions. Third, GCM are typically evaluated as ensembles; a group of models are simulated with different scenarios and they are evaluated as a group against data (Parker 2006, 2010; Lenhard and Winsberg 2010).

Lloyd argues that goodness-of-fit is not enough in evaluating our models. In addition to considering its "input-output profile," we must also "look under the hood" (Hausman 2008). Indeed, we must consider the different assumptions our models make. She writes,

> Numerous assumptions are made in the construction of any model. These include assumptions about which factors influence the changes in the system, what the ranges for the parameters are, and what the mathematical form of the laws is. Many of these assumptions have potential empirical content.... [W]hen empirical claims are then made about this model, the assumptions may have empirical significance. (1994, 147)

Model assumptions include the choice of state variables, choice of parameters and whether they take constant values or are random variables, and the laws of succession or coexistence and whether they are continuous or discrete, deterministic, or stochastic.[8] Clearly one can determine the values of the state variables and not test the assumptions of the model. Imagine we have two models that correctly predict that a state variable takes the same determinate value at some time t. Absent any other evidence, we cannot justifiably choose between the models. The evidence we have is consistent with the first's assumptions being true and the second's false, vice versa, and both being false. In some cases, however, we can test the assumptions of a model. Consider a simple example, the exponential growth model $dN/dt = rN$. One assumption of the model is that there is no intraspecific competition occurring in a population of a species. Biologists can sometimes determine that individuals of the same species compete. With regard to GCM, we must evaluate the laws postulated regarding the relations between variables and parameters. Climate scientists have done this insofar as they have tested assumptions of GCM such

as the Navier-Stokes equations, ideal gas law, and various parameterizations. However, we still need an explication of what testing *independent* assumptions means. This is a difficult topic and is crucial for understanding model robustness.

One might construe independent assumptions as assumptions that are independent in the traditional statistical sense (Lloyd 1994, 149). On this interpretation, events E_1 and E_2 are independent if and only if $P(E_2|E_1) = P(E_2) \times P(E_1)$. Or, if we are talking about random variables X and Y, with cumulative distribution functions $F_X(x)$ and $F_Y(Y)$ and probability densities $f_X(x)$ and $f_Y(y)$, then X and Y are independent if and only if the combined random variable (X, Y) has a joint cumulative distribution function $F_{X,Y}(x, y) = F_X(x)F_Y(y)$ (or a joint density $f_{X,Y}(x, y) = f_X(x)f_Y(y)$). Assumptions, or specifically their constituent propositions, are statistically independent insofar as the events or random variables they describe are independent. Insofar as our model assumptions do not concern experimental setups for events or random variables, I will construe models with independent assumptions as ones with *logically independent* assumptions.

Suppose we have a model with an assumption A_1 and another with assumption A_2 where A_1 does not entail A_2 or vice versa. For example, insofar as the Navier-Stokes equations are logically independent of the ideal gas law, then evidence for the one is independent of the other. *S*uppose we have a model with assumption A_1 and another with assumption A_2 and further suppose that A_1 entails A_2 but not conversely. If we test A_2 and determine that it is false, then we have tested A_1 and have determined it to be false. However, the converse need not follow. We could determine that A_1 is false and yet A_2 be true. For example, our equation of state for the ocean is such that density is a function of temperature, salinity, and pressure, or $\rho = \wp(T, S, P)$. We can characterize a coefficient of thermal expansion e_T which is the percent decrease in density per degree of temperature increase. If we have a small change in temperature and density relative to our reference values T_0 and ρ_0 and where salinity is not changing, we can linearly approximate the above function $\rho = \rho_0[1 - e_T(T - T_0)]$ (Neelin 2010, 77–78). If it turns out our assumption that ρ is a function of temperature, salinity, and pressure is incorrect, then so must be the linear approximation for ρ. However, the linear

approximation can be false when the assumption that ρ is a function of temperature, salinity, and pressure is true. So, even when we do not have well-defined experimental setups, it is reasonable to construe the notion of independently testing the assumptions of a model as shorthand for the testing of logically independent assumptions. If two or more assumptions are not logically independent of one another and we cannot properly interpret them as statistically independent, then we cannot test them independently of one another.

Incidentally, the testing of independent assumptions is related to but distinct from the notion of "ad hocness." Theories or models generally only make predictions when coupled with auxiliary hypotheses. An auxiliary hypothesis is considered ad hoc if it cannot be tested independently of the theory under consideration. For example, in the nineteenth century, Newtonian mechanics had successfully accounted for the orbits of most of the known planets in our solar system. However, there was one particularly difficult case, Uranus. Uranus' orbit was not as predicted by Newtonian mechanics. Astronomers concluded that either Newtonian mechanics was incorrect or they had made some mistake in the application of these laws. John Adams and Urbain Leverrier proposed that there was an unobserved planet of a certain size and distance beyond Uranus and they subsequently computed its expected orbit. They found that the orbit of Uranus was as Newtonian mechanics predicts when coupled with the additional auxiliary hypothesis. Eventually the unobserved planet, Neptune, was observed and Newtonian mechanics was credited with the success. The auxiliary hypothesis, "There is a planet of a certain size and at a certain location which gives rise to certain perturbations in Uranus' orbit," was not ad hoc since one could test it independently of Newton's law of gravitation and laws of motion. Eventually, one could use a telescope to detect Neptune. Ad hocness concerns dependent auxiliary assumptions *outside* and not *inside* our model.

Lloyd has argued that (c)—variety of evidence—is critical for model evaluation. Variety comes in, well, varieties and so we will consider two types. First, ideally, we would like our model's goodness-of-fit profile to be determined through a variety of data sets. For example, if a GCM fits Mt. Pinatubo data, tide gauge data, and average surface temperatures, then it is better confirmed than if it fit any one or two of these

data sets. Ideally, we would like not just different tokens of the same data type (e. g. different fits to volcano data) but tokens of different data types (e. g. different fits to volcano data, tide gauge data, etc.). Prima facie then, a model that fits a greater number of data types is more confirmed than one that fits fewer. Second, a model that has evidence for more independent assumptions is more confirmed than one that has evidence for fewer independent assumptions. I now turn to (d) model robustness.

10.3 Model Robustness

One way in which we can understand the epistemic significance of robustness is to consider how it can be used to answer a very simple question.[9] Suppose one says of a model, "Why accept this model even if its predictions are confirmed; we already know that it is false?" Robustness analysis equips us to reply, "Those idealizations don't matter; they are harmless since our model would make the same prediction without them." In what follows, I will articulate an account of model robustness due to the work of Richard Levins (1966), William Wimsatt (2007), Michael Weisberg (2006), mine (2011), and Elisabeth Lloyd (2010)—what we might term the "LWWOL" approach (Lloyd 2015).

Consider a family **M** of models; **M** is a model type. Each member M_i of **M** is divided into two non-empty subsets of assumptions A_i.[10] The first subset consists in the shared assumptions retained in each model of **M.** Since we are considering GCM, the most important shared component is the assumption of greenhouse gas forcing, *GHG*. The complement of the shared assumptions consists in those that vary between models, which includes parameterizations such as cloud formation and ocean mixing. We then have our predictions concerning the values of variables and parameters. One such important variable in GCM is average surface temperature T.[11] Finally, let us say that two models with are distinct just in case they contain statistically or logically independent assumptions. Last, let us define a notion of robustness with regard to GCM. Consider a set of models $\mathbf{M} = \{M_1, M_2,\ldots,M_n\}$ where each model is composed of *GHG* and at least one distinct A_i.

A prediction T is *robust* over $\mathbf{M}$ if for each $M_i \in \mathbf{M}$, M_i entails T; otherwise, it is *fragile*.

I focus on deductive entailment for simplicity, but we could provide probabilistic relations between T and M_i as well. Lloyd (2015) defines "robustness" differently than I do here. On her view, a prediction is robust when there is independent evidence for T, A_i for $i = 1, 2, \ldots, n$ (though we are unsure which is correct), and *GHG* (Lloyd 2015, 64). Her definition is logically stronger than the one offered here. If a result is robustness on her proposal, it, along with the other parts of the model, are confirmed whereas on the one I offered, that is left open.

Robustness can be given a causal gloss as well. Suppose we compare two models M_i and M_j where the former has an assumption that a causal factor C is present while the other lacks this assumption (or has the that C is absent, or takes zero as its value, etc.). Thus, C is *causally relevant* to T if T is fragile over $\mathbf{M} = \{M_i, M_j\}$, and C is *causally irrelevant* to T if T is robust over $\mathbf{M} = \{M_i, M_j\}$ when $i \neq j$. That is, C is causally irrelevant to T if adding or removing C from our model does not alter whether T is implied. Likewise, C is causally relevant to T if adding or removing C to our model does alter whether our model implies T. Of course, robustness analysis alone does not carry causal implications but can when appropriate causal information is included. Note that this is precisely what adding *GHG* to our GCM does. We cannot accurately account for increases in average surface temperature if we include only natural forcings in our GCM (Randall et al. 2007, 600).

By way of a summary then, when we consider a family of models in which its assumptions are divided into disjoint subsets of shared assumptions and assumptions that vary across models, we can determine whether a prediction is robust or not. If it is, then we can respond to our skeptic's question.

Question: Why accept T since it depends on a false assumption?
Answer: T does not depend; it is *robust*.

If we are to apply this approach to GCM, then we have to articulate the components of $\mathbf{M}$ in climate science and so we turn to this issue.

In a GCM, our family has a core consisting in the following: Navier-Stokes equations of motion, hydrostatic equation, continuity equation, thermodynamic energy equation, equations of state, and specifically greenhouse gas forcings.[12] However, there are a variety of assumptions that differ across models, including forcing agents, parameterizations, and grid type. We can see how GCMs vary by consider the following tables composed by Pirtle, et al. (2010) from Randall (2007, 597–599) (Tables 10.1 and 10.2).

Let's consider an example of model similarity and dissimilarity from the Hadley Centre with UKMO-HadCM3 and UKMO-HadGEM. Both models share the same GHG forcings but differ in other ways. For example, they differ with regard to aerosol forcings, atmospheric resolution, atmospheric layers, ocean resolution, ocean layers, and grid type. However, to conduct a robustness analysis, we need to find common predictions across models. Are there such predictions?

There are common predictions across these models (IPCC AR4 2007, 687). Consider the following:

T: Average surface temperatures have increased over the twentieth century.

We can see that this prediction is robust over models that include *GHG* forcing; specifically, 14 models simulated 58 times predict it. However, it is fragile with respect to models that include only natural forcings; specifically five models simulated 19 times fail to reproduce it (Fig. 10.2).

If one was suspicious of *T* because of an idealization with regard to forcings, atmospheric resolution, atmospheric layers, ocean resolution, ocean layers, or grid type, can one remove the doubt regarding those idealizations with robustness analyses?

If one requires that the models be exactly the same save one difference in A_i, then UKMO-HadCM3 and UKMO-HadGEM would not be susceptible to such a robustness analysis. In addition to differences with regard to atmospheric and oceanic layers, atmospheric and oceanic resolution, and grid type, these two model families differ with respect to forcings including aerosols but also land use. That is, there is no single

Table 10.1 Comparison of GCM in IPCC AR4 2007 (From Pirtle et al. 2010)

List of general circulation models included in the IPCC AR4, with key characteristics							
Model details			Atmospheric resolution (lat/long)	Atmospheric layers	Oceanic resolution (1 at/long)	Oceanic vertical layers	Grid type
Originating group(s)	Country	CMIP3 ID					
Beijing Climate Center	China	BCC-CM1	1.9° × 1.9°	16	1.9° × 1.9°	30	
Bjerknes Centre for Climate Research	Norway	BCCR-BCM2.0	1.9° × 1.9°	31	0.5–1.5° × 1.5°	35	
National Center for Atmospheric Research	USA	CCSM3	1.4° × 1.4°	26	0.3–1° × 1°	40	Eulerian spectral transform
Canadian Centre for Climate Modelling and Analysis	Canada	CGCM3.1(T47)	3.75° × 3.75°	31	1.9° × 1.9°	29	Spectral transform
Canadian Centre for Climate Modelling and Analysis	Canada	CGCM3.1(T63)	2.8° × 2.8°	31	1.4° × 0.94°	29	Spectral transform
Meteo-France/ Centre National de Recherches Meteorologiqu.es	France	CNRM-CM3	~1.9° × 0.9°	45	0.5–2° × 2°	31	Semi-lagrangian semi-implicit time integration with 30 mn time-step, 3 h time step for radiative transfer

(*continued*)

Table 10.1 (continued)

List of general circulation models included in the IPCC AR4, with key characteristics							
Model details			Atmospheric resolution (lat/long)	Atmospheric layers	Oceanic resolution (1 at/long)	Oceanic vertical layers	Grid type
Originating group(s)	Country	CMIP3 ID					
CSIRO Atmospheric Research	Australia	CSIRO-Mk3.0	~1.9° × 1.9°	18	0.8° × 1.9°	31	Spectral for some variables, lagrangian for others, leapfrog
Max Planck Institute for Meteorology	Germany	ECHAM5/MPI-0 M	~1.9° × 1.9°	31	1.5° × 1.5°	40	Spectral transform method, leapfrog timestep scheme
University of Bonn (Germany), KMA (Korea) and M&D Group[a]	G/K	ECHO-G	~3.9° × –3.9[a]	19	0.5°–2.8° × 2.8°	20	
LASG/Institute of Atmospheric Physics	China	FGOALS-g1 0	~2.8° × 2.8°	26	1° × 1°	16	Finite difference, semi-implicit time

(*continued*)

Table 10.1 (continued)

List of general circulation models included in the IPCC AR4, with key characteristics

Model details							
Originating group(s)	Country	CMIP3 ID	Atmospheric resolution (lat/long)	Atmospheric layers	Oceanic resolution (1 at/long)	Oceanic vertical layers	Grid type
US Dept. of Commerce/ NOAA/ Geophysical Fluid Dynamics Laboratory	USA	GFDL-CM2.1	2.0° × 2.5°	24	0.3–1° × 1°	–>	B-grid scheme
US Dept. of Commerce/ NOAA/ Geophysical Fluid Dynamics Laboratory	USA	GFDL-CM2.0	2.0° × 2.5°	24	0.3–1° × 1°		B-grid scheme
NASA/Goddard Institute for Space Studies	USA	GISS-AOM	3° × 4°	12	3° × 4°	16	C-grid scheme
NASA/Goddard Institute for Space Studies	USA	GISS-EH	4° × 5°	20	2° × 2°	16	Arakawa B-grid, among others
NASA/Goddard Institute for Space Studies	USA	GISS-ER	4° × 5°	20	4° × 5°	13	Arakawa B-grid, among others

(*continued*)

Table 10.1 (continued)

List of general circulation models included in the IPCC AR4, with key characteristics							
Model details			Atmospheric resolution (lat/long)	Atmospheric layers	Oceanic resolution (1 at/long)	Oceanic vertical layers	Grid type
Originating group(s)	Country	CMIP3 ID					
Institute for Numerical Mathematics	Russia	INM-CM3.0	4° × 5°	21	2° × 2.5°	33	Finite difference (arakawa 1972), semi-implicit
Institut Pierre Simon Laplace	France	IPSL-CM4	2.5° × 3.75°	19	2° × 2°	31	Finite difference equations, leapfrog time approach
University of Tokyo, NIES, and JAMSTEC[b]	Japan	MIROC3.2(hires)	~1.1° × 1.1	56	0.2° × 0.3°	47	Spectral transform
University of Tokyo, NIES, and JAMSTEC[b]	Japan	MIROC3.2(medres)	2.8° × 2.8°	20	0.5–1.4° × 1.4°	43	Spectral transform
Meteorological Research Institute	Japan	MRI-CGCM2.3.2	2.8° × 2.8°	30	0.5–2.0° × 2.5°	23	Spectral transform method, leapfrog timestep scheme, semi-implicit

(*continued*)

Table 10.1 (continued)

List of general circulation models included in the IPCC AR4, with key characteristics							
Model details			Atmospheric resolution (lat/long)	Atmospheric layers	Oceanic resolution (1 at/long)	Oceanic vertical layers	Grid type
Originating group(s)	Country	CMIP3 ID					
National Center for Atmospheric Research	USA	PCM	~2.8° × 2.8°	26	0.5–0.7° × 1.1°	40	Eulerian spectral transform
Hadley Centre for Climate Prediction and Research/Met Office	UK	UKM0-HadCM3	2.5° × 3.75°	19	1.25° × 1.25°	20	Arakawa B-grid, hybrid vertical coordinates. Eulerian advection scheme
Hadley Centre for Climate Prediction and Research/Met Office	UK	UKMO-HadGEMI	~1.3° × 1.9°	38	0.3–1.0° × 1.0°	40	Arakawa C-grid horizontally, Chaney Phillips grid vertically

Sources: IPCC Table 8.1 (Solomon et al. 2007, 597–599) and (PCMDI http://www-pcmdi.llnl.gov/ipcc/model_documentation/ipcc_model_documentation.php)

[a]Meteorological Institute of the University of Bono, Meteorological Research Institute of KMA, and Model and Data group

[b]Center for Climate System Research (The University of Tokyo), National Institute for Environment Studies, and Frontier Research Center for Global Change (IAMSTEC)

Table 10.2 Comparison of GCM in IPCC AR4 2007 (From Pirtle et al. 2010)

	Forcing agents						Forcing agents											
	Greenhouse gases						Aerosols										Other	
Model details CMIP3 I.D.	CO_2	CH_4	N_2O	Stratospheric ozone	Tropospheric ozone	CFCs	SO_4	Urban	Black carbon	Organic	Nitrate	1st Indirect	2nd Indirect	Dust	Volcanic	Sea salt	Land use	Solar
BCC-CM1	Y	Y	Y	Y	Y	Y	Y	0	0	0	0	0	0	0	Y	0	Y	Y
BCCR-BCM2.0	Y	Y	Y	Y	Y	Y	Y	Y	0	0	0	0	0	Y	0	Y	Y	Y
CCSM3	Y	Y	Y	Y	Y	Y	Y	0	Y	Y	0	0	0	Y	Y	Y	0	Y
CGCM3,1(T47)	Y	Y	Y	Y	Y	Y	Y	0	0	0	0	0	0	Y	Y	Y	Y	Y
CGCM31(T63)	Y	Y	Y	Y	Y	Y	Y	0	0	0	0	0	0	Y	y	Y	Y	Y
CNRM-CM3	Y	Y	Y	Y	Y	Y	Y	Y	G	0	0	0	0	Y	0	Y	0	0
CSIRO-Mk30	Y	Y	Y	Y	Y	Y	Y	0	0	0	0	0	0	Q	0	0	0	Y
ECHAM5/ MPI-OM	Y	Y	Y	Y	Y	Y	Y	0	0	0	0	Y	0	0	0	0	0	0
ECHO-G	Y	Y	Y	Y	Y	Y	Y	0	0	0	0	Y	0	0	Y	0	0	Y
FGOALS-g1.0	Y	Y	Y	Y	Y	Y	Y	0	0	0	0	0	0	0	0	0	0	Y
GFDL-CM2.1	Y	V	Y	Y	Y	Y	Y	0	Y	Y	0	0	0	Y	Y	Y	Y	Y
GFDL-CM2.0	Y	V	Y	Y	Y	Y	Y	0	Y	Y	0	0	0	Y	V	Y	Y	Y
GISS-AOM	Y	Y	Y	Y	Y	Y	Y	0	Y	Y	Y	0	Y	Y	y	Y	Y	Y
GISS-EH	Y	Y	Y	Y	Y	Y	Y	0	Y	Y	Y	0	Y	Y	Y	Y	Y	Y
GISS-ER	Y	Y	Y	Y	Y	Y	Y	0	Y	Y	Y	0	Y	Y	Y	Y	Y	Y
INM-CM30	Y	Y	Y	Y	Y	0	Y	0	0	0	0	0	0	0	Y	0	0	Y
IPSL-CM4	Y	Y	Y	0	0	Y	Y	0	0	0	0	Y	0	0	0	0	0	0
MIROC3.2 (hires)	Y	V	Y	Y	Y	Y	Y	0	Y	Y	0	Y	Y	Y	Y	Y	Y	Y
MIROC3.2 (medres)	Y	Y	Y	Y	Y	Y	Y	0	Y	Y	0	Y	Y	Y	Y	Y	Y	Y
MRI-CGCM2.32	Y	Y	Y	Y	Y	Y	Y	0	0	0	0	0	0	0	Y	0	0	Y
PCM	Y	Y	Y	Y	Y	Y	Y	0	0	0	0	0	0	0	Y	0	0	Y
UKM0-HadCM3	Y	Y	Y	Y	Y	Y	Y	0	G	0	0	Y	0	0	Y	0	0	Y
UKMO-HadGEM1	Y	Y	Y	Y	Y	Y	Y	0	Y	Y	0	Y	Y	0	Y	Y	Y	Y

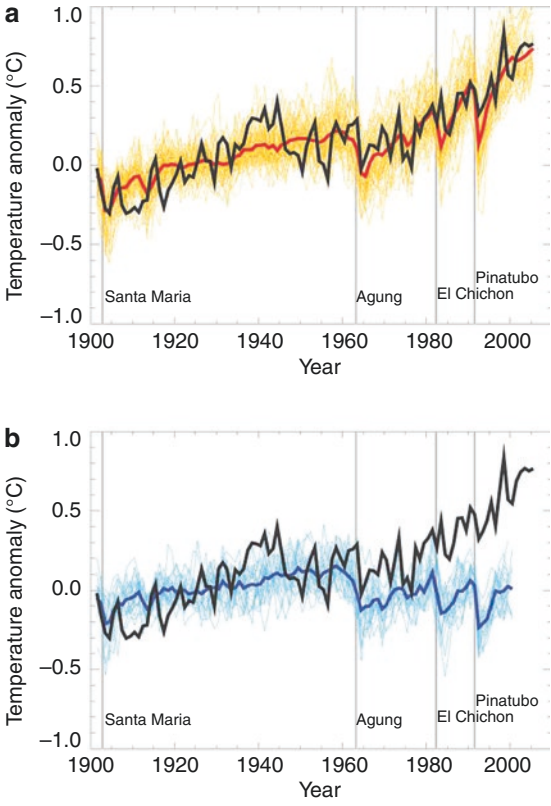

Fig. 10.2 Average surface temperatures compared with GCM with anthropocentric and natural forcing and with GCM with only natural forcing (From Randall et al. 2007)

assumption by which they differ. Consider a different example—the Beijing Climate Center's BCC-CM1 and the National Center for Atmospheric Research's PCM. They are exactly the same with respect to forcings save one, land use. Thus, one could use a robustness analysis with regard to this pair and T to remove doubts regarding models or parameterizations regarding land use. Unfortunately, even this example has a few differences with regard to non-forcing assumptions. Hence, if one is to relieve doubts regarding some model idealization and one requires a robustness analysis be done between two models that are the same except

with regard to that idealization, then climate modelers have work to do. However, this is work that they *can* do.

Though I think this is a reasonable explication of robustness and how it can be used to deal with worries concerning idealizations, there is a substantial worry that needs to be considered.

10.4 Contextualism and Two Types of Robustness Analysis

Suppose our skeptic objects to the above-mentioned robustness analysis as follows. In some cases, we replace an idealized assumption with another assumption and show the alternative model makes the same prediction. However, why should this robustness analysis relieve my worry when you have replaced one idealization with another idealization? We can articulate the argument in the format offered above as follows. Suppose for a prediction T and idealization A_i such that our member M of **M** with A_i entails P, we substitute A_j such that M with A_j entails P. We are also supposing the M of **M** differ *only* with regard to A_i and A_j. Hence, our worries regarding A_i are relieved by replacing it with A_j. However, A_j is idealized or it is true. If it is an idealization, then we must find another idealization A_k to replace it with such that it when conjoined with M implies T. As we can see this continues ad infinitum unless we can find an assumption that is true and when conjoined with M entails T. If we further assume idealization is inescapable as suggestion in §2, then it we cannot remove our skeptic's doubts by robustness analysis.

One response to the regress argument is what I will term a "contextualist" response. Epistemological contextualists often claim that whether one knows or is justified in believing a proposition depends on what standards are at work.[13] For example, Keith DeRose has suggested this with his bank thought experiment (DeRose 1992). Suppose a husband and wife are deciding whether to deposit a check on Friday or Saturday. She says that she knows that the bank is open on Saturday because she visited it two weeks ago and it was open until noon. Let's also suppose it will be open on Saturday. DeRose suggests we would claim the wife

knows the bank is open on Saturday provided nothing of grave significance would occur if it in fact it wasn't. However, if we suppose that it is a matter of grave importance that the check is deposited before Monday, then most of us would deny that the wife knows the bank is open on Saturday. One diagnosis of the conviction that she has knowledge in the one case but not the other is that there is a shift in epistemic standards.

So, our skeptic might be worried about a specific idealization, a specific set of idealizations, or idealization per se. As with DeRose's example, our skeptic about GCM raises the epistemic bar far higher than the skeptic regarding grid size. Thus, when conducting robustness analysis, we must distinguish between *relative* versus *absolute* robustness analyses. If we are worried about a specific idealized assumption, then we can remove this worry by replacing it with another assumption, which in conjunction with the substantial core implies this prediction. In this instance, it does not matter whether the replacing assumption is an idealization provided that the skeptic is not worried about it. For lack of a better term, I call this "relative" robustness analysis. Suppose however that our skeptic is not worried about just this idealization but any idealization per se. For example, insofar as the ideal gas law is idealized, our skeptic would worry about it too. The only way to remove this worry is to show that there is some true assumption when conjoined with the substantial core implies the prediction. I will call this the "absolute" robustness analysis. Thus, if we are worried about idealizations per se, we must perform absolute robustness analyses and if we are worried only about some idealizations and not others, then relative robustness analyses will do the trick. The regress argument thus assumes that if there is some idealized assumption A_i when coupled with M entails T that we doubt, then we will have doubts regarding any other idealized assumption A_j which similarly predicts T when conjoined with M. An epistemic contextualist would not accept this supposition.

We can be more precise in diagnosing the error in the regress argument. Consider the following inference schema:

1. $(M \,\&\, A_i) \rightarrow T$
2. $(M \,\&\, A_i)$
3. $\therefore T$

Let's consider a climate modeler and our model skeptic. Suppose both are justified in believing (1). Suppose that our climate modeler and model skeptic believe with justification that M but both doubt A_i. Since our climate modeler is not suspicious of idealizations per se but only this one, we can remove their doubt by replacing it with A_j and conducting the subsequent robustness analysis. However, this is not so with regard to our model skeptic since nothing short of a true assumption will resolve their doubt; that is, we need an absolute robustness analysis. In sum, the epistemic contextualist suggests our climate modeler is justified in believing T on the basis of her justification that (1), their justification that M, and a relative robustness analysis that replaces A_i with A_j. Our model skeptic is not justified in believing that T since unlike our modeler the relative robustness analysis does not remove their doubt.

One might object to the above epistemic contextualism by noting that different epistemic communities may have different standards in operation, but this is merely a sociological fact without normative significance. If invariantism is true, then our climate modeler, though operating with less exacting standards, is simply not justified in his belief that T.[14] One way of responding to this objection is by considering what contextual factors affect the epistemic status of claims. Philosopher Michael Williams (1996, 2001) has done much to explore these factors and we should consider some of what he says. First, there is what he terms *intelligibility constraints*. In order for doubt to even make sense, we must be entitled to believe some propositions as true. As Williams puts this point, "To be intelligible at all – and not just to be reasonable – questioning may need a *lot* of stage-setting" (2001, 160). Second, there are *methodological constraints* that require that certain doubts be excluded so that certain questions can be asked and answered. Those propositions that are exempted from doubt are the result of *methodological necessities*. For example, if one doubts whether the Earth existed five minutes ago, then one cannot engage in paleoclimatology. Or, if one wants to inquire into complex systems such as the Earth's climate, one must abstract and idealize.[15] Third, there are *economic necessities*. If we require that very unlikely errors should be ruled out, then our standards will be very high with regard to knowledge and

justification of certain propositions. However, if the benefits are great and it costs little, then our standards will be relaxed.[16]

As a matter of fact, I take it that climate scientists are not worried about all the idealizations present in their models. For example, though the background physical core of their models is idealized, I take it that it is sufficiently well tested to be immune to serious doubt. Put differently, if A_i is false, then $P(A_i) = 0$ and thus $P(M \,\&\, A_i|T) = 0$ given Bayes Theorem. Confirmation of our idealized models cannot even get off the ground. However, given the contextual standards as work, it is sometimes warranted to regard idealizations as "true enough" (Elgin 2004; Teller 2001). That is, $Pr(A_i) \neq 0$; assigning a zero prior probability would be unreasonable, and confirmation is possible. Note as well, this alleviates our having to alleviate worries about idealizations in a Millian manner (i.e. replacing the doubted assumptions one at a time with nothing else changed). Insofar as *all* of A_i are true enough—have non-zero probabilities—our models can be confirmed by varieties of evidence (Lloyd 2009).

However, it is surely correct that idealizations specific to GCMs do raise serious worries and thus when we can perform relative robustness analyses they should ameliorate the skeptic's worries. In addition, given intelligibility and methodological constraints, we cannot pursue questions of interest in the sciences if we do not allow for idealization. Model building presupposes that we idealize. I am reminded of something the pragmatist Charles Sander Peirce wrote, "Let us not pretend to doubt in philosophy what we do not doubt in our hearts."

10.5 Conclusion

In this chapter, first I provided a framework in which to understand model building and evaluation including that of GCM. Second, I offered an explication of model robustness and applied to climate modeling. Third, I considered an objection to the above account of robustness centering on the epistemic status of idealization arguing that a form of epistemic contextualism could turn back the objection and is independently plausible.

Notes

1. Climate scientists refer to general circulation models as "GCM"; however, when a model includes atmospheric and oceanic components, they are referred to as "AOGCM." For simplicity, I will refer to all such models as "GCM."
2. For a useful survey of the relevant processes, see Neelin (2010), chapter 2.
3. It is worth noting that not every abstraction is an idealization or every idealization an abstraction. For example, a representation might not include all of the causal variables but say only true things about the ones it includes. Likewise, a representation might not omit any causally relevant variable but distort what is says about them.
4. On the semantic view of theories, philosophers of science assume that models are abstract objects such relational structures, phase spaces, and so on. Here I assume they are propositions (though not propositions axiomatically arranged per the received view of theories). Of course, anything I say here can be understood in one's preferred view of theories and propositions.
5. For a useful discussion, see Neelin (2010), chapter 3.
6. I have argued that models also serve as heuristics for certain purposes (Odenbaugh 2005). That is, untested or disconfirmed models are used to explore possibilities, serve as simple baselines, and provide conceptual frameworks. Climate models of course can do this as well—for example, see the simple layer model in Archer (2012), chapter 2. However here I am concerned with model evaluation in the narrow sense, i.e. confirmation and disconfirmation.
7. For example, if one collects data from a system at a time and then a week later, are these different data sets? Presumably questions like this will partially depend on the questions one is asking.
8. The philosophical status of laws such as the conversation of mass and the ideal gas law is of course controversial. However, when modelers use the term "law," we need not assume that they mean what philosophers do, e.g. natural necessities or relations between universals.
9. My approach to model robustness is largely inspired by the work of William Wimsatt (2007) and has been developed in Odenbaugh (2011) and Alexandrova and Odenbaugh (2011). Additionally, Michael Strevens (2008), Michael Weisberg (2006), Jim Woodward (2006) have provided important analyses. With regard to climate modeling and robustness

analysis, I have been especially influenced by Elisabeth Lloyd (2010, 2015). For an interesting overview of model robustness in the context of climate modeling, see Wendy Parker (2011). Parker considers a variety of explications of model robustness; however, I would argue that the account of model robustness and the queries to which it is put is not found in her analysis and thus avoids her worries.

10. Strictly speaking, M_i will be sub-types of **M** since they will be unspecified.
11. With regard to GCM, our prediction will not be a *point* prediction; rather, it will be that some variable takes a value in some range. Or, it will be a configuration of such variables such that say average surface temperature is increasing over some set of times.
12. In effect, a set of subsidiary models (or model types) becomes a single model (or model type). Note that this means that whether a given assumption or set of them are ad hoc can change through time.
13. Epistemological contextualism is classified as substantive or semantic where the former concerns whether one knows or is justified in believing a proposition with respect to varying standards whereas the latter concerns whether "knowing" or "justification" is context-sensitive. Here I am only concerned with substantive epistemological contextualism.
14. Invariantism is simply the claim that correct epistemic standards do not change with context.
15. Williams argues that these types of constraints are not merely practical or due to relaxed standards but are the result of the "logic of inquiry."
16. If ethical or political costs of global climate change filter into model evaluation, then these norms can influence how skeptical we are (see Biddle and Winsberg 2010). For example, if we are reluctant to bear economic burdens through carbon taxes, then we may hold GCM to high standards. Alternately, if we are very worried about climate impacts on developing nations and future generations, we may want to err on the side of causation. In effect, our model skepticism becomes ethically and politically infused (see Odenbaugh 2010).

References

Alexandrova, Anna, and Jay Odenbaugh. 2011. Buyer Beware: Robustness Analyses in Economics and Biology. *Biology and Philosophy* 26 (5): 757–771. https://doi.org/10.1007/s10539-011-9278-y.

Archer, David. 2012. *Global Warming: Understanding the Forecast*. Hoboken: Wiley.

Biddle, Justin, and Eric Winsberg. 2010. Value Judgements and the Estimation of Uncertainty in Climate Modeling. In *New Waves in Philosophy of Science*, ed. P.D. Magnus and Jacob Busch, 172–197. Basingstoke: Palgrave Macmillan.

Callender, Craig, and Jonathan Cohen. 2006. There Is No Special Problem About Scientific Representation. *Theoria. Revista de Teoría, Historia Y Fundamentos de La Ciencia* 21 (1): 67–85.

DeRose, Keith. 1992. Contextualism and Knowledge Attributions. *Philosophy and Phenomenological Research* 52 (4): 913–929. https://doi.org/10.2307/2107917.

Elgin, Catherine Z. 2004. True Enough*. *Philosophical Issues* 14 (1): 113–131. https://doi.org/10.1111/j.1533-6077.2004.00023.x.

Hausman, Daniel. 2008. Why Look Under the Hood? In *Essays in Philosophy and Economic Methodology*. Cambridge: Cambridge University Press.

Lenhard, Johannes, and Eric Winsberg. 2010. Holism, Entrenchment, and the Future of Climate Model Pluralism. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 41 (3): 253–262. https://doi.org/10.1016/j.shpsb.2010.07.001. Special Issue: Modelling and Simulation in the Atmospheric and Climate Sciences.

Levins, Richard. 1966. The Strategy of Model Building in Population Biology. *American Scientist* 54 (4): 421–431.

Lloyd, Elisabeth Anne. 1994. *The Structure and Confirmation of Evolutionary Theory*. Princeton: Princeton University Press.

Lloyd, Elisabeth A. 2009. I—Elisabeth A. Lloyd: Varieties of Support and Confirmation of Climate Models. *Aristotelian Society Supplementary Volume* 83 (1): 213–232. https://doi.org/10.1111/j.1467-8349.2009.00179.x.

———. 2010. Confirmation and Robustness of Climate Models. *Philosophy of Science* 77 (5): 971–984.

Neelin, J. David. 2010. *Climate Change and Climate Modeling*. Cambridge: Cambridge University Press.

Odenbaugh, Jay. 2005. Idealized, Inaccurate but Successful: A Pragmatic Approach to Evaluating Models in Theoretical Ecology. *Biology and Philosophy* 20 (2–3): 231–255. https://doi.org/10.1007/s10539-004-0478-6.

———. 2011. True Lies: Realism, Robustness, and Models. *Philosophy of Science* 78 (5): 1177–1188. https://doi.org/10.1086/662281.

———. 2010. Philosophy of the Environmental Sciences. In *New Waves in the Philosophy of Science*, ed. P.D. Magnus and Jacob Busch. Basingstoke: Palgrave Macmillan.

Parker, W.S. 2006. Understanding Pluralism in Climate Modeling. *Foundations of Science* 11 (4): 349–368. https://doi.org/10.1007/s10699-005-3196-x.

Parker, Wendy S. 2010. Predicting Weather and Climate: Uncertainty, Ensembles and Probability. *Studies in History and Philosophy of Modern Physics* 3 (41): 263–272. https://doi.org/10.1016/j.shpsb.2010.07.006.

———. 2011. When Climate Models Agree: The Significance of Robust Model Predictions. *Philosophy of Science* 78 (4): 579–600. https://doi.org/10.1086/661566.

Pirtle, Zachary, Ryan Meyer, and Andrew Hamilton. 2010. What Does It Mean When Climate Models Agree? A Case for Assessing Independence Among General Circulation Models. *Environmental Science and Policy* 5 (13): 351–361. https://doi.org/10.1016/j.envsci.2010.04.004.

Randall, D.A., R.A. Wood, S. Bony, R. Colman, T. Fichefet, J. Fyfe, V. Kattsov, et al. 2007. Climate Models and Their Evaluation. In *Climate Change 2007: Report of the Intergovernmental Panel on Climate Change*, ed. Solomon Susan, D. Qin, Martin R. Manning, Z. Chen, M. Marquis, K. Avery, M. Tignor, and H. Miller. Cambridge, UK: Cambridge University Press.

Sorensen, Roy. 2012. Veridical Idealizations. In *Thought Experiments in Science, Philosophy, and the Arts*, ed. Melaine Frappier, Letitia Meynell, and James Brown. Routledge University Press.

Strevens, Michael. 2008. *Depth: An Account of Scientific Explanation*. Cambridge MA: Harvard University Press.

Teller, Paul. 2001. Twilight of the Perfect Model Model. *Erkenntnis* 55 (3): 393–415. https://doi.org/10.1023/A:1013349314515.

Weisberg, Michael. 2006. Robustness Analysis. *Philosophy of Science* 73 (5): 730–742. https://doi.org/10.1086/518628.

Williams, Michael. 1996. *Unnatural Doubts: Epistemological Realism and the Basis of Scepticism*. Princeton: Princeton University Press.

_____. 2001. *Problems of Knowledge*. Oxford University Press.

Wimsatt, William C. 2007. *Re-engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*. Cambridge, MA: Harvard University Press.

Woodward, Jim. 2006. Some Varieties of Robustness. *Journal of Economic Methodology* 13 (2): 219–240. https://doi.org/10.1080/13501780600733376.

Part III

Climate Models as Guides to Policy

11

Climate Model Confirmation: From Philosophy to Predicting Climate in the Real World

Reto Knutti

11.1 What Is a Climate Model, and Why Do We Need One?

As scientific problems get more complex and computers get faster, numerical models are getting ubiquitous. The reasons often cited for using models are that it is too complicated, time-consuming, impossible, or dangerous to do a real-world experiment. With the global climate, all of these criteria apply, although we are of course performing a very large and potentially dangerous and costly experiment by increasing atmospheric greenhouse gas concentrations to levels unprecedented in nearly a million years, changing land use, rerouting rivers and extracting groundwater, and polluting oceans and atmosphere. The difference is that it is not very controlled and coordinated, everything changes at the same time, and the experiment may be largely irreversible on human timescales (Solomon et al. 2009). A proper scientific experiment, however, should ideally be reproducible, multiple experiments should be possible to test

R. Knutti (✉)
Department of Environmental Systems Science, Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland

E.A. Lloyd, E. Winsberg (eds.), *Climate Modelling*,
https://doi.org/10.1007/978-3-319-65058-6_11

different hypotheses, and only one boundary condition should be changed at a time. A numerical model is useful here because it can be run multiple times (and often cheaply) to understand how the response changes for different boundary conditions or model configurations. It is through this process of "digging into" the problem that scientists learn about the behavior of the model, and ultimately the target system. Rather than doing one single experiment, we learn by sequentially and systematically testing different hypotheses or processes, comparing to data, and by changing the model to understand its behavior. But of course, the model remains a model, and we have to make inferences from what we learn from the model to the real world.

The above comments apply to many environmental situations where models are used. In this chapter, I will focus on models that attempt to represent the global climate system. They describe the atmosphere, ocean, sea ice, and land processes, and are based on fundamental principles of physics like conservation of energy, mass, and angular momentum. Apart from these well-established laws, there are, for example, small-scale or biological processes which have to be described in simplified and aggregated ways. Combined with initial conditions and boundary conditions, this system of equations is solved numerically on a spatial grid and the state of the climate is integrated forward in time on high-performance computers (Mueller 2010). Such models are used to understand processes, simulate the past climate, and predict the future.

11.2 Why Are Models Uncertain?

It is tempting to answer that this question makes no sense, because a model is *not* uncertain. It is also not wrong. A model, if properly specified, is a series of equations, assumptions, boundary and initial conditions, and if properly solved numerically produces deterministic results. In other words, the same model run twice on the identical hard- and software will produce bit-for-bit identical results, and therefore it is not uncertain. The exception may be where random numbers are used, but for the same sequence of random numbers, the result should still be identical. When people say a model is uncertain, they mean either how much

the results of the model vary as alternative plausible choices in the model are tested, or more commonly, how well the model represents the real world for the purpose in question. For the second, it may be more appropriate to call this "system representation uncertainty" to highlight that it is the relationship between the model and reality that is unclear, that is, how relevant or representative the model is to learn about the features of interest in reality. Models are representations of something else, and it is this representation uncertainty that I will explore in this chapter.

Predictions of various models are biased compared to the actual outcome of the target system for several reasons. The Earth climate system is influenced by essentially everything out there, from the flap of a butterfly wing to the gravitational field of Jupiter. It covers spatial processes from micrometers and fractions of seconds for cloud processes to thousands of kilometers and millions of years for the ocean circulation and movement of continents. There are various ways to categorize the different factors affecting the model results, but a straightforward separation of uncertainty is into model structural error, numerical approximations, parameterizations, natural variability due to initial conditions, emission scenario, boundary conditions, and observational data uncertainty.

First, the structure of the model differs from reality in that the model can only describe a subset of the components and interactions that exist. All models are incomplete, but that is often misinterpreted as being "wrong." It is the very purpose of a model to describe a simplified and reduced form of that open system, a form that can be "experimented" with in a controlled way. Every model of the Earth has to draw a line somewhere and ignore certain components, scales, or interactions. Most models used for predicting the climate of the next century, for example, do not contain an interactive description of ice sheets, and continents are assumed to be fixed, because those parts are assumed not to vary strongly over that period.

Second, the equations for climate models cannot be solved analytically. To solve them numerically the Earth has to be divided into finite grid cells with typical dimensions of tens to hundreds of kilometers for global models. Solving equations on a grid introduces numerical inaccuracies due to the coarse grid and the limited precision with which the computer handles numbers (Mueller 2010). That however is rarely the dominating

problem. It is the fact that many relevant processes happen on scales much smaller than the grid scale. Resolution of a few kilometers is required to explicitly simulate atmospheric convection, and realistic cloud properties only emerge with resolutions of meters. A doubling of the horizontal resolution increases the computational cost by about an order of magnitude. So even if computing speeds continue to increase as they did in the past, it will be many decades for global models to reach resolutions of meters.

Third, the limited resolution implies that many small-scale processes have to be parameterized. A parameterization is a description of the effect of small-scale processes in terms of the available large-scale quantities, without actually resolving the processes (McFarlane 2011). Parameterizations are needed in cases where the actual processes are well known but too complex to simulate, or if their effect is observed but the underlying laws are not sufficiently well understood at the scales resolved, or both. For example, tides and small-scale mixing dissipate large amounts of energy in the ocean, and are parameterized as a diffusion or advection term in coarse resolution models (Knutti et al. 2000). Such parameterizations are often termed "closures" because they close the energy or water cycle on the small scale. The smaller the resolution, the more of those small-scale processes can be simulated explicitly, and the fewer parameterizations are needed. Important parameterizations in global models include cloud microphysics, the boundary layer, radiation, and atmospheric convection. Other parameterizations include the growth of plants, often described as plant functional types, that is, relationships that express how well a plant grows as a function of temperature, moisture, light, and maybe other conditions. In contrast to small-scale mixing where computational cost is the limiting factor, a parameterization of a plant or animal species is limited by understanding. There is simply no fundamental equation to describe how a tree grows. Some parameterizations start with a functional form based on or inspired by physical laws; for example, diffusion for any small-scale mixing through advection and turbulence, with the parameters chosen to match certain observed fields. Others like the plant functional types are largely empirical fits to data. In other words, a parameterization is largely a practical computational simplification, or an empirical description of a poorly known effect. Some

are optional, for example, vegetation could simply be assumed to be constant, whereas other closure schemes are critical. The choice of the parameter value for a given model structure (including the form of parameterizations) is connected to the first two issues discussed above. Whether a parameterization is needed or not depends on the model structure and resolution. The optimal value for a parameterization also depends on resolution.

The fourth main source of uncertainty is natural variability, the fact that weather is unpredictable on timescales beyond a few days. The climatological mean state (or the probability of all states) is largely predictable over several decades, but on shorter timescales natural variability associated with ocean atmosphere interaction and atmospheric circulation (i.e., essentially weather) can be large (Deser et al. 2012a, b; Fischer et al. 2013; Knutti and Sedláček 2013; Mahlstein et al. 2011, 2012), and is largely unpredictable. Uncertainty due to natural variability arises from the fact that the initial conditions for the model are not exactly known. But even if they were, the model would deviate from reality because of simplifications made in the structure, resolution, and parameterizations.

The fifth source of uncertainty for climate prediction are the scenarios, assumptions on future population, energy use, air pollution, land use, and policies (Moss et al. 2010). Such factors do not follow any laws of nature, and are often thought of as choices humanity can make. Model results are therefore often presented as projections, that is, changes in climate conditional on a specific scenario of human decisions (reflected in energy use, etc.).

Finally, boundary conditions like the bathymetry of the ocean or the solar constant are not perfectly known. Observations are also uncertain, but those usually do not enter the model (except in data assimilation) but are used to develop and evaluate the model.

The first three, the choices in the model structure, resolution, and parameterizations, together represent the epistemic uncertainty. They are the core of the representational uncertainty in the model, and for each purpose in questions we need to ask whether these simplifications are appropriate. They reflect missing, incomplete, or imprecise knowledge and technical limitations that at least in principle can be improved.

Natural variability and human behavior are uncertainties that are inherent in the system. But conceptually they are different in that variability is truly unpredictable beyond on timescales of decades and longer, whereas the choice of a pathway for humanity is a choice we make.

The contribution of the different sources of uncertainty to uncertainty in predictions depends on the variable and scale (temporal and spatial) (Masson and Knutti 2011b). The contributions also change over time with natural variability being approximately constant in absolute terms, but decreasing in relative terms as the forced climate change signal (with its uncertainty) emerges from variability (Hawkins and Sutton 2009, 2011; Knutti and Sedláček 2013; Mahlstein et al. 2011, 2012). Other separations are possible, but for many purposes the separation of model structure, numerical approximation, parameterization and parameters, natural variability, and emission scenarios are adequate.

11.3 Why Do We Need More than One Climate Model?

Different purposes require different types of models. Simple models are used to explore many scenarios and for probabilistic projections (Meinshausen et al. 2009; Rogelj et al. 2012), intermediate complexity models are often used for paleoclimate simulations that extend over thousands of years (Claussen et al. 2002), and high-resolution ocean atmosphere models are used to simulate climate change over the twentieth and twenty-first centuries. But even for a particular question and set of processes, different models exist. Strictly they are incompatible; they cannot be true at the same time. But they are usually seen as complementary, because they represent different plausible (although not necessarily equally plausible) approximations to the target system, given some computational constraints, limited and uncertain observations, and incomplete understanding of all processes (Knutti 2008a; Parker 2006). The hope is that we learn more from an ensemble of models than from a single model. For example, there are several ways to parameterize atmospheric convection, and no scheme is clearly superior to the others for all

climatic states, and parameters are not well constrained. As a result, the most recent Coupled Model Intercomparison Projection Phase 5 (CMIP5) (Taylor et al. 2012) included more than 40 models, although some of those are largely duplicates of others, and the set of models cannot be interpreted as necessarily being a representative sample of the uncertainty (see below).

11.4 Model Evaluation, Confirmation, Robustness, and Variety of Evidence

A model can never be proven to be correct or true in a strict sense, that is, as a complete and accurate representation of the real world (Oreskes et al. 1994). But that does not ask the right question for a model. Instead, confirmation is better thought of as a gradual process in building up confidence. But what exactly are we confirming? Are we accumulating confidence in the model itself, or its adequacy for purpose for a specific prediction, or its relationship to the target system for some purpose?

The common way to evaluate a model is to compare in a quantitative way the fit of modeling results to observed data (Flato et al. 2013; Gleckler et al. 2008; Jun et al. 2008; Knutti 2008a, b; Schaller et al. 2011). The climate modeling community mostly uses the term "evaluation", whereas hydrologists often use "validation", but in practice they mean essentially the same. Successful instances of model fit are often uncritically interpreted as confirming the model, but they really are "model performance metrics", that is, numbers that quantify the agreement between simulated and observed data, and it remains to be discussed what they imply for "model quality" for a purpose (Huber et al. 2011; Knutti et al. 2010a, b). Instances of fit could be the result of compensating biases, or overfitting, or could simply be unimportant if the evaluated quantity is unrelated to the prediction of interest. Instances of misfit could result from the fact that the model simulates a different quantity than that observed or from biases in observations, or of different processes in models and observations.

Lloyd (2010) argues that "a model with many instances of fit is much better supported and has a higher probability under a preferred confirmation function than a model with only one or two instances," and that multiple instances of fit of distinct variables provides "stronger evidence for a model than either one of the instances considered individually". This "variety of evidence" argument is largely a Bayesian interpretation of increasing the probability for a hypothesis being correct by accumulating independent data. She also refers to "variety of evidence" when discussing Weisberg (2006), who, in the context of robustness analysis, states that "if a sufficiently heterogeneous set of models for a phenomenon all have the common structure, then it is very likely that the real-world phenomenon has a corresponding causal structure. This would allow us to infer that when we observe the robust property in a real system, then it is likely that the core structure is present and that it is giving rise to the property". Weisberg (2006) refers to a "corresponding causal structure," Lloyd (2009) to "common core (causal) structure", and "evidence for the model" (Lloyd 2010). At first sight, they both appear to argue for support of the model itself, or core parts of it (see Fig. 11.1a). But Lloyd (2010) on another instance clarifies that by saying: "I will use the shorthand of 'models' being confirmed, instead of 'theoretical hypotheses using the models' being confirmed, and will often leave off the purposes."

For a physicist, the model in a narrow sense can be just a mathematical construct of rules, and boundary conditions, or in practice a compiled program based on thousands of lines of code. In this case, it fundamentally cannot and does not need to be confirmed. If the model in a wider sense is a representation of the target system, then it is an *interpreted* mathematical structure, where at least some terms in the model are supposed to represent (or map to) particular features of the world. In this case, we confirm the hypothesis that the world has a similar causal structure in certain respects.

The actual core of the model is a set of bytes in the computer versus air and water moving in the atmosphere, so there is no material similarity. What could be confirmed is either the prediction of a model (i.e., the numerical result generated by the model) or that the model structure resembles a target system in some particular respect. But in practice, even confirming such core relationships is difficult, as it would require us to

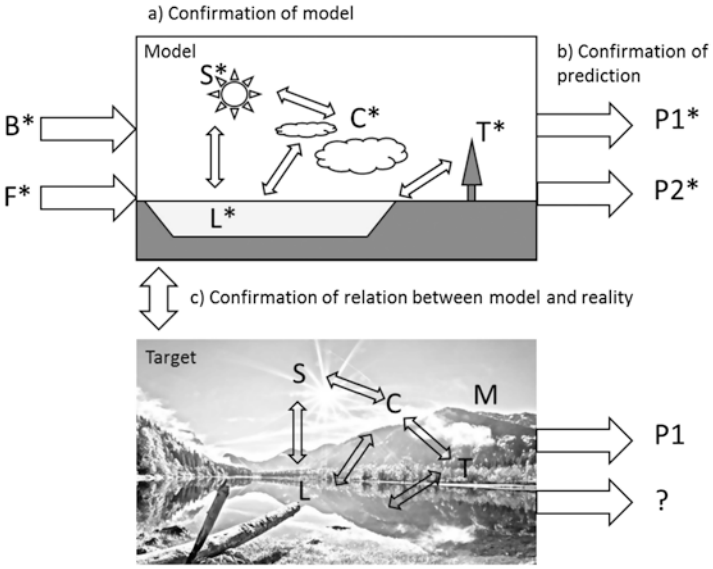

Fig. 11.1 (*Top*) A model of the climate with the sun (S*), clouds (C*), a lake (L*), and trees (T*) that takes some boundary conditions (B*) and forcing (F*) to predict several quantities ((P1*, P2*), *bottom*) the corresponding target system, with the main difference that it includes more that the model (e.g., mountains (M)) and only some parts are observed (P1) but not others (P2). The question is whether we confirm (**a**) the model (equation, structure), (**b**) its prediction, (**c**) the relationship between the two, or a combination, e.g., the model structure being sufficiently similar to the target such that P1* is an adequate estimate of P1

build a set of models with the same causal core relationship and vary everything around it, which is practically impossible. We cannot exclude that all models have a similar simplification that cause a certain behavior. In fact some features are known to be similarly wrong in most if not all models (e.g., the simulations of the Innertropical Convergence Zone [ITCZ]), so the argument of robustness, although used often (Fischer and Knutti 2013; Knutti and Sedláček 2013), needs to be made carefully because of common structural biases and model dependencies (Knutti et al. 2013; Parker 2011, 2013; Pirtle et al. 2010).

Parker (2009a) stresses that we do not provide support for a model as such, but for its adequacy for a particular purpose. Testing adequacy for purpose for a forecast is straightforward in the case of repeated verification

of simulation results. If we predict tomorrow's weather for several decades, and if the prediction repeatedly matches what actually happened within the desired accuracy in certain quantities (usually defined as skill scores), then we have good evidence that the model is adequate for predicting tomorrow's weather (as measured by the defined scores, and under the assumption that tomorrow's weather is predictable in the same way as it was in the past). But Parker (2009a) argues that we can also confirm (i.e., support, but not guarantee to be true) the claim that the model is adequate for making skillful predictions of that sort in a broader range of cases that have not yet been checked.

In most cases, we do not have the option of repeated verification of simulation results, or in fact no verification at all. We may have a prediction for the typical seasonal cycle in temperature at different grid points (prediction P1, see Fig. 11.1) that we can evaluate with observations, but we want to predict sea ice cover in the year 2100 (prediction P2, see Fig. 11.1), for which there are no observations. So what we need is an argument that success on P1 supports the hypothesis that it will be successful on P2. Parker (2009a) discusses many of the difficulties in finding the criteria that would ensure adequacy for purpose, and understanding the relevant processes that determine P1 and P2 is one of them. I argue here that process understanding is the primary route to determine adequacy for purpose, and will flesh this out more in the following.

Before we start, we need to define what we mean by what is adequate. For some questions, no model will be adequate. A weaker form would be to demonstrate skillfulness, which could be defined as the model providing more information than one would have had without the model. Skillfulness would imply there is a benefit of using the model information, but it may still not be adequate given certain standards. For example, cell phone tracking information is skillful in that it hints to a person being somewhere at a given point in time, but it is inadequate for the purpose of declaring a person responsible for a murder, whereas DNA of a suspect found on a dead body may be sufficiently adequate for that purpose. But let's assume that for a given purpose, we can define the criteria for adequacy; for example, we need a forecast of global temperature for the year 2050 with an uncertainty less than 0.5 °C.

Because there is no direct evaluation of the forecast, the key is the similarity of the relevant processes. I argue what we try to confirm is that *for the particular purpose of interest, (1) the relevant quantitative relationships or interaction between different parts or variables that emerge from the inner structure of the model are sufficiently similar to those in the target system, (2) they will remain so over time and beyond the range where data is available for evaluation, and (3) no important part or interaction, either known or unknown, is missing*.

This is the core argument in its general form, and it deserves a more detailed discussion. The first point, the similarity of the relationship, means that that "things behave the same way": a change in quantity or component A corresponds or leads to a change in B that sufficiently closely matches reality (see Fig. 11.1). Material similarity is irrelevant; that is, whether the computer is built from the same material as the target system is unimportant. Whether the model actually resolves the processes connecting A and B is also not relevant, except that if it does not, then point 2 becomes more questionable (see below). So the model is not "similar" to the target in the sense that the parameterization might not resemble anything in the real world, but it may be considered similar in its structure because its parts and behavior are similar. The argument is usually based on both the parts and the interactions. For example, a model could have several pools of carbon with different reservoir sizes corresponding to soil, leaves, stems, and so on, and fluxes representing certain processes that connect them and exchange carbon, like photosynthesis or respiration. Even though these carbon pools are massive oversimplifications of the real world, they reflect that there are different reservoirs that interact in different ways.

The second point is important because we can directly test the model for adequacy in predicting P1, but we actually want to know P2 (see Fig. 11.1). Therefore, the relationship between P1 and P2 needs to be correctly captured beyond the range of both the variables and time where they can directly be tested with observations. Note that the relationship can change; for example, the sea ice albedo feedback (snow and ice melting leading to darker surface, lower reflectance, stronger warming, and amplified melting) changes in magnitude as the area with snow and ice

decreases in the future, but we need to argue that it changes in the model and the target in a similar way based on our understanding of the processes and how they are implemented. The argument about the relationships emerging from the inner structure is trying to address this point. If the inner structure is a polynomial fit, a neural network, or some other statistical tool, then there is little reason to be confident that relationships hold beyond the range where the model is trained and evaluated. They may hold over time but not once the model is applied outside the range of values where it was calibrated. Statistical methods are very powerful at capturing patterns from data, but are often poor at extrapolation. If the core structure (i.e., the correspondence of components like reservoirs of carbon, water masses, and variables) has similarities to the real world, and the interactions are described as (or approximated from) known and well-understood physical principles (conservation laws, chemical reactions, scaling arguments…), then those are likely to hold over a wide range of parameters and climate states, even if the fit to data gets worse. Scientists often prefer approximations based on basic principles over statistical fits. Even if the latter may show better predictions for P1, their relationship to P2 is less clear. Note that many approximations are fine within a range (e.g., linearizations for small perturbations) but eventually fail, a behavior known even for aspects that we would consider as fundamental laws (e.g., gravity, where Newton's law get inappropriate for very small scales). Criteria for confidence in a relationship often include that the response to a small perturbation (e.g., within the observed warming over the last decades) fits with observations, and the response to extreme perturbations is physically plausible (e.g., the snow albedo feedback vanishes when no snow is left).

Point 3 is the question whether we have sufficient understanding of what processes are relevant for the question at hand, and how well those are described in the model. Some processes may not be well observed (e.g., ice sheets before about 1990), some may be observed but poorly understood (e.g., some ecosystems), and some may be quite well understood but are too hard to compute (fluid dynamics on small scales, the mixing effect of which needs to parameterized). We need to argue that those processes not represented in the model are not relevant to the

question at hand, and that we can quantify the effect of those that are poorly represented or understood.

One may argue that this argument about process understanding simply shifts the problem to confirming individual processes which one might consider a submodel. Of course, similar questions arise, but the hope is that individual processes are easier to evaluate because they only work on certain scales, because they are more closely linked to a small number of physical, chemical, or biological processes, and because they can more easily be constrained by observations. For example, photosynthesis and its dependence on various parameters like temperature and humidity could be measured in a greenhouse under controlled conditions, and for various types of plants.

Parker (2009b) argues that "in a computer simulation study, then, scientists learn first and foremost about the behavior of the programmed computer; from that behavior, taking various features of it to represent features of some target system, they hope to infer something of interest about the target system." This is close to the argument made here in the sense that we try to infer the behavior of the real world from a simpler model. But while strictly correct, I argue that we do not want to learn about the behavior of the computer, but about our model. The computer is just a stupid (but fast) machine that calculates what we could do in our head based on the assumptions and equations that we wrote down, if we were fast enough. The computer is made of some material of course, but in contrast to Parker (2009b), I consider this as a purely practical and irrelevant nuisance, and I think about computer simulations more like thought experiments in that we are exploring the consequences of assumptions, represented in the form of a model. The computer introduces numerical errors, but in practice, this is rarely the dominant uncertainty, and given enough resources, it can be minimized if needed. It is an imperfection in much the same way as any measurement in a material experiment is imperfect. We should not ignore it but if needed we can minimize it. We can learn from computer simulations, but they do not generate new knowledge in the sense that all the information is already put into the system, pre-specified as rules. But the knowledge may only be implicit, so we do generate knowledge in the sense that we become aware of it. By changing the rules of the model, we can learn about the

system (either the model or the target), but the simulation does not tell us the rules. The most we can tell is which set of rules is consistent with aspects of the observed behavior in the real system, that is, which set of rules is adequate to describe some aspect of a real system (and in that sense, the computer can inspire us to discover some relationships). To support this, we could build a simple model in which two quantities are linearly related, and we would not need a computer to solve it and make a prediction. Conceptually, I would argue, there is no difference to a complex climate model. We need an equation, some data to calibrate the model, and some boundary or initial conditions, and we learn about consistency of our model with reality. The fact that the complex model needs a computer is mostly a practical nuisance.

11.5 Simplicity Versus Complexity, and the Purpose of Models

A large number of models of different types and complexity exist, from simple box models to full three-dimensional general circulation models (GCMs) that describe the atmosphere, ocean, ice, and land (Claussen et al. 2002). The tendency, in most cases, is to make them more complicated as time progresses. The assumption, often implicit, is that the model will get more realistic by adding more "stuff" to address shortcomings, and eventually that its behavior will converge to the real target system. The model, once it describes everything, would become purpose-independent. In reality, we often realize that the model may get more realistic in terms of matching observed data, but less useful to provide insight; it gets too expensive to operate, too tedious to maintain, and there are too many things happening at the same time so that the scientific understanding is limited.

Held (2005) argues that climate modelers have not been very successful in building hierarchies of models to trace certain behavior across different types of models. He suggests that if we managed, like the biologists, to define the "*E. coli* or *Drosophila melanogaster* of climate models", that is, archetypes of models used by many, we may be better positioned to

understand their most fundamental limitations. Along similar lines, Stevens and Bony (2013) argue for "a deeper understanding and better representation of the coupling between water and circulation, rather than a more expansive representation of the Earth System". Simplicity implies that most of the behavior is explained by a basic process or physical law that is well understood, and that can be traced to parts or processes in the model, or even the governing equations. Many complex systems are governed by simple relationships on the large scale (Held 2005, 2014). Economic models for example often are highly idealized. We may not believe the exact numbers they produce, but they tell us something about the emerging behavior of a few core assumptions. The quote "models are for insight, not numbers" captures this well. At the opposite end of the spectrum, for some purposes we may not know which processes are most relevant, either because our understanding is insufficient, or because they interact in nonlinear ways, or exhibit threshold behavior. We hope that if we describe all processes and parts separately in sufficient detail and accurately, then the emerging behavior of the model, the sum of all parts, should reflect reality. Rather than arguing about what is relevant and needs to be evaluated, the underlying assumption is one of convergence of the model behavior to the real world. Unfortunately, this strategy often fails. Models are stubborn and do their own thing; the complexity becomes overwhelming. The model may get more realistic in some aspect, but with every addition the degrees of freedom get larger, and other limitations become apparent.

I don't think a case can be made for whether simpler or more complex models are more useful, except that each is useful for some purposes but not others, but the purpose discussion often gets lost. But I do think in the battle between simple, targeted, and selective representation on one hand and completeness on the other hand, the climate modeling community is pushing too far toward the latter. We build, use, and present complex models as if they were reality (Lahsen 2005), and forget (or at least are not explicit about the fact) that they are just tools that reflect parts of a real system. Depending on those parts, they are more or less useful to answer specific questions. Predictions of the climate in 2050 may require a full-blown Earth system model with as many processes represented as possible, but our insight into processes and uncertainties

may improve more with *E. coli* climate models. If we complain that a simple model has failed, then in fact we have failed to remember that science and models are targeted toward certain questions, and that we may have considered the wrong set of processes. Understanding model failures may provide more insight than an infinite amount of output of the most complex and expensive model we can build.

11.6 Why Do We Believe Models?

Sometimes we are tempted to believe what the models' simulations are saying is true for the real world (Lahsen 2005), but of course we should not without carefully evaluating the model first. We can believe what the model is predicting in the model world, because it follows from the rules that we specified, but believing that this also applies to the target system requires the similarity of the relationships in model and target system as described above.

There are various ways to accumulate support for these relationships. One is to focus on processes and feedbacks, one at the time, test whether they match between model and reality where they can be evaluated with observations, and provide an argument that this will hold beyond the range tested (Bony et al. 2006). If all processes are tested successfully that way, then the assumption is that the model as a whole will hold as well. We could provocatively call this the nerd's approach. The difficulty here is that the interaction between processes can be subtle, complex, and nonlinear, such that inevitable small biases in one process will propagate into large biases of the whole system. It is common that when different components of a model are put together (e.g., the ocean, atmosphere, and sea ice), they perform worse than they did when tested individually. The other problem is that there are so many parts to the model that it needs a large number of nerds. Because they often do not talk to each other, and none of them understands and cares about the sum of all parts (called the "love factor" by Bjorn Stevens), again the overall performance is not guaranteed. Therefore, the model as a whole needs to be evaluated as well. There is no question that our understanding of many relevant processes in the climate system has greatly improved over the last decades,

and that the representation of those processes in models is much more comprehensive than it used to be. Earlier models, for example, treated the land surface as fixed, whereas now there are very detailed process descriptions of land atmosphere coupling, hydrology, vegetation dynamics, carbon and nitrogen cycles, all the way to urban models embedded in climate models. The question is whether the skill of models as a whole has improved as much as our understanding of the parts.

Another approach is that of "brute force applied statistician," who with the help of a massive computing infrastructure will attempt to increase complexity and resolution of the model and perform an exhaustive evaluation on all possible data, hoping that the model will converge to reality. The assumption, broadly speaking, is that if the model matches all the data, then the underlying structure must be correct, because it is implausibly unlikely to get such a good match by chance or for the wrong reason if the amount of data is much larger than the degrees of freedom in the model. The problems are that this is technically challenging and computationally expensive. It is fundamentally impossible to argue that we will fully converge due to limited observations, natural variability in the observations, and a lack of observations for the actual prediction. We may converge on the things we observed, but not on the prediction if there are no observations that tell us enough about some processes that matter. There may be a process that only matters in the future (e.g., methane hydrates on the ocean ground), so the model could match almost perfectly for the things observed but still be biased for the prediction. The kinds of model evaluation routinely performed are to compare the climatological mean state of the model with observations (e.g., the monthly rainfall at each location) (Gleckler et al. 2008; Reichler and Kim 2008), the variability (e.g., the magnitude and time evolution of the El Niño Southern Oscillation in the tropical Pacific) (van Oldenborgh et al. 2005), the trends observed over the industrial period (e.g., the decline of Arctic sea ice) (Stroeve et al. 2007, 2012), or the response of the model to specific events like large volcanic eruptions (Boer et al. 2007; Gleckler et al. 2006; Soden et al. 2002; Trenberth and Dai 2007). With each generation, models continue to better represent the mean climate state and variability (Knutti et al. 2013; Reichler and Kim 2008), and the amount and quality of data improves as well.

Another comparison to data is by looking at climate states before the industrial period (e.g., the climate response to solar variations in the Holocene, or the climate of the last ice age, or periods even further back; Braconnot et al. 2012; Hargreaves and Annan 2009; Lean et al. 1995; Lean 2010). The advantages are that those provide partly independent information not used in model development, and that the climatic differences are large relative to today. Limitations are that the relevant processes might be different (e.g., ice sheets can be assumed constant for the next decades, but not for an ice age), and that the boundary conditions (radiative forcing) and the data (e.g., sea surface temperature) are limited in time and space and derived from proxy data, which introduces large uncertainties in what "reality" was.

A third approach is to find so-called emerging constraints, that is, to find clear relationships between observables (P1) and predictions (P2) across a wide range of models. If we have data to constrain P1 sufficiently well, then that provides a constraint on the prediction P2 through the relationship found, if we think we have the processes explaining the relationship roughly right in the model. For example, past sea ice trends relate strongly to future sea ice trends (Boé et al. 2009; Mahlstein and Knutti 2012), seasonal albedo feedbacks relate to long-term albedo feedbacks (Hall and Qu 2006), short-term relationships between tropical temperature and CO_2 growth rates constrain the long-term feedbacks (Cox et al. 2013), and some climatological features relate to feedbacks and climate sensitivity (Fasullo and Trenberth 2012; Huber et al. 2011; Sherwood et al. 2014). One could call this the "pragmatic ignoramus" approach, because a priori there is no process understanding needed if the relationship is strong enough. But such relationships may appear by chance and because of structural similarity of models, or they may only hold for certain classes of models but not others (Caldwell et al. 2014; DelSole and Shukla 2009; Masson and Knutti 2013). They are only powerful if we understand why the relationships appear, and if we can argue that the underlying processes are well understood and represented in the models.

A mix of two and three above are methods of detection and attribution, which use a combination of models and statistics to extract an emerging signal or pattern of change that can be attributed to a specific

cause. For example, models are run separately with greenhouse gases only over the historical period, and then again with natural forcings, and with aerosols. Because the temperature response to the total forcing (i.e., all factors) can be approximated as the sum of the response to the individual responses, this provides a way to estimate the warming attributable to greenhouse gases alone in the past decades (Stott et al. 2000). The global energy budget provides another way to estimate the contributions of warming by individual forcings (Huber and Knutti 2012). The total warming over the twentieth century can be reproduced with a variety of model parameters due to compensating effects of stronger aerosol forcing with higher feedbacks and climate sensitivity (Kiehl 2007; Knutti 2008b; Knutti and Hegerl 2008), and does not provide a strong constraint on future warming. The past warming attributable to greenhouse gases, however, is closely related to the warming attributable to greenhouse gases in the future in every model. Therefore, the individual warming contributions from different forcings can serve as a better emerging constraint (Allen et al. 2000; Frame et al. 2006; Knutti et al. 2002; Rogelj et al. 2012; Stott and Kettleborough 2002). The causal relationships are clearer in such cases, but additional steps are needed to separate the signal attributable to specific drivers from other variations.

We also have to recognize that there are very few pieces of actual raw observations that can be used for model evaluation. Most measurements have to be processed in various ways, aggregated over time and space, and calibrated between instruments. In some cases, the instrument, for example, a satellite, measures reflectance of a wavelength, from which we infer a temperature. All of those steps rely on models in one way or the other. Some reanalysis data is even produced by a weather model that assimilates observations, and is thus an interpolated dataset with biases that may be similar to those in climate models. So there is a spectrum between the actual data that is measured, and the model, and they meet somewhere, but the transition is gradual.

Finally, there is the philosopher who will think hard about the model but not touch it, and hope for the best. Of course, none of the above is meant in a depreciatory way. The different methods are not exclusive and none is superior to the others, and it is only by fruitful interaction that we can provide support that the models are telling us something useful

about the real world, and where the limitations are from both a conceptual and a practical point of view.

11.7 Model Calibration

Model calibration or tuning is unavoidable in climate models. Certain parameters in the model have no analogue in reality and must be chosen (with bounds of course) to maximize agreement with data. In many cases, it is not the parameter itself that is constrained, but the effect of the parameterization on the overall simulation. Calibration is common in simple models but has rarely been discussed publicly for complex models. Some argue that tuning undermines the credibility of models, because it could result in the model getting the right effect for the wrong reason, a point raised a few years ago when compensating biases in climate sensitivity and radiative forcing was found across climate models (Kiehl, 2007; Knutti 2008b). From a Bayesian point of view, however, calibration is a natural way to obtain a prediction given some observations available. Observed warming trends or mean climate are used routinely in simple models and scaling methods (Knutti et al. 2002; Meinshausen et al. 2009; Rogelj et al. 2012: Rowlands et al. 2012; Stott and Kettleborough 2002), or methods that weight models a posteriori (Smith et al. 2009; Tebaldi et al. 2004, 2005). The important points are that we need to make sure the constraints are sufficiently strong to be informative, that the degrees of freedom in the model are small compared to the amount of data to avoid overfitting, and that we do not use the agreement of model and data where calibration has occurred as evidence for model quality or adequacy for predictive purposes (it is of course evidence for the model to be adequate for fitting the data, but that is not our primary purpose). Model agreement in this case tells us little about the model having the correct structure, only that the model is consistent with data. But one should not misinterpret this as the models having no predictive power. If we can make a case that there are no other equally consistent models to explain the data, then the models are powerful in explaining the past, and likely in predicting the future through some emerging constraints. This is the idea underlying the detection and attribution work,

and its relation to predicting future warming (Allen et al. 2000; Stott et al. 2013; Stott and Kettleborough 2002).

There are some methodological questions that have not been explored much so far. For example, it seems justified to calibrate a sea ice parameter to get a good representation of a sea ice model when forced with observed ocean and atmosphere boundary conditions. But once the sea ice model is coupled to the ocean and atmosphere model, its performance will be worse because the ocean and atmosphere models are providing biased boundary conditions. Is it acceptable then to change the sea ice parameter to correct for those? Many would argue it is, because we are still calibrating the same quantity, just in a different environment. But is it acceptable to change the sea ice parameter to improve the overall global climate model (including ENSO maybe) even if it makes the sea ice agreement worse? Some would argue this is not justified, because it produces the right effect for the wrong reason, which undermines the credibility of the model as a reliable representation of the target, but others would argue that it is justified because it improves the model overall. Despite some philosophical work (Lenhard and Winsberg 2010), this question of balance between evaluating the performance of the whole model versus small parts has not received much attention in the climate community.

In practice, the computational cost and complexity of GCMs prevent extensive and systematic parameter calibration in a coupled model, except to some degree in distributed setups (Rowlands et al. 2012). Model evaluation on the other hand is comparably cheap and does feed back into model development. A senior colleague put it this way: "I am obviously not advocating trying to tune and tweak to reproduce exactly what happened in the past. I am sure we would not be able to do that anyway. Models are stubborn about what they want to do. I am suggesting that we should not ignore important changes that have happened in the past but are not simulated in the models." It would be strange to argue that models are not tuned at all, because we value models more if they seem to be "right" even without extensive or explicit tuning; to an extent, we may have tuned them unconsciously during the development process, since some of the past data are known to us during the process. It is important for the modeling community to discuss model calibration and document

it (Mauritsen et al. 2012), as this is part of the process of understanding why the models behave the way they do. It is also important to clarify many misconceptions about model tuning by those unfamiliar with model development.

11.8 Challenges

Models are evaluated extensively against data (Flato et al. 2013), but even such assessments are largely silent about what those instances of fit or misfit to data imply. The main difficulty is that model evaluation must be specific to the purpose. The purpose determines which processes matter, on what spatial and temporal scales, and it is therefore difficult to make general statements about implications regarding skill or adequacy. How to weight different pieces of agreement or mismatch is subject to debate, and it is this translation of model performance metrics into model quality for a purpose where we struggle (Knutti 2008a; Knutti et al. 2010a; Parker 2009a). Ultimately, we do not want to measure fit to observations (P1* matching P1) but evaluate relationships, the internal covariance structure of the models. We need to make sure the models do the right thing for the right reason, because we want to use them beyond the range they have been evaluated. We have greatest confidence in models where we understand the processes behind the results, and where we can argue that models represent them well enough.

The same difficulty about defining model quality implies that it is not obvious to decide which models in an ensemble are better and should be given more weight, if any (Knutti et al. 2010a, b; Tebaldi and Knutti 2007). Overfitting to observations has been shown to actually decrease skill, in particular when there are very few models (Weigel et al. 2010). Strong emerging constraints are quite rare, and do not always hold across structurally different models (Knutti et al. 2006, 2010b; Masson and Knutti 2013; Sanderson 2013). One possible explanation of this is that we have essentially used more of the available data to evaluate and constrain models already, and thus the data provides no further constraint either because its uncertainty is too large, or because structural model issues are the limiting factor (Sanderson and Knutti 2012). In Bayesian

words, the ensemble is already conditional on the observations, and the data therefore cannot constrain it further.

On the level of an ensemble of models, there are at least three open questions. The first relates to the interpretation of the ensemble in a statistical sense. The fact that the average of simulation results from several models often agrees better with observations than any single model (Gleckler et al. 2008; Knutti et al. 2010b; Reichler and Kim 2008; Sanderson and Knutti 2012) may suggest that models are randomly distributed around the truth such that the errors cancel when averaging. This "truth plus error" interpretation conveniently implies that projections get ever more certain with more models (Knutti et al. 2010b; Lopez et al. 2006; Tebaldi et al. 2004), but at least in the limit of a very large number of models it is not defensible. The "truth plus error" interpretation is equivalent to say that we care about the uncertainty in the model mean response. The alternative interpretation is that reality is "indistinguishable" from the set of models; that is, every model realization is an equally plausible future (Annan and Hargreaves 2011), in which case the uncertainty does not depend on the number of models. Reality may be more complicated than picking one of the two, because an ensemble may change its characteristics over time. The compensation of errors is indeed more pronounced than expected by chance in CMIP5 (i.e., there is an element of "truth plus error"), yet the "indistinguishable" interpretation is clearly preferred for the future (Sanderson and Knutti 2012).

The second issue is the number of independent models is quite small. Of course, all models are dependent in the sense that they describe the same system and use the same basic equations. However, some models also use the same parameterization, or make similar simplifications, that is, are similarly "wrong". In some cases, they even share code, or in the extreme case a model can be submitted several times to an intercomparison with just minor changes (e.g., resolution, or fixed vs. interactive chemistry). Such model similarity is clear not only from knowing the code, but from analyzing the simulated climate (Knutti et al. 2013; Masson and Knutti 2011a; Pennell and Reichler 2011). The sharing of code and ideas is not a problem, but if it is not considered then the ensemble will be biased toward duplicate models. It may also artificially

increase significance in correlations in emerging constraints (Caldwell et al. 2014).

The third issue is that model intercomparisons are based on a set of models that is sampled neither systematically nor randomly. It is often called an ensemble of opportunity, and cannot a priori be interpreted as the kind of sample from which a statistician would usefully estimate uncertainty, in much the same way as the response of 30 people riding in the same train car cannot be interpreted as the public's opinion on any topic. The ensemble may be too wide (i.e., not reflecting our actual belief about the uncertainty) if some models are performing badly, or too narrow if all models are missing certain processes. An ensemble that is too wide can be narrowed more easily by observational constraints, but an ensemble that is overconfident would require extrapolation beyond the range of models, which is harder (e.g., if all models are ignoring the effect of methane hydrates, it is very difficult to say anything about it). It turns out that at least for some quantities, the range across models is probably not too far off from an assessed uncertainty (Collins et al. 2013), but given the issues listed above, it gets increasingly difficult to justify the view of model democracy for future climate projections (Knutti 2010).

Finally, on decadal timescales the natural variability associated with ocean atmosphere interaction and atmospheric circulation is large, in particular for variables of the water cycle (Deser et al. 2012a, b; Fischer et al. 2013; Knutti and Sedláček 2013; Mahlstein et al. 2011, 2012). These largely unpredictable deviations from an underlying anthropogenic change pose a major challenge for model evaluation, for estimating model robustness, and for near term projections (Knutti and Sedláček 2013; Schaller et al. 2011; Sedláček and Knutti 2014; Tebaldi et al. 2011).

11.9 Conclusions

Numerical models are often seen as inferior to "real data" from experiments, or even discarded as useless by construction. The quote "Garbage in, garbage out", which refers to the fact that computers can quickly produce large amounts of erroneous or irrelevant data, captures that well. However, that view misses the fact that all predictions and most inferences

about data are based on some form of a model. The model may be as trivial as a linear relationship as an underlying theory or assumption, but it is still a model, that is, a representation of some features of the real world as its intended target system. To put any numbers into context, to turn numbers into insight, we need to assume some conceptual or numerical model. In my view, the fact that the more complex models need a computer is only a practical nuisance.

Computer models do not have a life on their own. It is not the model that is wrong, or the computer that does something crazy; it is our assumptions and theories that may not be adequate for the purpose we use them, or the way we instruct the computer to solve the equations. Models as such are not good or bad, right or wrong, or uncertain; they are just better or worse in representing certain aspects of reality, and thus more or less useful and relevant to inform us about reality for a certain purpose. I argue that we do not confirm the model itself in a narrow sense, nor its prediction, because the model is what it is, and the prediction is an inevitable consequence from it. We try to confirm that the relationship between the model and the target system is sufficiently straightforward and understood that the model serves as a representation of the target, such that the insight from the model can be used to infer something about the target system, e.g., that the model is adequate to within a specified uncertainty range to make a statement about the target for a specific purpose. Uncertainty is inevitable because all models are idealizations, and the uncertainty range is determined by a form of uncertainty propagation of the known structural limitations of the models, parameter uncertainties, and data uncertainties in the prediction and inherent variability in the natural and social systems. In practice, such uncertainty analysis is hard. The propagation of uncertainty may be nonlinear and complex. If only the input were uncertain, it may be possible to apply uncertainty propagation through the model, but the model structure, that is, the propagation itself, is also not fully known.

So looking ahead into a world with ever-increasing amounts of data and computational power, how should we allocate our resources? To quantify uncertainty, should we have massive numbers of simulations with perturbed parameters to explore all possible outcomes of future climate? Should we have more groups developing models independently, or

should we rather combine our efforts into developing fewer but "better" models? Should we increase resolution, or include more processes? Should we even abandon the idea of developing only one model in an institution and build different ones, some for scientific insight and others for prediction, where the latter would strictly focus on predictive skill for a particular purpose? And is there even agreement about the purpose before we build those models? What is the value of trying to have more detailed spatial information versus reducing the uncertainties in large-scale changes? How do we value robustness if agreement may be partly due to shared biases? Model spread in the latest ensemble has decreased where observations are available but not for future projections, so we converge on observations where we have them but not elsewhere (Knutti and Sedláček 2013). The same effect can be seen on the range of equilibrium climate sensitivity, which is almost unchanged in at least three decades (Knutti and Hegerl 2008). Do we have more confidence in newer models even if there are persistent biases? And even if the uncertainties in projections remain the same? One might argue that we do, since we trust the inner workings and relationships more if the models describe more of the relevant processes in a more physical way. In Donald Rumsfeld's words, one may argue, we have converted some of the "unknown unknowns" to "known unknowns," so we may be more certain that the uncertainties in the models truly reflect our understanding, data, and ability to synthesize that in models (Knutti and Sedláček 2013).

I leave many of the above questions unanswered. They have challenged us for many years and will continue to do so. In most cases they boil to down to understanding what matters to define skill, that is, how to link performance metrics of model agreement with data on one hand to quality metrics on the other hand, which tell us whether the model is adequate to tell us something about the real world. So inevitably, the answers will be specific to the questions, and will require us to make arguments that in one way or the other are based on process understanding and how that understanding is reflected in the model. In some cases, the model may be fine to inform decisions. In others, we may realize that the model is not informative, and we should be able to say so. Or we may realize that the information is not required if the question is framed differently (Dessai and Hulme 2004; Lempert 2013). In any case, I believe that the

critical debates about the actual value of models for various purposes and their uncertainties in informing us about the real world do not happen often enough, although admittedly they have started in various places (Curry and Webster 2011; Dessai and Hulme 2004; Held 2014; Knutti 2008a, 2010; Knutti et al. 2010a, b; Knutti and Sedláček 2013; Lloyd 2010; Mearns 2010; Parker 2006, 2009a, b; Smith 2002; Weisberg 2006).

Progress in simulating climate in a single model has been slower than many have hoped, but at least it can clearly be demonstrated for present-day mean climate and for individual processes (Flato et al. 2013; Knutti et al. 2013). How to interpret and use the results of a model for future changes, and how to use ensembles of models for decision-making, is far less obvious. It requires more than the next numerical algorithm or dataset; it requires overcoming the gap between the climate modeling community and other fields, including social sciences, and in the end decision-making about climate adaptation and mitigation. There is no argument against looking at models, and indeed models already provide a wealth of large-scale information that we trust. But there may be an argument not to use the information for making decisions for some more complex phenomena when we conclude that the models are not capturing the relevant processes adequately. Uncertainties in predicting future climate are likely to persist for some time, and waiting for the perfect prediction will not be the best strategy. In this respect, climate change is not different from many other aspects in daily life where decisions need to be made based on incomplete knowledge and uncertainties need to be managed.

Acknowledgments I thank Christoph Baumberger, Gertrude Hirsch Hadorn, Lisa Lloyd, Maria Rugenstein, Wendy Parker, and Eric Winsberg for constructive comments and discussions, which have helped to clarify my thinking and have greatly improved this manuscript.

References

Allen, Myles R., Peter A. Stott, John F.B. Mitchell, Reiner Schnur, and Thomas L. Delworth. 2000. Quantifying the Uncertainty in Forecasts of Anthropogenic Climate Change. *Nature* 407 (6804): 617–620.

Annan, J.D., and J.C. Hargreaves. 2011. Understanding the CMIP3 Multimodel Ensemble. *Journal of Climate* 24 (16): 4529–4538.

Boé, Julien, Alex Hall, and Xin Qu. 2009. September Sea-Ice Cover in the Arctic Ocean Projected to Vanish by 2100. *Nature Geoscience* 2 (5): 341–343.

Boer, G.J., M. Stowasser, and K. Hamilton. 2007. Inferring Climate Sensitivity from Volcanic Events. *Climate Dynamics* 28 (5): 481–502.

Bony, Sandrine, Robert Colman, Vladimir M. Kattsov, Richard P. Allan, Christopher S. Bretherton, Jean-Louis Dufresne, Alex Hall, et al. 2006. How Well Do We Understand and Evaluate Climate Change Feedback Processes? *Journal of Climate* 19 (15): 3445–3482.

Braconnot, Pascale, Sandy P. Harrison, Masa Kageyama, Patrick J. Bartlein, Valerie Masson-Delmotte, Ayako Abe-Ouchi, Bette Otto-Bliesner, and Yan Zhao. 2012. Evaluation of Climate Models Using Palaeoclimatic Data. *Nature Climate Change* 2 (6): 417–424.

Caldwell, Peter M., Christopher S. Bretherton, Mark D. Zelinka, Stephen A. Klein, Benjamin D. Santer, and Benjamin M. Sanderson. 2014. Statistical Significance of Climate Sensitivity Predictors Obtained by Data Mining. *Geophysical Research Letters* 41 (5): 1803–1808.

Claussen, Martin, L. Mysak, A. Weaver, Michel Crucifix, Thierry Fichefet, M.-F. Loutre, S. Weber, et al. 2002. Earth System Models of Intermediate Complexity: Closing the Gap in the Spectrum of Climate System Models. *Climate Dynamics* 18 (7): 579–586.

Collins, Matthew, Reto Knutti, Julie Arblaster, J.-L. Dufresne, Thierry Fichefet, Pierre Friedlingstein, Xuejie Gao, et al. 2013. Long-Term Climate Change: Projections, Commitments and Irreversibility. In *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, 1029–1136. Cambridge/New York: Cambridge University Press.

Cox, Peter M., David Pearson, Ben B. Booth, Pierre Friedlingstein, Chris Huntingford, Chris D. Jones, and Catherine M. Luke. 2013. Sensitivity of Tropical Carbon to Climate Change Constrained by Carbon Dioxide Variability. *Nature* 494 (7437): 341–344.

Curry, Judith A., and Peter J. Webster. 2011. Climate Science and the Uncertainty Monster. *Bulletin of the American Meteorological Society* 92 (12): 1667–1682.

DelSole, Timothy, and Jagadish Shukla. 2009. Artificial Skill Due to Predictor Screening. *Journal of Climate* 22 (2): 331–345.

Deser, Clara, Reto Knutti, Susan Solomon, and Adam S. Phillips. 2012a. Communication of the Role of Natural Variability in Future North American Climate. *Nature Climate Change* 2 (11): 775–779.

Deser, Clara, Adam Phillips, Vincent Bourdette, and Haiyan Teng. 2012b. Uncertainty in Climate Change Projections: The Role of Internal Variability. *Climate Dynamics* 38 (3–4): 527–546.

Dessai, Suraje, and Mike Hulme. 2004. Does Climate Adaptation Policy Need Probabilities? *Climate Policy* 4 (2): 107–128.

Fasullo, John T., and Kevin E. Trenberth. 2012. A Less Cloudy Future: The Role of Subtropical Subsidence in Climate Sensitivity. *Science* 338 (6108): 792–794.

Fischer, E.M., and R. Knutti. 2013. Robust Projections of Combined Humidity and Temperature Extremes. *Nature Climate Change* 3 (2): 126–130.

Fischer, Erich M., Urs Beyerle, and Reto Knutti. 2013. Robust Spatially Aggregated Projections of Climate Extremes. *Nature Climate Change* 3 (12): 1033–1038.

Flato, Gregory, Jochem Marotzke, Babatunde Abiodun, Pascale Braconnot, Sin Chan Chou, William J. Collins, Peter Cox, et al. 2013. Evaluation of Climate Models. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. *Climate Change 2013* 5: 741–866.

Frame, D.J., D.A. Stone, P.A. Stott, and M.R. Allen. 2006. Alternatives to Stabilization Scenarios. *Geophysical Research Letters* 33 (14). http://onlinelibrary.wiley.com/doi/10.1029/2006GL025801/full.

Gleckler, P.J., T.M.L. Wigley, B.D. Santer, J.M. Gregory, K. AchutaRao, and K.E. Taylor. 2006. Volcanoes and Climate: Krakatoa's Signature Persists in the Ocean. *Nature* 439 (7077): 675–675.

Gleckler, P.J., K.E. Taylor, and C. Doutriaux. 2008. Performance Metrics for Climate Models. *Journal of Geophysical Research: Atmospheres* 113 (D6): D06104. https://doi.org/10.1029/2007JD008972.

Gleckler, P.J., K.E. Taylor, and C. Doutriaux. 2008. Performance Metrics for Climate Models. *Journal of Geophysical Research* 113: 1–20. https://doi.org/10.1029/2007JD008972.

Hall, Alex, and Xin Qu. 2006. Using the Current Seasonal Cycle to Constrain Snow Albedo Feedback in Future Climate Change. *Geophysical Research Letters* 33 (3). http://onlinelibrary.wiley.com/doi/10.1029/2005GL025127/full.

Hargreaves, J.C., and J.D. Annan. 2009. On the Importance of Paleoclimate Modelling for Improving Predictions of Future Climate Change. *Climate of the Past* 5 (4): 803–814.

Hawkins, Ed, and Rowan Sutton. 2009. The Potential to Narrow Uncertainty in Regional Climate Predictions. *Bulletin of the American Meteorological Society* 90 (8): 1095–1107.

———. 2011. The Potential to Narrow Uncertainty in Projections of Regional Precipitation Change. *Climate Dynamics* 37 (1–2): 407–418.

Held, Isaac M. 2005. The Gap Between Simulation and Understanding in Climate Modeling. *Bulletin of the American Meteorological Society* 86 (11): 1609–1614.

Held, Isaac. 2014. Simplicity Amid Complexity. *Science* 343 (6176): 1206–1207. https://doi.org/10.1126/science.1248447.

Huber, Markus, and Reto Knutti. 2012. Anthropogenic and Natural Warming Inferred from Changes in Earth's Energy Balance. *Nature Geoscience* 5 (1): 31–36.

Huber, Markus, Irina Mahlstein, Martin Wild, John Fasullo, and Reto Knutti. 2011. Constraints on Climate Sensitivity from Radiation Patterns in Climate Models. *Journal of Climate* 24 (4): 1034–1052.

Jun, Mikyoung, Reto Knutti, and Douglas W. Nychka. 2008. Local Eigenvalue Analysis of CMIP3 Climate Model Errors. *Tellus A* 60 (5): 992–1000.

Kiehl, Jeffrey T. 2007. Twentieth Century Climate Model Response and Climate Sensitivity. *Geophysical Research Letters* 34 (22). http://onlinelibrary.wiley.com/doi/10.1029/2007GL031383/full.

Knutti, Reto. 2008a. Should We Believe Model Predictions of Future Climate Change? *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 366 (1885): 4647–4664.

———. 2008b. Why Are Climate Models Reproducing the Observed Global Surface Warming So Well? *Geophysical Research Letters* 35 (18). http://onlinelibrary.wiley.com/doi/10.1029/2008GL034932/full.

———. 2010. The End of Model Democracy? *Climatic Change* 102 (3–4): 395–404.

Knutti, Reto, and Gabriele C. Hegerl. 2008. The Equilibrium Sensitivity of the Earth's Temperature to Radiation Changes. *Nature Geoscience* 1 (11): 735–743.

Knutti, Reto, and Jan Sedláček. 2013. Robustness and Uncertainties in the New CMIP5 Climate Model Projections. *Nature Climate Change* 3 (4): 369–373.

Knutti, Reto, Thomas F. Stocker, and Daniel G. Wright. 2000. The Effects of Subgrid-Scale Parameterizations in a Zonally Averaged Ocean Model. *Journal of Physical Oceanography* 30 (11): 2738–2752.

Knutti, Reto, Thomas F. Stocker, Fortunat Joos, and Gian-Kasper Plattner. 2002. Constraints on Radiative Forcing and Future Climate Change from Observations and Climate Model Ensembles. *Nature* 416 (6882): 719–723.

Knutti, Reto, Gerald A. Meehl, Myles R. Allen, and David A. Stainforth. 2006. Constraining Climate Sensitivity from the Seasonal Cycle in Surface Temperature. *Journal of Climate* 19 (17): 4224–4233.

Knutti, Reto, Gabriel Abramowitz, Matthew Collins, Veronika Eyring, Peter J. Gleckler, Bruce Hewitson, and Linda Mearns. 2010a. Good Practice Guidance Paper on Assessing and Combining Multi Model Climate Projections. In *Meeting Report of the Intergovernmental Panel on Climate Change Expert Meeting on Assessing and Combining Multi Model Climate Projections*, ed. Thomas Stocker, Qin Dahe, G.K. Plattner, M. Tignor, and P.M. Midgley. Bern: IPCC Working Group I Technical Support Unit, University of Bern, Switzerland.

Knutti, Reto, Reinhard Furrer, Claudia Tebaldi, Jan Cermak, and Gerald A. Meehl. 2010b. Challenges in Combining Projections from Multiple Climate Models. *Journal of Climate* 23 (10): 2739–2758.

Knutti, Reto, David Masson, and Andrew Gettelman. 2013. Climate Model Genealogy: Generation CMIP5 and How We Got There. *Geophysical Research Letters* 40 (6): 1194–1199.

Lahsen, Myanna. 2005. Seductive Simulations? Uncertainty Distribution Around Climate Models. *Social Studies of Science* 35 (6): 895–922.

Lean, Judith L. 2010. Cycles and Trends in Solar Irradiance and Climate. *Wiley Interdisciplinary Reviews: Climate Change* 1 (1): 111–122.

Lean, Judith, Juerg Beer, and Raymond S. Bradley. 1995. Reconstruction of Solar Irradiance Since 1610: Implications for Climate Change. *Geophysical Research Letters* 22 (23): 3195–3198.

Lempert, Robert. 2013. Scenarios That Illuminate Vulnerabilities and Robust Responses. *Climatic Change* 117 (4): 627–646.

Lenhard, Johannes, and Eric Winsberg. 2010. Holism and Entrenchment in Climate Model Validation. *Studies in History and Philosophy of Modern Physics* 41: 253–262.

Lloyd, Elisabeth A. 2009. Varieties of Support and Confirmation of Climate Models. *Aristotelian Society Supplementary Volume* 83: 213–232. https://doi:10.1111/j.1467-8349.2009.00179.x

———. 2010. Confirmation and Robustness of Climate Models. *Philosophy of Science*: 971–984.

Lopez, Ana, Claudia Tebaldi, Mark New, Dave Stainforth, Myles Allen, and Jamie Kettleborough. 2006. Two Approaches to Quantifying Uncertainty in Global Temperature Changes. *Journal of Climate* 19 (19): 4785–4796.

Mahlstein, I., R. Knutti, S. Solomon, and R.W. Portmann. 2011. Early Onset of Significant Local Warming in Low Latitude Countries. *Environmental Research Letters* 6: 34009. https://doi.org/10.1088/1748-9326/6/3/034009.

Mahlstein, Irina, Robert W. Portmann, John S. Daniel, Susan Solomon, and Reto Knutti. 2012. Perceptible Changes in Regional Precipitation in a Future Climate. *Geophysical Research Letters* 39: 1–5. https://doi.org/10.1029/2011GL050738.

Masson, D., and R. Knutti. 2011a. Climate Model Genealogy. *Geophysical Research Letters* 38: L08703. https://doi.org/10.1029/2011GL046864.

Masson, David, and Reto Knutti. 2011b. Spatial-Scale Dependence of Climate Model Performance in the CMIP3 Ensemble. *Journal of Climate* 24: 2680–2692. https://doi.org/10.1175/2011JCLI3513.1.

———. 2013. Predictor Screening, Calibration, and Observational Constraints in Climate Model Ensembles: An Illustration Using Climate Sensitivity. *Journal of Climate* 26: 887–898. https://doi.org/10.1175/JCLI-D-11-00540.1.

Mauritsen, Thorsten, Bjorn Stevens, Erich Roeckner, Traute Crueger, Monika Esch, Marco Giorgetta, Helmuth Haak, et al. 2012. Tuning the Climate of a Global Model. *Journal of Advances in Modeling Earth Systems* 4. https://doi.org/10.1029/2012MS000154.

McFarlane, Norman. 2011. Parameterizations: Representing Key Processes in Climate Models Without Resolving Them. *Wiley Interdisciplinary Reviews: Climate Change* 2: 482–497. https://doi.org/10.1002/wcc.122.

Mearns, Linda O. 2010. The Drama of Uncertainty. *Climatic Change* 100 (1): 77–85.

Meinshausen, Malte, Nicolai Meinshausen, William Hare, Sarah C.B. Raper, Katja Frieler, Reto Knutti, David J. Frame, and Myles R. Allen. 2009. Greenhouse-Gas Emission Targets for Limiting Global Warming to 2 °C. *Nature* 458 (7242): 1158–1162.

Moss, Richard H., Jae A. Edmonds, Kathy A. Hibbard, Martin R. Manning, Steven K. Rose, Detlef P. Van Vuuren, Timothy R. Carter, et al. 2010. The Next Generation of Scenarios for Climate Change Research and Assessment. *Nature* 463 (7282): 747–756.

Mueller, Peter. 2010. Constructing Climate Knowledge with Computer Models. *Wiley Interdisciplinary Reviews: Climate Change* 1 (4): 565–580.

Oreskes, Naomi, Kristin Shrader-Frechette, Kenneth Belitz, and others. 1994. Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences. *Science* 263 (5147): 641–646.

Parker, Wendy S. 2006. Understanding Pluralism in Climate Modeling. *Foundations of Science* 11 (4): 349–368.

———. 2009a. Does Matter Really Matter? Computer Simulations, Experiments, and Materiality. *Synthese* 169 (3): 483–496.

———. 2009b. II—Confirmation and Adequacy-for-Purpose in Climate Modelling. *Aristotelian Society Supplementary Volume* 83: 233–249. https://doi:10.1111/j.1467-8349.2009.00180.x

———. 2011. When Climate Models Agree: The Significance of Robust Model Predictions. *Philosophy of Science* 78 (4): 579–600.

———. 2013. Ensemble Modeling, Uncertainty and Robust Predictions. *Wiley Interdisciplinary Reviews: Climate Change* 4 (3): 213–223.

Pennell, Christopher, and Thomas Reichler. 2011. On the Effective Number of Climate Models. *Journal of Climate* 24 (9): 2358–2367.

Pirtle, Zachary, Ryan Meyer, and Andrew Hamilton. 2010. What Does It Mean When Climate Models Agree? A Case for Assessing Independence Among General Circulation Models. *Environmental Science & Policy* 13 (5): 351–361.

Reichler, Thomas, and Junsu Kim. 2008. How Well do Coupled Models Simulate Today's Climate? *Bulletin of the American Meteorological Society* 89 (3): 303–311.

Rogelj, Joeri, Malte Meinshausen, and Reto Knutti. 2012. Global Warming Under Old and New Scenarios Using IPCC Climate Sensitivity Range Estimates. *Nature Climate Change* 2 (4): 248–253.

Rowlands, Daniel J., David J. Frame, Duncan Ackerley, Tolu Aina, Ben B.B. Booth, Carl Christensen, Matthew Collins, et al. 2012. Broad Range of 2050 Warming from an Observationally Constrained Large Climate Model Ensemble. *Nature Geoscience* 5 (4): 256–260.

Sanderson, Benjamin M. 2013. On the Estimation of Systematical Error in Regression-Based Predictions of Climate Sensitivity. *Climatic Change* 118 (3–4): 757–770.

Sanderson, Benjamin M., and Reto Knutti. 2012. On the Interpretation of Constrained Climate Model Ensembles. *Geophysical Research Letters* 39 (16). http://onlinelibrary.wiley.com/doi/10.1029/2012GL052665/full.

Schaller, N., I. Mahlstein, J. Cermak, and R. Knutti. 2011. Analyzing Precipitation Projections: A Comparison of Different Approaches to Climate Model Evaluation. *Journal of Geophysical Research: Atmospheres* 116 (D10). https://doi.org/10.1029/2010JD014963/full.

Sedláček, Jan, and Reto Knutti. 2014. Half of the World's Population Experience Robust Changes in the Water Cycle for a 2 °C Warmer World. *Environmental Research Letters* 9 (4): 44008.

Sherwood, Steven C., Sandrine Bony, and Jean-Louis Dufresne. 2014. Spread in Model Climate Sensitivity Traced to Atmospheric Convective Mixing. *Nature* 505 (7481): 37–42.

Smith, Leonard A. 2002. What Might We Learn from Climate Forecasts? *Proceedings of the National Academy of Sciences* 99 (suppl 1): 2487–2492.

Smith, Richard L., Claudia Tebaldi, Doug Nychka, and Linda O. Mearns. 2009. Bayesian Modeling of Uncertainty in Ensembles of Climate Models. *Journal of the American Statistical Association* 104 (485): 97–116.

Soden, Brian J., Richard T. Wetherald, Georgiy L. Stenchikov, and Alan Robock. 2002. Global Cooling after the Eruption of Mount Pinatubo: A Test of Climate Feedback by Water Vapor. *Science* 296 (5568): 727–730.

Solomon, Susan, Gian-Kasper Plattner, Reto Knutti, and Pierre Friedlingstein. 2009. Irreversible Climate Change Due to Carbon Dioxide Emissions. *Proceedings of the National Academy of Sciences* 106: 1704–1709. https://doi:10.1073/pnas.0812721106.

Stevens, Bjorn, and Sandrine Bony. 2013. What Are Climate Models Missing? *Science* 340 (6136): 1053–1054.

Stott, Peter A., and James A. Kettleborough. 2002. Origins and Estimates of Uncertainty in Predictions of Twenty-First Century Temperature Rise. *Nature* 416 (6882): 723–726.

Stott, Peter A., S.F.B. Tett, G.S. Jones, M.R. Allen, J.F.B. Mitchell, and G.J. Jenkins. 2000. External Control of 20th Century Temperature by Natural and Anthropogenic Forcings. *Science* 290 (5499): 2133–2137.

Stott, Peter, Peter Good, Gareth Jones, Nathan Gillett, and Ed Hawkins. 2013. The Upper End of Climate Model Temperature Projections Is Inconsistent with Past Warming. *Environmental Research Letters* 8 (1): 14024.

Stroeve, Julienne, Marika M. Holland, Walt Meier, Ted Scambos, and Mark Serreze. 2007. Arctic Sea Ice Decline: Faster than Forecast. *Geophysical Research Letters* 34 (9). http://onlinelibrary.wiley.com/doi/10.1029/2007GL029703/full.

Stroeve, Julienne C., Vladimir Kattsov, Andrew Barrett, Mark Serreze, Tatiana Pavlova, Marika Holland, and Walter N. Meier. 2012. Trends in Arctic Sea Ice Extent from CMIP5, CMIP3 and Observations. *Geophysical Research Letters* 39 (16). http://onlinelibrary.wiley.com/doi/10.1029/2012GL052676/full.

Taylor, Karl E., Ronald J. Stouffer, and Gerald A. Meehl. 2012. An Overview of CMIP5 and the Experiment Design. *Bulletin of the American Meteorological Society* 93 (4): 485–498.

Tebaldi, Claudia, and Reto Knutti. 2007. The Use of the Multi-Model Ensemble in Probabilistic Climate Projections. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 365 (1857): 2053–2075.

Tebaldi, Claudia, Linda O. Mearns, Doug Nychka, and Richard L. Smith. 2004. Regional Probabilities of Precipitation Change: A Bayesian Analysis of Multimodel Simulations. *Geophysical Research Letters* 31 (24). http://onlinelibrary.wiley.com/doi/10.1029/2004GL021276/full.

Tebaldi, Claudia, Richard L. Smith, Doug Nychka, and Linda O. Mearns. 2005. Quantifying Uncertainty in Projections of Regional Climate Change: A Bayesian Approach to the Analysis of Multimodel Ensembles. *Journal of Climate* 18 (10): 1524–1540.

Tebaldi, Claudia, Julie M. Arblaster, and Reto Knutti. 2011. Mapping Model Agreement on Future Climate Projections. *Geophysical Research Letters* 38 (23). http://onlinelibrary.wiley.com/doi/10.1029/2011GL049863/full.

Trenberth, Kevin E., and Aiguo Dai. 2007. Effects of Mount Pinatubo Volcanic Eruption on the Hydrological Cycle as an Analog of Geoengineering. *Geophysical Research Letters* 34 (15). http://onlinelibrary.wiley.com/doi/10.1029/2007GL030524/full.

Van Oldenborgh, Geert Jan, S.Y. Philip, and Matthew Collins. 2005. El Niño in a Changing Climate: A Multi-Model Study. *Ocean Science* 1 (2): 81–95.

Weigel, Andreas P., Reto Knutti, Mark A. Liniger, and Christof Appenzeller. 2010. Risks of Model Weighting in Multimodel Climate Projections. *Journal of Climate* 23 (15): 4175–4191.

Weisberg, Michael. 2006. Robustness Analysis. *Philosophy of Science* 73 (5): 730–742.

12

Uncertainty in Climate Science and Climate Policy

Jonathan Rougier and Michel Crucifix

12.1 Introduction

This chapter, written by a statistician and a climate scientist, describes our view of the gap that exists between current practice in mainstream climate science, and the practical needs of policymakers charged with exploring possible interventions in the context of climate change. By 'mainstream' we mean the type of climate science that dominates in universities and research centres, which we will term 'academic' climate science, in contrast to 'policy' climate science; aspects of this distinction will become clearer in what follows.

In a nutshell, we do not think that academic climate science equips climate scientists to be as helpful as they might be, when involved in

J. Rougier (✉)
School of Mathematics, University of Bristol, Bristol, UK

University Walk, Charlotte, NC, USA

M. Crucifix
Université catholique de Louvain, Louvain-la-Neuve, Belgium

E.A. Lloyd, E. Winsberg (eds.), *Climate Modelling*,
https://doi.org/10.1007/978-3-319-65058-6_12

climate policy assessment. Partly, we attribute this to an over-investment in high-resolution climate simulators, and partly to a culture that is uncomfortable with the inherently subjective nature of climate uncertainty.

In Sect. 12.2, we discuss current practice in academic climate science, in relation to the needs of policymakers. Section 12.3 addresses the apparently common misconception (among climate scientists) that uncertainty is something 'out there' to be quantified, much like the strength of meridional overturning circulation. Section 12.4, the heart of the chapter, addresses the core needs of the policymaker and focuses on three strictures for the climate scientist wanting to help her: answer the question, own the judgement, and be coherent. Section 12.5 concludes with a brief reflection.

We have taken the opportunity in this chapter to be a little more polemical than we might be in an academic paper, and maybe a little more exuberant in our expressions. We have also ignored the technical details of practical climate science, something we are both involved in day to day, choosing instead to look at the larger picture. We believe that our observations are valid more widely than just with regard to climate science; for example, many of them would apply with little modification in many areas of natural hazards and in radiological or ecotoxicological risk assessment (Rougier et al. 2013). But they seem most pertinent in climate science, which outstrips the other areas in terms of funding. For example, the UK's Natural Environment Research Council (NERC), whose vision is to 'advance knowledge and understanding of the Earth and its environments to help secure a sustainable future for the planet and its people', allocates 40% of its science budget to climate science and earth system science (NERC Annual Report and Accounts 2010–11, p. 40).

12.2 Different Modes of Climate Science

For our purposes, the telling feature of climate science is that it gained much of its momentum in the era before climate change became a pressing societal concern. Consequently, when policymakers turned to climate science for advice, they encountered a well-developed academic field

whose focus was more towards explanation than prediction. Explanation, in this context, is verifying that observable regularities in the climate system are emergent properties of the basic physics. Largely this is through the interplay between observation and dynamical climate simulation. As the resolution of climate simulators increases, more observed regularities fall into the 'explained' category. The El Niño Southern Oscillation (ENSO) is getting closer to falling into this category, for example (Guilyardi et al. 2009).

Thus, for investment, the dominant vector in academic climate science has been to improve the spatial and temporal resolution of the solvers in climate simulators. Supporting evidence can be found in meteorology. It is argued that one of the contributory factors to measurable improvements in weather forecasting over the last 30 years is higher-resolution solvers, although the quantification of this is confounded by simultaneous improvements in understanding the physics, in the amount of data available for calibration, and in techniques for data assimilation (Kalnay 2002, ch. 1). Setting these confounders aside, it seems natural to assert that higher-resolution solvers will lead to better climate simulators. And indeed, we would not deny this, but we would also question whether in fact it is resolution that is limiting the fidelity of climate simulators.

The reason that we are suspicious of arguments about climate founded on experiences in meteorology is the presence of biological and chemical processes in the earth system that operate on climate policy but not weather timescales. We believe that the acknowledgement of biogeochemistry as a full part of the climate system distinguishes the true climate scientist from the converted meteorologist. Our lack of understanding of climate's critical ecosystems mocks the precision with which we can write down and approximate the Navier-Stokes equations. The problem is, though, that putting ecosystems into a climate simulator is a huge challenge, and progress is difficult to quantify. It introduces *more* uncertain parameters, and, by replacing prescribed fields with time-evolving fields, it can actually make the performance of the simulator worse, until tuning is successfully completed (and there is no guarantee of success). Newman (2011) provides a short and readable account of the difficulties of biology, in comparison to physics.

On the other hand, spending money on higher-resolution solvers requires *fewer* parameterisations of sub-grid-scale processes, and so reduces the challenge of tuning. This activity has a well-documented provenance, and a clear motivation within a coherent science plan. And we cannot resist pointing out another immediate benefit: one can show the funder a more realistic-looking ocean simulation ('now at 0.5° resolution!')—although in fact resolutions as high as 0.1° do not fool experienced oceanographers. But while this push to higher resolutions is natural for meteorology, with its forecast horizon measured in days, for climate we fear that it blurs the distinction between what one *can* simulate and what one *ought* to simulate for policy purposes.

So how might the investment be directed differently? For climate policy, it is necessary to enumerate what might happen under different climate interventions: do nothing, monetise carbon, regulation for contraction and convergence, geo-engineering, and so on. And each of these interventions must be evaluated for a range of scenarios that capture future uncertainty about technology, economics, and demographics. For each pair of intervention and scenario, there is a range of possible outcomes, which represent our uncertainty about future climate. Uncertainty here is 'total uncertainty': only the intervention and the scenario are specified—the policymaker does not have the luxury of being able to pick and choose which uncertainties are incorporated and which are ignored.

Internal variability, part of the natural variability of the climate system, can be estimated from high-resolution simulators, but it is only a tiny part of total uncertainty. Over centurial scales, it is negligible compared to our combined uncertainty of the behaviour of the ice-sheets, and the marine and terrestrial biosphere. This uncertainty can be assessed with the assistance of climate simulators, if it is possible to run them repeatedly under different configurations of the simulator parameters and modules, where these configurations attempt to span the range of not-implausible climate system behaviours. To construct a tableau such as the one in Fig. 12.1 will require a minimum of $4 \times 3 \times 100 \times 90$ model-years of simulation, say, 120,000 model-years, including spin-up. The 100 is the number of different simulator configurations that might be tried, and the 90 is the number of years until 2100. Of course, 100 is

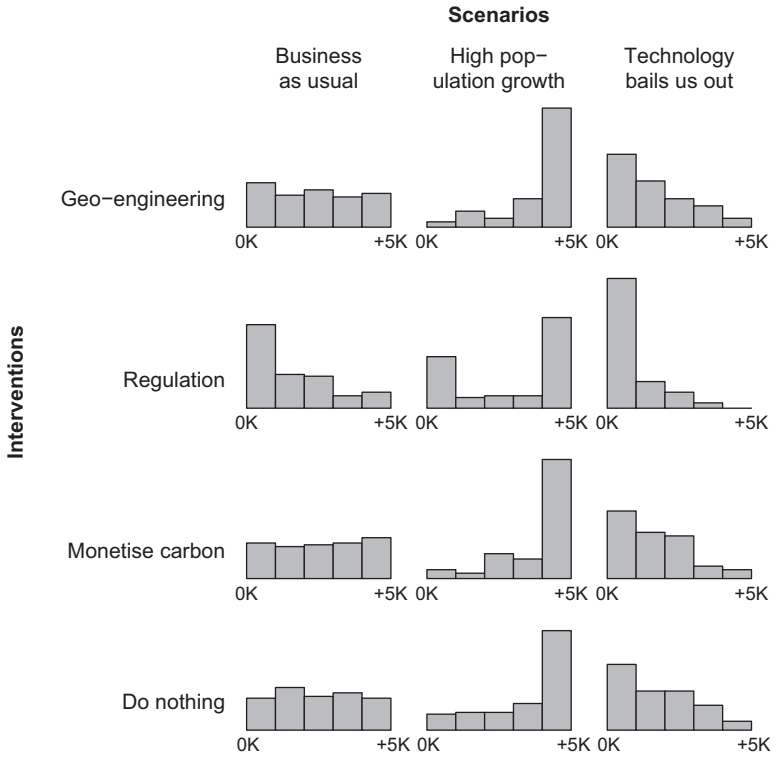

Fig. 12.1 Policy tableau, showing the effect of different possible interventions under different scenarios. These frequency histograms might in this case measure simulated global warming by 2100 under different not-implausible simulator configurations, but more generally they would measure losses, inferred from simulated distributions for weather in 2100. Please note that these histograms are *completely fictitious!*

woefully small for the number of configurations. There are more than 100 uncertain parameters in a high-resolution climate simulator (Murphy et al. 2004). Admittedly only some of these will turn out to be important but we cannot rule out interactions among the parameters. There is a well-developed statistical field for this type of analysis; see, for example, Santner et al. (2003).

Note that this is a designed experiment, deliberately constructed to be informative about uncertainty. It is completely different from assembling

an ad hoc collection of simulator runs, such as the CMIP3 or CMIP5 multimodel ensembles, in the same way that a carefully stratified sample of 100 people is far more informative about a population than simply selecting the next 100 people that pass a particular lamp post. In the absence of designed experiments, though, climate scientists who want to assess uncertainty will have to use the ad hoc ensemble. The various types and uses of currently available ensembles of climate simulator runs are reviewed in Parker (2010) and Murphy et al. (2011).

So what is the status of these policy-relevant designed experiments? Current 'IPCC class' simulators (with a solver resolution of about 1°) run at about 100 model-years per month of wall-clock time. So starting now, an experiment to assess uncertainty in 2100 for policy purposes will be finished in about 100 years, if it is performed at one research centre. But this might be reduced to 10 years if the runs were shared out across all centres, or even less factoring in faster computers and no increase in resolution. Thus, these IPCC class simulators could be very helpful for assessing uncertainty and supporting policymakers, but this requires a cap on solver resolution, and careful coordination across research centres. In contrast, the current uncoordinated approach, with its apparent commitment to spending CPU cycles on a few runs of high-resolution climate simulators, will force climate scientists in 2020 to base their future climate assessments on ad hoc ensembles.

12.3 The Nature of Uncertainty About Climate

In this chapter, we confine our discussion of climate uncertainty quantification to the assessment of probabilities. There are, of course, several interpretations of probability. L.J. Savage wrote of 'dozens' of different interpretations of probability (Savage 1954, p. 2), and he focused on three main strands: the Objective (or Frequentist), the Personalistic, and the Necessary. This tripartite classification is widely accepted among statisticians, and discussed, with embellishments, in the initial chapters of Walley (1991) and Lad (1996). Not to be outdone, Hájek (2012)

notes that philosophers of probability now have six leading interpretations of probability.

Of all of these interpretations, however, we contend that only the Personalistic interpretation can capture the 'total uncertainty' inherent in the assessment of climate policy. Our uncertainty about future climate is predominantly *epistemic* uncertainty—the uncertainty that follows from limitations in knowledge and resources. The hallmark of epistemic uncertainty is that it could, in principle, be reduced with further introspection, or further experiments. As one of the key drivers of research investment in climate science is to reduce uncertainty, this epistemic interpretation of 'total uncertainty' must be uncontentious. It rules out the Objective (classical, frequency, propensity) interpretation, and leaves us with Personalistic and Necessary (also termed logical) interpretations.

The Necessary interpretation asserts that there are principles of reasoning that extend Boolean logic to uncertainty, and that these principles are in fact the calculus of probability and Bayesian conditioning. This interpretation is formally attractive, but invokes additional principles to 'fill in' those initial probabilities that are mandated by conditioning—which are generally referred to as 'prior' probabilities in a Bayesian context. These are to be based on self-evident properties of the inference, such as symmetries. Examples are discussed in Jaynes (2003); see, for example, his elegant resolution of Bertrand's problem (sec. 12.4.4). However, it is hard to know how one might discover and apply these properties in an assessment of, say, the maximum height of the water in the Thames Estuary in 2100. Thus, starting with Frank Ramsey, and finding eloquent champions in Bruno de Finetti and L.J. Savage, among others, the Personalistic interpretation has provided an operational subjective definition of probability, in terms of betting rates (see, e.g., Ramsey 1931; de Finetti 1964; Savage 1954; Savage et al. 1962). De Finetti's late writings are both subtle and discursive; Lad (1996) attempts to corral them.

Not everyone will find the Personalistic definition of probability compelling. But at least it provides a very clear answer to the question, 'What do You mean when You state that $\Pr(A) = p$?' A brief answer is that, if betting for a small amount of money, such as £1, You would be agreeable to staking up to £p in a gamble to receive £1 if A turns out to be true and nothing if A turns out to be false. There are other operationalisations as

well, which are very similar but not psychologically equivalent; see, for example, the discussion in Goldstein and Wooff (2007, sec. 2.2). Our view is that an operationalisation of Personalistic probability is highly desirable, and a useful thing to fall back on, but not in itself the yardstick by which all probabilities are assessed. But, if someone provides a probability p for a proposition A, it might be a good idea to ask him if he would be prepared to bet £p on A being true: the answer could be very revealing.

However, many physical scientists seem to be very uncomfortable with the twin notions that uncertainty is subjective (i.e., it is a property of the mind), and that probabilities are expressions of personal inclinations to act in certain ways. At least part of the problem concerns the use of the word 'subjective', about which the first author has written before (Rougier 2007, sec. 2). This word is clearly inflammatory. We suggest that some scientists have confused the Mertonian scientific norm of 'disinterestedness' with the notion of 'objectivity', and then taken subjectivity to be the antithesis of objectivity, and thus to be avoided at all costs. L.J. Savage was sensitive to this confusion and hence favoured 'Personalistic'. De Finetti strongly favoured 'subjective', about which Jeffrey (2004, p. 76, footnote 1) commented on 'the lifelong pleasure that de Finetti found in being seen to give the finger to the establishment'.

Confusion about 'subjectivity' is just a digression, though. What is abundantly clear is that climate scientists are not ready to accept that climate uncertainties are Personalistic. Their every reference to '*the* uncertainty' commits an error which the physicist E.T. Jaynes called the 'mind projection fallacy':

> an almost universal tendency to disguise epistemological statements by putting them into a grammatical form which suggests to the unwary an ontological statement. To interpret the first kind of statement in the ontological sense is to assert that one's own private thoughts and sensations are realities existing externally in Nature. (Jaynes 2003, p. 22)

Jaynes is an example of a physicist who embraced the essential subjectivity of uncertainty: he advocated the Necessary interpretation, plus the additional principle of maximising Shannon entropy to extend limited

judgements to probabilities. Paris (1994) provides a detailed assessment of the properties of this entropy-maximising approach, among others.

One very stealthy manifestation of the Mind Projection Fallacy is the substitution of 'assumptions' for 'judgements' when discussing uncertainty. Assumptions typically refer to simplifications we assert about the system itself. It is perfectly acceptable to *assume* that, for example, the hydrostatic approximation holds: this is a statement that actual ocean behaves a lot like a slightly different ocean that is much simpler to analyse. You cannot *assume*, though, that the maximum water level in the Thames Estuary in 2100 has a Gaussian distribution. Instead, You may judge it appropriate to represent Your uncertainty about the maximum water level with a Gaussian distribution. This is rather wordy, unfortunately, which is perhaps why it is so easy to lapse in this way.

Consider the uncertainty assessment guidelines for the forthcoming IPCC report (Mastrandrea et al. 2010). Nowhere in the guidelines was it thought necessary to define 'probability'. Either the authors of the guidelines were not aware that this concept was amenable to several different interpretations, or that they were aware of this, and decided against bringing it out into the open. One can imagine, for example, that an opening statement of the form 'In the context of climate prediction, probability is an expression of subjective uncertainty and it can be quantified with reference to a subject's betting behaviour' would have caused great consternation—so much the better!

We can hardly suppose that the omission of a definition for the key concept in such an important and high-profile document was made in ignorance. And yet the mind projection fallacy is in evidence throughout. It looks as though the authors have deliberately chosen *not* to acknowledge the essential subjectivity of climate uncertainty, and to suppress linguistic usage that would indicate otherwise. This should be termed 'monster denial' in the taxonomy of Curry and Webster (2011). Choosing not to rock the boat is convenient for academic climate scientists. But it makes life difficult for policymakers, who are tasked with turning uncertainties into actions. For policymakers, the meaning of '$\Pr(A) = p$' is of paramount importance, and they need to know if ten different climate scientists mean it ten different ways.

12.4 The Risk Manager's Point of View

In any discussion of uncertainty and policy, it is helpful to label the key players (Smith 2010, ch. 1). Conventionally, the person who selects the intervention is the *risk manager*, who represents a particular set of stakeholders. These stakeholders, who are funding the risk manager, and will also fund the intervention that she selects, will appoint an *auditor*, whom the risk manager must satisfy. This framework, of a risk manager who must satisfy an auditor, is a simple way to abstract from the complexities of any particular decision. It emphasises that the risk manager is an agent who must defend her selection, and this has important consequences for the way in which she acts.

The risk manager is surely uncertain about future climate, and its implications. For concreteness, suppose that her concern is about the maximum height of water in the Thames Estuary in the year 2100. If asked, she might say, 'Really, I've no idea, perhaps not lower than today's value, and not more than two metres higher'. But she is not obliged to make such an assessment in isolation: she can consult an *expert*. Put simply, her expert is someone whose judgements she accepts as her own (see Lad 1996, sec. 6.3 for a discussion). So one task of the risk manager is to select her expert, and she must do this in such a way that the auditor is satisfied with the selection process, and with the elicitation process. When seen from the other side, it follows that scientists who want to be involved in climate policy are competing with each other to be selected as one of the risk manager's experts. Therefore, they must demonstrate their grasp of the risk manager's needs. Likewise for climate scientists who are competing for policy-tagged funding.

We highlight the following three risk managers' needs, as posing particular challenges for academic climate scientists.

12.5 Answer the Question

As already discussed, the risk manager needs an assessment of 'total uncertainty'. It can be difficult for the climate scientist to assess his total uncertainty about future climate because of academic climate science's

focus on consuming CPU cycles in higher-resolution solvers, rather than designed replications across alternative not-implausible configurations of simulator parameters and modules. This leaves the willing-to-engage climate scientist ill-equipped to answer questions about ranges for future climate values, because he has nothing other than intuition to guide him on the consequence of the limitations in our knowledge. Unfortunately, his intuition may be tentative at best when reasoning about a dynamical system as complex as the climate system, on centurial timescales.

In this case, the climate scientist may end up specifying very wide intervals which, although honest, do not advance the risk manager because they swamp any 'treatment effect' that might arise from different choices of intervention. This honest climate scientist may well be passed over in favour of other experts who advertise their smaller uncertainty as a putative measure of their superior expertise. This type of competition is extensively discussed in Tetlock (2005), in the context of political and economic forecasting, and the parallels with climate forecasting seem very strong.

How to make the uncertainties smaller? One way is to qualify them with conditions. If these conditions are specified in the question, then of course this is fine. If the risk manager, for example, wants to know about the height of the water in the Thames Estuary under the 'Technology bails us out' scenario, then in it goes. But everything else is suspect. Sometimes the qualification is overt; for example, one hears 'assuming that the simulator is correct' quite frequently in verbal presentations, or perceives the presenter sliding into this mindset. This is so obviously a fallacy that he might as well have said 'assuming that the currency of the US is the jam doughnut'. The risk manager would be justified in treating such an assessment as meaningless. After all, if the climate scientist is not himself prepared to assess the limitations of the simulator, then what hope is there for the risk manager?

As Tetlock (2005) documents, though, often the qualifications are implicit, and only ever appear at the point where the judgement has been shown to be wrong, for example, 'Well, of course I was assuming that the simulator was correct'. The risk manager is not going to be able to winkle out all of these implicit conditions at the start of the process, but other climate scientists might be able to. Thus the elicitation process must be

very carefully structured to ensure that, by the time that the experts finally deliver their probabilities, as many as possible of the implicit qualifications have been exposed and undone. This usually involves a carefully facilitated group elicitation, typically extending over several days. Interestingly, Tetlock did not use group elicitations in his study, but they are standard in environmental science areas such as natural hazards; see, for example, Cooke and Goossens (2000), Aspinall (2010), or Aspinall and Cooke (2013).

Scientists working in climate, and philosophers too we expect, often receive requests to complete online surveys about future climate. These surveys are desperately flawed by responses missing 'not at random'. But even were they not, their results ought to be treated with great circumspection, given the experience in natural hazards of how much difference a careful group elicitation can make, in comparing experts' probabilities at the start and at the finish of the process.

12.6 Own the Judgement

This is in fact another type of qualification, where the climate scientist does not present his own judgement, but someone else's. A classic example would be 'according to the recent IPCC report'. As far as the climate scientist is concerned, these qualified uncertainty assessments are consequence-free, and they ought to be judged by the risk manager as worthless, since nothing is staked.

The IPCC reports are valuable sources of information, but no one owns the judgements in them. Only a very naive risk manager would take the IPCC assessment reports as their expert, rather than consulting a climate scientist, who had read the reports, and also knew about the culture of climate science, and about the IPCC process. This is not to denigrate the IPCC, but simply to be appropriately realistic about its sociological and political complexities, in the face of the very practical needs of the risk manager. These complexities are well-recognised, and a decision by the risk manager to adopt the IPCC reports as her expert can hardly be blame-free. As a marketing ploy, the decision to buy IBM

computers was said to be blame-free in the 1970s and 80s: 'nobody ever got fired for buying IBM equipment'—how hollow that sounds now!

The challenge with owning the judgement in climate science is the complexity of the science itself. There are three main avenues for developing quantitative insights about future climate: (i) computer simulation, (ii) contemporary data collected mainly from field stations, ocean sondes, and satellites, but also slightly older data from ships' logbooks, and (iii) palaeoclimate reconstruction from archives such as ice and sediment cores, speleothems, boreholes, and tree rings. Each of these is a massive exercise in its own right, involving large teams of people, large amounts of equipment, and substantial numerical processing. Judgements about future climate at high spatial and temporal resolution come mainly from computer simulation, but one must not forget that these simulators have been tuned and critiqued against contemporary data and, increasingly, palaeoclimate reconstructions.

Wherever there is a high degree of scientific complexity, there is a large opportunity for human error. With computer simulation, an often-overlooked opportunity for error is the wrapping of the computational core for a specific task; for example, performing a time-slice experiment for the Mid-Holocene at a particular combination of simulator parameter values. Whereas the computational core of the simulator is used time and again, and one might hope that large errors will have been picked up and corrected and committed back to the repository, the wrapper is often used only once. It tends to be poorly documented, often existing as a loose collection of scripts which are passed around from one scientist to another. It is easy to load the wrong initialisation file or boundary file, and also easy to extract the wrong summary values from the gigabytes of simulator output. 'Easy' in this case equates to 'if you have done an experiment like this, you will be aware of at least one mistake that you made, spotted, and corrected'. The correction of this type of mistake can take weeks of effort, as it is tracked backwards from the alarming simulator output to its source in the underlying code.

At the other end of the modelling spectrum, there are phenomenological models of low-dimensional properties of climate and its impacts. See, for example, Crucifix (2012), who surveys dynamical models of glacial

cycles, or Lorenz et al. (2012), who study the welfare value of reducing uncertainty, notably in the presence of a climate tipping point. There are several advantages to such models. First, they are small enough to be coded by the scientist himself, and can be carefully checked for code errors. Thus the scientist can himself be fairly sure that the interesting result from his simulator is not an artefact of a mistake in the programming. Second, they are often tractable enough to permit a formal analysis of their properties. For example, they might be qualitatively classified by type, or explicitly optimised, or might include intentional agents who perform sequences of optimisations (such as risk managers). Third, they are quick enough to execute that they can be run for millions of model-years. Hence, the scientist can use replications to assimilate measurements (including tuning the parameters) and to assess uncertainty, both within a statistical framework (e.g., using the sequential approach of Andrieu et al. 2010).

Of course, 'big modellers' will be scornful of the limited physics (biology, chemistry, economics, etc.) that these phenomenological models contain, although they must be somewhat chastened by the inability of their simulators to conclusively outperform simple statistical procedures in tasks such as ENSO prediction (Barnston et al. 2012). But the real issue is one of ownership. A single climate scientist cannot own an artefact as complex as a large-scale climate simulator, and it is very hard for him to make a quantitative assessment of the uncertainty that is engendered by its limitations. We advocate spending resources on designed experiments to support the climate scientist in this assessment, but we also note that a scientist *can* own a phenomenological model, and the judgements that follow from its use.

12.7 Be Coherent

Tetlock (2005, p. 7) has a similar requirement. In this context, 'coherent' has a technical meaning, which is to say, 'don't make egregious mistakes in probabilistic reasoning'. This needs to be said, because it is more honoured in the breach than the observance.

For example, Gigerenzer (2003) provides a vivid account of how doctors, who ought to be good at uncertainty assessment, often struggle with even elementary probability calculations, and how this compromises the notion of informed consent to medical procedures. As another example, the '*P* -value fallacy'—inferring that the null hypothesis is false because the *P* -value is small—is endemic in applied statistics (see, e.g., Goodman 1999; Ioannidis 2005; Goodman and Greenand, 2007; Ioannidis, 2007). It is very similar to the Prosecutor's fallacy in Law (see, e.g., Gigerenzer 2003, ch. 9). These fallacies serve to remind us that people are not very good when reasoning about uncertainty, and that they can easily be misled by fallacious arguments (that violate the probability calculus), sometimes intentionally.

Tetlock (2005, ch. 4) also notes another aspect of coherence, which is to appropriately update opinions in the light of new information. He emphasises the use of Bayes's Theorem, and demonstrates that his experts did not make the full adjustment that was indicated by Bayesian conditioning. While there are psychological explanations for under-adjustment, we would also note that the probability calculus and Bayesian conditioning is only a *model* for reasoning about uncertainty, and not the sine qua non.

Probabilistic inference owes its power to the unreasonable demands of its axioms, notably the need to quantify an additive (probability) measure on a sufficiently rich field of propositions. This point was very clearly expressed by Savage (1954, notably sec. 2.5), in his contrast between the small world in which one assesses probabilities and performs calculations, and the grand world in which one makes choices. He writes, 'I am unable to formulate criteria for selecting these small worlds and indeed believe that their selection may be a matter of judgment and experience about which it is impossible to enunciate complete and sharply defined general principles . . . On the other hand it is an operation in which we all necessarily have much experience, and one in which there is in practice considerable agreement' (pp. 16–17).

A similar point is made by Howson and Urbach (2006, ch. 3), who defend precise probabilities as a *model* for reasoning against more complex variants in terms of 'the explanatory and informational dividends obtained from their use within simplifying models of *uncertain* inference'

(p. 62, original emphasis). Howson and Urbach present an instructive analogy with deductive logic, whose poor representation of implication requires that we use it thoughtfully when reasoning about propositions that are either true or false (p. 72). Thus, in reasoning about uncertainty, grand world probabilities will be informed by small world calculations such as Bayesian conditioning, but need not be synonymous with them. The Temporal Sure Preference condition of Goldstein (1997) provides one way to connect these two worlds (see also Goldstein and Wooff 2007, sec. 3.5).

So, for climate scientists, and the risk managers they are hoping to impress, the moral of *be coherent* is that (i) it is very easy to make mistakes when reasoning about uncertainty, (ii) strict adherence to the rules of the probability calculus (and perhaps the assistance of a professional statistician) will minimise these, and (iii) although probability calculations are highly informative, no one should be overly impressed by an uncertainty assessment that is a precise implementation of fully probabilistic Bayesian conditioning—one would expect this to be simplistic.

12.8 Reflection

Suppose that you were one of a group of climate scientists, interested in playing an active role in climate policy, and able to meet the three strictures outlined in the previous three sections in Sect. 12.4. You have all embraced subjective uncertainty, and have been summoned, willingly, to a carefully facilitated expert elicitation session. After two intense but interesting days, your 95% equi-tailed credible interval for the maximum height of water in the Thames Estuary in 2100 is 0.5–2.75 m higher than today. This is wider than your initial interval, as you came to realise, during the elicitation process, that there were uncertainties which you had not taken into account.

Suppose that this has recently happened, and you are reflecting on the process, and wondering what information might have made a large difference to your uncertainty assessment, and that of your fellow experts. In particular, you imagine being summoned back in the year 2020, to reassess your uncertainties in the light of eight years of climate science

progress. Would you be saying to yourself, 'Yes, what I really need is an *ad hoc* ensemble of about 30 high-resolution simulator runs, slightly higher than today's resolution'? Let's hope so, because right now, that is what you are going to get.

But we think you would be saying, 'What I need is a designed ensemble, constructed to explore the range of possible climate outcomes, through systematically varying those features of the climate simulator that are currently ill-constrained, such as the simulator parameters, and by trying out alternative modules with qualitatively different characteristics'. Obviously, you would prefer higher resolution to the current resolution, but you don't see squeezing another 0.25° out of the solver as worth sacrificing all the potential for exploring uncertainty inherent in our limited knowledge of the earth system's dynamics, and its critical ecosystems. We would like to see at least one of the large climate modelling centres commit to providing this information by 2020, on their current simulator, operating at a resolution that permits hundreds of simulator runs per scenario (a resolution of about 2°, we hazard). Research funders have the power to make this happen, but for some reason they have not yet perceived the need.

References

Andrieu, C., A. Doucet, and R. Holenstein. 2010. Particle Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society, Series B* 72 (3): 269–302. With Discussion, 302–342.

Aspinall, W.P. 2010. A Route to More Tractable Expert Advice. *Nature* 463: 294–295.

Aspinall, W.P., and R.M. Cooke. 2013. Quantifying Scientific Uncertainty from Expert Judgment Elicitation. In Rougier et al. (2013), Chapter 4.

Barnston, A.G., M.K. Tippett, M.L. L'Heureux, S. Li, and D.G. DeWitt. 2012. Skill of Real-Time Seasonal ENSO Model Predictions During 2002–11: Is Our Capability Increasing? *Bulletin of the American Meteorological Society* 93 (5): 631–651.

Cooke, R.M., and L.H.J. Goossens. 2000. Procedures Guide for Structured Expert Judgement in Accident Consequence Modelling. *Radiation Protection Dosimetry* 90 (3): 303–309.

Crucifix, M. 2012. Oscillators and Relaxation Phenomena in Pleistocene Climate Theory. *Philosophical Transactions of the Royal Society, Series A*, Reprint Available at arXiv:1103.3393v1.

Curry, J.A., and P.J. Webster. 2011. Climate Science and the Uncertainty Monster. *Bulletin of the American Meteorological Society* 92 (12): 1667–1682.

de Finetti, B. 1964. Foresight: Its Logical Laws, Its Subjective Sources. In *Studies in Subjective Probability*, ed. H. Kyburg and H. Smokler, 93–158. New York: Wiley. (2nd ed., New York: Krieger, 1980).

Gigerenzer, G. 2003. *Reckoning with Risk: Learning to Live with Uncertainty*. London: Penguin.

Goldstein, M. 1997. Prior Inferences for Posterior Judgements. In *Structures and Norms in Science. Volume Two of the Tenth International Congress of Logic, Methodology and Philosophy of Science, Florence, August 1995*, ed. M.L.D. Chiara, K. Doets, D. Mundici, and J. van Benthem, 55–71. Dordrecht: Kluwer.

Goldstein, M., and D.A. Wooff. 2007. *Bayes Linear Statistics: Theory & Methods*. Chichester: Wiley.

Goodman, S. 1999. Toward Evidence-Based Medical Statistics. 1: The *p*-value Fallacy. *Annals of Internal Medicine* 130: 995–1004.

Goodman, S., and S. Greenland. 2007. Why Most Published Research Findings Are False: Problems in the Analysis. *PLoS Medicine* 4(4): e168. A Longer Version of the Paper Is Available at http://www.bepress.com/jhubiostat/paper135

Guilyardi, E., A. Wittenberg, A. Fedorov, M. Collins, C. Wang, A. Capotondi, G.J. van Oldenborgh, and T. Stockdale. 2009. Understanding El Niño in Ocean—Atmosphere General Circulation Models: Progress and Challenges. *Bulletin of the American Meteorological Society* 90 (3): 325–340.

Hájek, A. 2012. Interpretations of Probability. In ed. E.N. Zalta, *The Stanford Encyclopedia of Philosophy (Summer Edition)*. Forthcoming URL http://plato.stanford.edu/archives/sum2012/entries/probability-interpret/

Howson, C., and P. Urbach. 2006. *Scientific Reasoning: The Bayesian Approach*. 3rd ed. Chicago: Open Court Publishing Co.

Ioannidis, J.P.A. 2005. Why Most Published Research Findings Are False. *PLoS Medicine* 2 (8): e124. See also Goodman and Greenland (2007) and Ioannidis (2007).

———. 2007. Why Most Published Research Findings Are False: Author's Reply to Goodman and Greenland. *PLoS Medicine* 4 (6): e215.

Jaynes, E.T. 2003. *Probability Theory: The Logic of Science*. Cambridge: Cambridge University Press.

Jeffrey, R.C. 2004. *Subjective Probability: The Real Thing*. Cambridge: Cambridge University Press. Unfortunately This First Printing Contains Quite a Large Number of Typos.

Kalnay, E. 2002. *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge: Cambridge University Press.

Lad, F. 1996. *Operational Subjective Statistical Methods*. New York: Wiley.

Lorenz, A., M.G.W. Schmidt, E. Kriegler, and H. Held. 2012. Anticipating Climate Threshold Damages. *Environmental Modeling and Assessment* 17: 163–175.

Mastrandrea, M.D., C.B. Field, T.F. Stocker, O. Edenhofer, K.L. Ebi, D.J. Frame, H. Held, E. Kriegler, P.R. Matschoss K.J. Mach, G.-K. Plattner, G.W. Yohe, and F.W. Zwiers. 2010. Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties. Technical report, Intergovernmental Panel on Climate Change (IPCC).

Murphy, J.M., D.M.H. Sexton, D.N. Barnett, G.S. Jones, M.J. Webb, M. Collins, and D.A. Stainforth. 2004. Quantification of Modelling Uncertainties in a Large Ensemble of Climate Change Simulations. *Nature* 430: 768–772.

Murphy, J., R. Clark, M. Collins, C. Jackson, M. Rodwell, J.C. Rougier, B. Sanderson, D. Sexton, and T. Yokohata. 2011. *Perturbed Parameter Ensembles as a Tool for Sampling Model Uncertainties and Making Climate Projections*. In Proceedings of ECMWF Workshop on Model Uncertainty, 20–24 June 2011, pp. 183–208. Available Online, http://www.ecmwf.int/publications/library/ecpublications/_pdf/workshop/2011/Model_uncertainty/Murphy.pdf

Newman, T.J. 2011. Life and Death in Biophysics. *Physical Biology* 8: 1–6.

Paris, J.B. 1994. *The Uncertain Reasoner's Companion: A Mathematical Perspective*. Cambridge: Cambridge University Press.

Parker, W.S. 2010. Predicting Weather and Climate: Uncertainty, Ensembles and Probability. *Studies in History and Philosophy of Modern Physics* 41: 263–272.

Ramsey, F.P. 1931. Truth and Probability. In *Foundations of Mathematics and Other Essays*, ed. R.B. Braithwaite, 156–198. London: Kegan, Paul, Trench, Trubner, & Co.

Rougier, J.C. 2007. Probabilistic Inference for Future Climate Using an Ensemble of Climate Model Evaluations. *Climatic Change* 81: 247–264.

Rougier, J.C., R.S.J. Sparks, and L.J. Hill, eds. 2013. *Risk and Uncertainty Assessment for Natural Hazards*. Cambridge: Cambridge University Press.

Santner, T.J., B.J. Williams, and W.I. Notz. 2003. *The Design and Analysis of Computer Experiments*. New York: Springer.

Savage, L.J. 1954. *The Foundations of Statistics*. New York: Dover. Revised 1972 Edition.

Savage, L.J., et al. 1962. *The Foundations of Statistical Inference*. London: Methuen.

Smith, J.Q. 2010. *Bayesian Decision Analysis: Principle and Practice*. Cambridge: Cambridge University Press.

Tetlock, P.E. 2005. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton/Oxford: Princeton University Press.

Walley, P. 1991. *Statistical Reasoning with Imprecise Probabilities*. London: Chapman & Hall.

13

Communicating Uncertainty to Policymakers: The Ineliminable Role of Values

Eric Winsberg

13.1 Introduction

Over the last several years, there has been an explosion of interest and attention devoted to the problem of Uncertainty Quantification (UQ) in climate science—that is, to giving quantitative estimates of the degree of uncertainty associated with the predictions of global and regional climate models. The technical challenges associated with this project are formidable: the real data sets against which model runs are evaluated are large, patchy, and involve a healthy mixture of direct and proxy data; the computational models themselves are enormous, and hence the number of model instances that can be run is minuscule and sparsely distributed in the solution space that needs to be explored; the parameter space that we would like to sample is vast and multidimensional;

E. Winsberg (✉)
Department of Philosophy, University of South Florida, Tampa, FL, USA

E.A. Lloyd, E. Winsberg (eds.), *Climate Modelling*,
https://doi.org/10.1007/978-3-319-65058-6_13

and the structural variation that exists amongst the existing set of models is substantial but poorly understood. Understandably, therefore, the statistical community that has engaged itself with this project has devoted itself primarily to overcoming some of these technical challenges.

So why is UQ so important in climate science? What goals are we trying to meet with UQ, and are they likely to be met? Those who are interested in these questions might benefit from a close look at some of the recent philosophical literature on the role of social values in science. UQ, I suggest, is first and foremost a tool for communicating knowledge from experts to policymakers. Experts, in this case, climate scientists and climate modelers, have knowledge about the climate. In one sense, therefore, they are the people who ought to be considered best situated to make decisions about what we ought to do in matters related to climate. But in another sense, they are not.

Consider the fact that we often evaluate the wisdom of pursuing various climate adaptation strategies, such as: how to manage the problem of glacial lake outburst floods, one of the many possible dangers of regional climate changes. These floods occur when a dam (consisting of glacier ice and a terminal moraine) containing a glacial lake fails. Should a local community threatened by a possible flood replace the terminal moraine with a concrete dam? The answer to this question depends in part on the likelihood of the glacier melting and the existing (natural) dam bursting, which climate scientists, who have the most expertise about the future of the local regional climate, would be in the best position to address. It also surely depends, however, on the cost of building the dam, and on the likely damage that would ensue if the dam were to break. Just as much, it might depend on how the relevant stakeholders weigh the present costs against the future damages. And so while, on the one hand, we would like the people making the decision to have the most expertise possible, we also, on the other hand, want the decision to be made by people who represent our interests, whoever "we" might be. Making decisions about, for example, climate adaptation strategies, therefore, requires a mixture of the relevant expertise and the capacity to represent the values of the people on whose behalf one is making the

decision.[1] But there is rarely any single group of people who obviously possess both of these properties.

UQ, as we will see in what follows, is in principle one way in which these different capacities can be kept separate. One clear motivation for solving the problems of UQ, in other words, is to maintain this division of labor between the epistemic and the normative—between the people who have the pure scientific expertise and the people with the legitimate ability to represent the values of the relevant stakeholders. And so if we want to understand where the need to produce quantitative estimates of uncertainty comes from, we need to delve into the role of social values in the administration of scientific expertise.

13.2 Science and Social Values

What do we mean, first of all, by "social values"? Social values, I take it, are the estimations of any agent or group of agents of what is important and valuable—in the typical social and ethical senses—and of what is to be avoided, and to what degree. What value does one assign to economic growth, on the one hand, and to the degree to which we would like to avoid various environmental risks, on the other? In the language of decision theory, by social values we mean the various marginal utilities one assigns to events and outcomes. The point of the word "social" in "social values" is primarily to flag the difference between these values and what Ernan McMullin once called "epistemic values," like simplicity, fruitfulness, and so forth (1983). But I do not want to beg any questions about whether or not values that are paradigmatically ethical or social can or cannot or should or should not play important epistemic roles. So, I prefer not to use that vocabulary. I talk instead about social and ethical values when I am referring to things that are valued for paradigmatically social or ethical reasons. I do not carefully distinguish, in this chapter, between the social and the ethical.[2]

It is uncontroversial that social and ethical values play a role in science. When we set constraints on experimentation, for example, or

when we decide which projects to pursue and which projects to ignore, these decisions uncontroversially reflect social values. But the philosophically controversial question about social and ethical values is about the degree to which they are involved (or better put: the degree to which they are necessarily involved, or inevitably involved, and perhaps most importantly: uncorrectibly involved) in the appraisal of hypotheses or in reaching other conclusions that are internal to science, and that necessarily also involves scientific expertise. This is the question, after all, of the degree to which the epistemic and the normative can be kept apart.

This is a question of some importance because we would like to believe that only experts should have a say in what we ought to believe about the natural world. But we also think that it is *not* experts, or at least not experts qua experts, who should get to say what is important to us, or what is valuable or has utility. Such a division of labor, however, is only possible to the extent that the appraisal of scientific hypotheses, and the consideration of other matters that require scientific expertise, can be carried out in a manner that is free of the influence of social and ethical values.

Philosophers of science of various stripes have mounted a variety of arguments to the effect that the epistemic matter of appraising scientific claims of various kinds cannot be kept free of social and ethical values. Here, we will be concerned only with one such line of argument—one that is closely connected to the issue of UQ—that goes back to the midcentury work of statistician C. West Churchman (1949, 1953) and philosopher of science Richard Rudner (1953).[3] This line of argument is now frequently referred to as the argument from inductive risk. It was first articulated by Rudner in the following schematic form:

1. The scientist qua scientist accepts or rejects hypotheses.
2. No scientific hypothesis is ever completely (with 100% certainty) verified.
3. The decision to either accept or reject a hypothesis depends upon whether the evidence is sufficiently strong.

4. Whether the evidence is *sufficiently* strong is "a function of the *importance*, in a typically ethical sense, of making a mistake in accepting or rejecting the hypothesis."
5. Therefore, the scientist qua scientist makes value judgments.

Rudner's oft-repeated example compared two hypotheses: (1) that a toxic ingredient of a drug is not present in lethal quantity in some resource, (2) that a certain lot of machine stamped belt buckles is not defective. Rudner's conclusion was that "how sure we need to be before we accept a hypothesis will depend upon how serious a mistake it would be" to accept it and have it turn out false (1953, p. 2). We can easily translate Rudner's lesson into an example from climate science: consider a prediction that, given future emissions trends, a certain regional climate outcome will occur. Should we accept the hypothesis, say, that a particular glacial lake dam will burst in the next 50 years? Suppose that if we accept the hypothesis, we will replace the moraine with a concrete dam. But whether we want to build the dam will depend not only on our degree of evidence for the hypothesis, but also on how we would measure the severity of the consequences of building the dam, and having the glacier not melt, vs. not building the dam, and having the glacier melt. Rudner would have us conclude that as long as the evidence is not 100% conclusive, we cannot justifiably accept or reject the hypothesis without making reference to our social and ethical values.

The best-known reply to Rudner's argument came from logician and decision theorist Richard Jeffrey (1956). Jeffrey argued that the first premise of Rudner's argument—that it is the proper role of the scientist qua scientist to accept and reject hypotheses—is false. The proper role of scientists, he urged, is to assign probabilities to hypotheses with respect to the currently available evidence. Others—for example, policymakers—can attach values or utilities to various possible outcomes or states of affairs and, in conjunction with the probabilities provided by scientists, decide how to act.

In providing this response to Rudner, Jeffrey was making it clear that an important purpose of probabilistic forecasts is to separate practice from theory and the normative from the epistemic, so that social values can be relegated entirely to the domain of practice, and cordoned off

from the domain of scientific expertise. If the scientist accepts or rejects a hypothesis, then Rudner has shown that normative considerations cannot be excluded from that decision process. In contrast, if scientists don't have to bring any normative considerations to bear when they assign probabilities to a hypothesis, then the normative considerations can be cordoned off. It should now be clear why I said at the beginning that UQ is first and foremost a tool for communicating knowledge from experts to policymakers. It is a tool for dividing our intellectual labor. If we were entirely comfortable simply letting experts qua experts decide for us how we should act, then we would not have such an acute need for UQ.

It is clear, however, that Jeffrey did not anticipate the difficulties that modern climate science would have with the task that he expected to be straightforward and value free, the assignment of probability with respect to the available evidence. There are many differences between the kinds of examples that Rudner and Jeffrey had in mind and the kinds of situations faced by climate scientists. For one, Rudner and Jeffrey discuss cases in which we need the probability of the truth or falsity of a single hypothesis, but climate scientists generally are faced with having to assign probability distributions over a space of possible outcomes. I believe, however, that the most significant difference between the classic kind of inductive reasoning Jeffrey had in mind (in which the probabilities scientists are meant to offer are their subjective degrees of belief based on the available evidence) and the contemporary situation in climate science is the extent to which epistemic agency in climate science is distributed across a wide range of scientists and tools.

Here, I am pursuing a theme that is at the heart of much of my work on computationally intensive science (2010): that this new kind of science requires of philosophers new ways of thinking about old epistemological issues. These kinds of claims can also be found in the work of Elisabeth Lloyd (in this volume and elsewhere, 2012, 2015), where she argues that recent developments in science require that we adopt what she calls "complex empiricism."

I will return to the issue of how climate science differs from the kind of science envisioned by Jeffrey later (especially in Sect. 13.6), but for now, we should turn to what I would claim are typical efforts in climate science to deliver probabilistic forecasts and see how they fare with respect

to Jeffrey's goal of using probabilities to divide labor between the epistemic and the normative.

13.3 Uncertainty in Climate Science

Where do probabilistic forecasts in climate science come from? We should begin with a discussion of the sources of uncertainty in climate models. There are two main sources that concern us here: *structural model uncertainty* and *parameter uncertainty*. While the construction of climate models is guided by basic science—science in which we have a great deal of confidence—these models also incorporate a barrage of auxiliary assumptions, approximations, and parameterizations, all of which contribute to a degree of uncertainty about the predictions of these models. This source of uncertainty is often called "structural model uncertainty."

Next, complex models involve large sets of parameters or aspects of the model that have to be quantified before the model can be used to run a simulation of a climate system. We are often highly uncertain about what the best value for many of these parameters is, and hence, even if we had at our disposal a model with ideal (or perfect) structure, we would still be uncertain about the behavior of the real system we are modeling, because the same model structure will make different predictions for different values of the parameters. Uncertainty from this source is called "parameter uncertainty."[4]

Most efforts in contemporary climate science to measure these two sources of uncertainty focus on what one might call "sampling methods." In practice, in large part because of the high computational cost of each model run, these methods are extremely technically sophisticated, but in principle they are rather straightforward.

I can best illustrate the idea of sampling methods with an example regarding parameter uncertainty: consider a simulation model with one parameter and several variables.[5] If one had a data set against which to benchmark the model, one could assign a weighted score to each value of the parameter based on how well it retrodicted values of the variables in the available data set. Based on this score, one could then assign a probability to each value of the parameter. Crudely speaking, what we are

doing in an example like this is observing the frequency with which each value of the parameter is successful in replicating known data—how many of the variables does it get right? with how much accuracy? over what portion of the time history of the data set?—and then weighting the probability of the parameter taking this value in our distribution in proportion to how well it had fared in those tests.

The case of structural model uncertainty is similar. The most common method of estimating the degree of structural uncertainties in the predictions of climate models is a set of sampling methods called "ensemble methods," which examine the degree of variation in the predictions of the existing set of climate models. By looking at the average prediction of the set of models and calculating their standard deviation, one can produce a probability distribution for every value that the models calculate.

13.4 Some Worries About the Standard Methods

There are reasons to doubt, however, that these simple methods for estimating structural model uncertainty and parameter uncertainty are conceptually coherent. Signs of this are visible in the results that have been produced. These signs have been particularly well noted by climate scientists Claudia Tebaldi and Reto Knutti (2007). Tebaldi and Knutti have noted, in the first instance, that many studies founded on the same basic principles produce radically different probability distributions. One of their very illustrative charts shows a comparison of four different attempts to quantify the degree of uncertainty associated with the predictions of climate models for a variety of scenarios, regions, and predictive tasks. Tebaldi and Knutti note the wide range of the various estimates.

Beyond the graphical display of the wide variety of possible results one can get from ensemble averages, there are various statistical analyses one can perform on ensemble sample characteristics that cast doubt on their reliability for naïve statistical analysis. These are summarized in Tebaldi and Knutti. I quote their conclusions here:

> Recent coordinated efforts, in which numerous general circulation climate models have been run for a common set of experiments, have produced large data sets of projections of future climate for various scenarios. Those multimodel ensembles sample initial conditions, parameters, and structural uncertainties in the model design, and they have prompted a variety of approaches to quantifying uncertainty in future climate change … This study outlines the motivation for using multimodel ensembles and discusses various challenges in interpreting them. Among these challenges are that the number of models in these ensembles is usually small, their distribution in the model or parameter space is unclear, and that extreme behavior is often not sampled … While the multimodel average appears to still be useful in some situations, these results show that more quantitative methods to evaluate model performance are critical to maximize the value of climate change projections from global models. (2007, p. 2053)

Indeed, I would argue that there are four reasons to suspect that ensemble methods are not a conceptually coherent set of methods:

1. Ensemble methods either assume that all models are equally good, or they assume that the set of available methods can be relatively weighted.
2. Ensemble methods assume that, in some relevant respect, the set of available models represent something like a sample of independent draws from the space of possible model structures.
3. Climate models have shared histories that are very hard to sort out.
4. Climate modelers have a herd mentality about success.

I discuss each of these four reasons in what follows. But, first, consider a simple example that mirrors all four: suppose that you would like to know the length of a barn. You have one tape measure and many carpenters. You decide that the best way to estimate the length of the barn is to send each carpenter out to measure the length and then take the average. There are four problems with this strategy. First, it assumes that each carpenter is equally good at measuring. But what if some of the carpenters have been drinking on the job? Perhaps you could weight the degree to which their measurements play a role in the average in inverse proportion to how much they have had to drink. But what if, in addition to

drinking, some have also been sniffing from the fuel tank? How do you weight these relative influences? Second, you are assuming that each carpenter's measurement is independently scattered around the real value. But why think this? What if there is a systematic error in their measurements? Perhaps there is something wrong with the tape measure that systematically distorts them. Third (and relatedly), what if all the carpenters went to the same carpentry school, and they were all taught the same faulty method for what to do when the barn is longer than the tape measure? And fourth, what if, before recording their value, each carpenter looks at the running average of the previous measurements, and if theirs deviates too much, they tweak it to keep from getting the reputation as a poor measurer?

All of these sorts of problems play a significant role—both individually, but especially jointly—in making ensemble statistical methods in climate science conceptually troubled. I will now discuss the role of each of them in climate science in detail:

1. *Ensemble methods either assume that all models are equally good, or they assume that the set of available methods can be relatively weighted.*

If you are going to use an ensemble of climate models to produce a probability distribution, you ought to have some grounds for believing that all of them ought to be given equal weight in the ensemble. Failing that, you ought to have some principled way to weight them. But no such thing seems to exist. While there is widespread agreement among climate scientists that some models are better than others, quantifying this intuition seems to be particularly difficult. It is not difficult to see why.

As Peter Gleckler et al. (2008) note, no single metric of success is likely to be useful for all applications. Their beautiful illustrations show the success of various models for various prediction tasks. It is fairly clear that while there are some unambiguous flops on the list, there is no unambiguous winner, nor a clear way to rank them.

2. *Ensemble methods assume that, in some relevant respect, the set of available models represent something like a sample of independent draws from the space of possible model structures.*

This is surely the greatest problem with ensemble statistical methods. The average and standard deviation of a set of trials is only meaningful if those trials represent a random sample of independent draws from the relevant space—in this case, the space of possible model structures. Many commentators have noted that this assumption is not met by the set of climate models on the market. In fact, I would argue, it is not clear what this would even mean in this case. What, after all, is the space of possible model structures? And why would we want to sample randomly from this? After all, we want our models to be as physically realistic as possible, not random. Perhaps we are meant to assume, instead, that the existing models are randomly distributed around the ideal model, in some kind of normal distribution, on analogy to measurement theory. But modeling isn't measurement, and so there is very little reason to think this assumption holds.[6]

3. *Climate models have shared histories that are very hard to sort out.*

Large clusters of the climate models on the market have shared histories, which is one reason for doubting that existing models are randomly distributed around an ideal model.[7] Some of them share code. Scientists move from one lab to another and bring ideas with them. Various parts of climate models come from a common toolbox of techniques, and so forth. Worse still, we do not even have a systematic understanding of these interrelations. So, it is not just the fact that most current statistical ensemble methods are naive with respect to these effects; it's also that it is far from obvious that we have the background knowledge we would need to eliminate this naïveté and therefore account for them statistically.

4. *Climate modelers have a herd mentality about success.*

Most climate models are highly tunable with respect to some of their variables, and to the extent that no climate lab wants to be the oddball on the block, there is significant pressure to tune one's model to the crowd. This kind of phenomenon has historical precedent.[8] In 1939, Walter Shewhart published a chart of the history of measurement of the speed of light. The chart shows a steady convergence of measured values that is not

well explained by their actual success. Myles Allen puts the point like this: "If modeling groups, either consciously or by 'natural selection,' are tuning their flagship models to fit the same observations, spread of predictions becomes meaningless: eventually they will all converge to a delta-function" (2008).

13.5 The Inevitability of Values: Douglas *contra* Jeffrey

What should we make of all of these problems from the point of view of the Rudner–Jeffrey debate? This much should be clear: Jeffrey's goal of separating the epistemic from the normative cannot be achieved using UQ based on statistical ensemble methods. But Heather Douglas's (2000) discussion of the debate about science and values should have made this clear from the beginning.[9]

Douglas noted a flaw in Jeffrey's response to Rudner: scientists often have to make methodological choices that do not lie on a continuum. Suppose I am investigating the hypothesis that substance X causes disease D in rats. I give an experimental group of rats a large dose of X and then perform biopsies to determine what percentage has disease D. How do I perform the biopsy? Suppose that there are two staining techniques I could use. One is more sensitive and the other is more specific—one produces more false positives and the other more false negatives. Which one should I choose? Douglas notes that which one I choose will depend on my inductive risk profile. To the extent that I weigh more heavily the consequences of saying that the hypothesis is false if it is in fact true, I will chose the stain with more false positives, and vice versa. But that, of course, depends on my social and ethical values. Social and ethical values therefore play an inevitable role in science.

Now, inevitability is always relative to some fixed set of background conditions, and the set of background conditions Douglas assumes include the use of something like classical statistical methods. If I have some predetermined level of confidence, *alpha*, say .05, then which staining method I use will raise or lower, respectively, the likelihood that the

hypothesis will be accepted. What if, on the other hand, all toxicologists were good Bayesians of the kind that Jeffrey almost surely had in mind? What is the argument that they could not use their expert judgment, having chosen whatever staining method they like, to factor in the specificity and sensitivity of the method when they use the evidence they acquire to update their degrees of belief about the hypothesis? In principle, surely they could. By factoring the specificity and sensitivity of the method into their degrees of belief, they are essentially eliminating or "screening out" the influence of the social or ethical values that otherwise would have been present. And if they could do this, social and ethical values, at least the kind that normally play a role in the balance of inductive risks, would not *have* to play a role in their assessments of the probabilities.[10] Let us call this the Bayesian response to the Douglas challenge (BRDC).

Back to climate science: another way to look at the problem with ensemble statistical methods is that they have no hope of skirting Douglas's challenge and hence no hope of fulfilling their intended role—to divide the epistemic from the normative. To the extent that we use sampling methods and ensemble averages, we are doomed to embed past methodological choices of climate modelers into our UQ. And, for just the reasons that Douglas highlights, along with some others, methodological choices often *need* to reflect judgments of social and ethical values.

There are at least two ways in which methodological choices in the construction of climate models will often ineliminably reflect value judgments in the typically social or ethical sense.

1. Model choices have reflected balances of inductive risk.
2. Models have been optimized, over their history, to particular purposes, and to particular metrics of success.

The first point should be obvious from our discussion of Douglas. When a climate modeler is confronted with a choice between two ways of solving a modeling problem, she may be aware that each choice strikes a different balance of inductive risks with respect to a problem that concerns her at the time. Choosing which way to go, in such a circumstance, will have to reflect a value judgment. This will always be true so long as a

methodological choice between methods A and B are not epistemologically *forced* in the following sense: while option A can be justified on the grounds that it is *less* likely to predict, say, outcome O, than B is when O *will not* in fact occur, option B could also be preferred on the grounds that it is *more* likely to predict O if O *will* in fact occur. So, to return to our old example, if the central question is whether or not some glacial dam will burst, there will often be a modeling choice that will make it less likely to predict that the dam will burst when it fact it won't, and a different modeling choice that will make it less likely to predict that the dam won't burst when in fact it will. In such a situation, neither choice will be "objectively correct," since the correct choice will depend on which of the above two situations is deemed more undesirable.

As to the second point, when a modeler is confronted with a methodological choice, she will have to decide which metric of success to use when evaluating the likely success of the various possibilities. And it is hard to see how choosing a metric of success will not reflect a social or ethical value judgment, or possibly even a response to a political pressure, about which prediction task is more "important" (in a not purely epistemic sense.) Suppose choice A makes a model that looks better at matching existing precipitation data, but choice B better matches temperature data. A modeler will need to decide which prediction task is more important in order to decide which method of evaluation to use and that will influence the methodological choice she makes.

13.6 Three Features of Climate Models

The discussion thus far should make two things clear. First, ensemble sampling approaches to Uncertainty Quantification (UQ) are founded on conceptually shaky ground. Second, and perhaps more importantly, they do not enable UQ to fulfill its primary function, namely, to divide the epistemic from the normative in the way that Jeffrey expected probabilistic forecasts to do. And they fail for just the reasons that Douglas has made perspicuous: because they ossify past methodological choices (which themselves can reflect balances of inductive risk and other social and ethical values) into "objective" probabilistic facts.

This raises, of course, the possibility that climate UQ could respond to these challenges with something akin to the BRDC: by adopting a thoroughly Bayesian approach to quantifying probabilities. Recall the problem faced by Douglas' hypothetical toxicologist. If she is looking for statistically significant evidence that substance X is causing disease D at some predetermined level of "statistical significance," then a particular choice of staining method will either raise or lower the probability of finding that result. But if she has some prior probability for the hypothesis, and updates it in response to the evidence acquired in the biopsies, then the choice of staining method needn't influence those probabilities. Similarly, one might hope, the Bayesian climate scientist might avoid the fundamental problem of any approach founded on "objective" ensemble averaging: that past methodological choices become features of the ensemble and hence exert a pull on the estimated uncertainties.

Indeed, this approach has been endorsed by several commentators.[11] Unfortunately, the role of genuinely subjective Bayesian approaches to climate UQ has been primarily in theoretical discussions of what to do; they have not been widely drawn on to produce actual estimates that one sees published and that are delivered to policymakers. Here, I identify some of the difficulties that might explain why these methods are not used in the field. Genuinely Bayesian approaches to UQ in climate science, in which the probabilities delivered reflect the expert judgment of climate scientists rather than observed frequencies of model outputs, face several difficulties. In particular, the difficulties arise as a consequence of three features of climate models: their massive size and complexity; the extent to which epistemic agency in climate modeling is distributed, in time and space, and across a wide range of individuals; and the degree to which methodological choices in climate models are generatively entrenched. Let me take each of these features in turn.

Size and Complexity

Climate models are enormous and complex. Take one of the state-of-the-art American models, NOAA's GFDL CM2.x. The computational model itself contains over a million lines of code. There are over a thousand

different parameter options. It is said to feature modules that are "constantly changing" and as well as hundreds of initialization files that contain "incomplete documentation" (Dunne 2006, p. 00). It is also said to contain novel component modules written by over 100 different people. Just loading the input data into a simulation run takes over two hours. Using over 100 processors running in parallel, it takes weeks to produce one model run out to the year 2100 and months to reproduce thousands of years of paleoclimate (Dunne 2006). Storing the data from a state of the art global climate model (GCM) every five minutes can produce tens of terabytes per model year.

Another aspect of the models' complexity is their extreme "fuzzy modularity" (Lenhard and Winsberg 2010). In general, a modern state-of-the-art climate model is a model with a theoretical core that is surrounded and supplemented by various submodels that themselves have grown into complex entities. Their overall interaction determines the dynamics—and these interactions are themselves quite complex. The coupling of atmospheric and oceanic circulation models, for example, is recognized as one of the milestones of climate modeling (leading to so-called coupled general circulation models). Both components had an independent modeling history, including an independent calibration of their respective model performance. Putting them together was a difficult task because the two submodels now interfered dynamically with each other.[12]

Today, atmospheric GCMs have lost their central place and given way to a deliberately modular architecture of coupled models that comprise a number of highly interactive submodels, like atmosphere, oceans, or ice cover. In this architecture, the single models act (ideally!) as interchangeable modules.[13] This marks a turn from a reliance on one physical core—the fundamental equations of atmospheric circulation dynamics—to the development of a more networked picture of interacting models from different disciplines (see Küppers and Lenhard 2006).

In sum, climate models are made up of a variety of modules and submodels. There is a module for the general circulation of the atmosphere, a module for cloud formation, for the dynamics of sea and land ice, for effects of vegetation, and many more. Each of them, in turn, includes a mixture of principled science and parameterizations. And it is the interaction of these components that generates the overall observable dynamics

in simulation runs. The results of these modules are not first gathered independently and then only after that synthesized. Rather, data are continuously exchanged between all modules during the runtime of the simulation.[14] The overall dynamics of one global climate model is the complex result of the interaction of the modules—not the interaction of the results of the modules. This is why I modify the word "modularity" with the warning flag "fuzzy" when I talk about the modularity of climate models: due to interactivity and the phenomenon of "balance of approximations," modularity does not break down a complex system into separately manageable pieces.[15]

Distributed Epistemic Agency

Climate models reflect the work of hundreds of researchers working in different physical locations and at different times. They combine incredibly diverse kinds of expertise, including climatology, meteorology, atmospheric dynamics, atmospheric physics, atmospheric chemistry, solar physics, historical climatology, geophysics, geochemistry, geology, soil science, oceanography, glaciology, paleoclimatology, ecology, biogeography, biochemistry, computer science, mathematical and numerical modeling, time series analysis, and so forth.

Epistemic agency in climate science is not only distributed across space (the science behind model modules comes from a variety of labs around the world) and domains of expertise but also across time. No state-of-the-art, coupled atmosphere-ocean GCM (AOGCM) is literally built from the ground up in one short surveyable unit of time. They are assemblages of methods, modules, parameterization schemes, initial data packages, bits of code, coupling schemes, and so forth that have been built, tested, evaluated, and credentialed over years or even decades of work by climate scientists, mathematicians, and computer scientists of all stripes.[16]

No single person, indeed no group of people in any one place, at one time, or from any one field of expertise, is in a position to speak authoritatively about any AOGCM in its entirety.[17]

Methodological Choices are Generatively Entrenched

In our (2010), Johannes Lenhard and I argued that complex climate models acquire an intrinsically historical character and show path-dependency. The choices that modelers and programmers make at one time about how to solve particular problems of implementation have effects on what options will be available for solving problems that arise at a later time. And they will have effects on what strategies will succeed and fail. This feature of climate models, indeed, has lead climate scientists such as Smith (2002) and Palmer (2001) to articulate the worry that differences between models are concealed in code that cannot be closely investigated in practice.

Of course the modelers could—in principle—re-work the entire code. The point is, however, that in even moderately complex cases, this is not a viable option for practical reasons. At best, this would be far too tedious and time-consuming. Conceivably, we would not even know how to proceed. So in the end, each step in the model building process, and how successful it might be, could very well depend on the particular way previous steps were carried out—because the previous steps are unlikely to be completely disentangled and redone.

This is the sense in which modeling choices are generatively entrenched. Modeling choices that are made early in the model construction process have effects on the models at later times in unpredictable ways. And the success of modeling choices at later times depends in unpredictable ways on earlier modeling choices.

13.7 Summary

To summarize then, state-of-the-art global climate models are highly complex, they are the result of massively distributed epistemic labors, and they arise from a long chain of generatively entrenched methodological choices whose effects are epistemically inscrutable. These three features, I would now argue, make the BRDC very difficult to pull off with respect to climate science.

13.8 The Failure of the BRDC in Climate Science

Recall how the BRDC is meant to go: Rudner argues that the scientist who accepts or rejects hypotheses has to make value judgments. Jeffrey replies that she should only assign probabilities to hypotheses on the basis of the available evidence, and, in so doing, avoid making value judgments. Douglas argues that scientists make methodological choices, and that these choices will become embedded in the mix of elements that give rise to estimates of probabilities that come from classical, as opposed to Bayesian, statistics. Since those methodological choices will involve a balance of inductive risks, the scientist cannot avoid value judgments. The BRDC suggests that scientists avoid employing any deterministic algorithm that will transmit methodological choices into probabilities (like employing a classical statistical hypothesis test in the toxicology case, or employing ensemble averages in the climate case), and should instead rely on their expert judgment to assess what the appropriate degree of belief in a hypothesis is given that a particular methodological choice is made and resultant evidence acquired. The probabilities such a scientist would offer should be the scientist's subjective degree of belief, one that has been conditionalized on the available evidence.

Unfortunately, large groups of individuals, distributed across space and time, do not possess subjective degrees of belief. Subjective Bayesian probabilities need to be "owned" by one individual epistemic agent (Parker 2011), or, at the very least, by manageably small epistemic groups.[18] But the three features of global climate models I have pointed to—that they are highly complex, are the result of massively distributed epistemic labors, and arise from a long chain of generatively entrenched methodological choices—make it seem implausible, at least to me, that any individual epistemic agent[19] will ever be in good position to have a useful degree of expert judgment of the kind required to implement the BRDC.[20] The BRDC precisely requires that *one epistemic agent* be capable of making an informed judgment about how every single methodological choice on which a climate model is built ought to influence his or her degree of belief in a hypothesis that he or she is evaluating with the

use of that model. But how can we expect any individual, or well-defined group of experts, to do this successfully when faced with massively complex models, built over large expanses of space and time, and built on methodological choices that have become generatively entrenched, and hence epistemically inscrutable?

The argument thus far, then, can be summarized as follows. Climate science, and the construction of climate models, like almost all of science, is full of unforced methodological choices. And like in the rest of science, these choices often reflect priorities with respect to predictive power, and balances of inductive risk. There is nothing new here. It is plausible to suppose, moreover, that Jeffrey understood this to be a feature of much of science, and still believed, pace Douglas, that the subjective Bayesian had available a defense of value-free science: once the methodological choices are made, the scientists qua scientist can update her degree of belief in any relevant hypothesis in light of the evidence that comes from those methodological choices—and that updating can be free of the canonically social or ethical values that guided those methodological choices in the first place. Or at least, so a modern Jeffrien is entitled to maintain. So why is climate science different? It is different because of the size, complexity, socially cooperative origin, and historical path dependency of climate modes. And it is different because climate experts, in light of the individually limited role that they play in the socially extended activity of building climate knowledge, can only arrive at posterior degrees of belief in ways that are fundamentally mediated by the complex models that they build. And they are incapable of sorting out the ways in which past methodological choices are influencing, through their entrenchment in the very models that mediate their inferences, the ways in which they could possibly arrive at those posterior degrees of belief. Their judgments about climate uncertainties, therefore, whether they come from "objective" ensemble methods, or from their subjective judgments, cannot be free from the social values that guide methodological choices everywhere in the sciences.

13.9 Values in the Nooks and Crannies

At this point in the discussion, it might be natural for a reader to ask for a specific example of a social, political, or ethical value that has influenced a methodological choice in the history of climate modeling. It is easy to give a couple of potted examples. In previous work, I have focused on the extent to which climate models have been optimized, over their history, to particular purposes, and to particular metrics of success.[21] I gave the example that, in the past, modelers had perhaps focused on the metric of successfully reproducing known data about global mean surface temperature, rather than other possible metrics. I speculated that they might have done so because of a social and political climate in which the concern was about "global warming," a phrase that is now being supplanted by the phrase "anthropogenic climate change."

But I now think it was a mistake to focus on particular historical claims about specific motives and choices. I want to focus instead on the fact that climate modeling involves literally thousands of unforced methodological choices.[22] Many crucial processes are poorly understood, many compromises in the name of computational exigency need to be made, and so forth. All one needs to see is that, as in the case of the biopsy stain, no unforced methodological choice can be defended in a value vacuum. If one asks, "Why parameterize this process rather than try to resolve it on the grid?" or "Why use this method for modeling cloud formation?" it will rarely be the case that the answer can be "because that choice is objectively better than the alternative." Rather, most choices will be better in some respects and worse in other respects than their alternatives, and the preference for the one over the other will reflect the judgment that this or that respect is more important. Some choices will invariably increase the probability of finding a certain degree of climate variation, while its alternative will do the opposite—and so the choice that is made can be seen as reflecting a balance of inductive risks.

All we need to argue here is that many of the choices made by climate modelers had to have been unforced in the absence of a relevant set of values—that in retrospect, such choices could only be defended against *some set of predictive preferences* and *some balance of inductive risks.* In other

words, any rational reconstruction of the history of climate science would have to make mention of predictive preferences and inductive risks at pain of making most of these choices seem arbitrary. But what I want to be perfectly clear about here (in a way that I think I have not been in earlier work) is that I do not mean to attribute to the relevant actors these psychological motives, nor any particular specifiable or recoverable set of interests.[23] I am not in the business of making historical, sociological, or psychological claims. I have no idea why individual agents made the choices that they made—and indeed it is part of my argument that these facts are mostly hidden from view. In fact, for many of the same reasons that these methodological choices are immune from the BRDC, they are also relatively opaque to us from a historical, philosophical and sociological point of view. They are buried in the historical past under the complexity, epistemic distributiveness, and generative entrenchment of climate models.

Some readers may find that this makes my claim about the value-ladenness of climate models insufficiently concrete to have any genuine bite. One might ask: "Where are the actual values?" Some readers, in other words, might be craving some details about how agents have been specifically motivated by genuine concrete ethical or political considerations. They might be tempted to think that I have too abstractly identified the role of values here to be helpful. But this is to miss the dialectical structure of my point. The very features that make the BRDC implausible make this demand unsatisfiable. No help of the sort that "finds the hidden values" can be forthcoming on my account. The social, political, and ethical values that find their way into climate models cannot be recovered in bite-sized pieces.

Recall that we began this whole discussion with a desire to separate the epistemic from the normative. But we have now learned that, with respect to science that relies on models that are sufficiently complex, epistemically distributed, and generatively entrenched, it becomes increasingly difficult to tell a story that maintains that kind of distinction. And without being able to provide a history that respects that distinction, there is no way to isolate the values that have been involved in the history of climate science.

One consequence of the blurred distinction between the epistemic and the normative in our case is that the usual remarks philosophers often make about the value-ladenness of science do not apply here. Those who make the claim that science is value laden often follow up with the advice that scientists ought to be more self-conscious in their value choices and that they ought to ensure that their values reflect those of the people they serve. Or they suggest implementing some system for soliciting public opinions or determining public values and making that the basis for these determinations. But on the picture I am painting, neither of these options is really possible. The bits of value-ladenness lie in all the nooks and crannies; they might very well have been opaque to the actors who put them there, and they are certainly opaque to those who stand at the end of the long, distributed, and path-dependent process of model construction. In the case of the biopsy stains I can say "consumer protection is always more important than corporate profits! Even in the absence of epistemologically forcing considerations, the toxicologist should choose the stain on the left!" But in the climate case, the situation is quite different. We can of course ask for a climate science that does not reflect systematic biases, unlike one cynically paid for by the oil industry. But this demand for a science that reflects the "right values" cannot go "all the way down" into all those nooks and crannies. In those relevant respects, it becomes terribly hard to ask for a climate science that reflects "better" values.[24]

13.10 Conclusion

So what could Climate Science—its practitioners, its public consumers, and the policymakers who rely on it—do? One very sensible response to a state of affairs in which there is no principled and value-neutral way to assign a precise probability distribution to climate outcomes is to refrain from giving one—certainly from giving one that is derived in a simplistic way from the distribution of modeling results that come from the set of models on the market. This is what Wendy Parker has urged in response to some of my earlier work. Perhaps, she argues, what we have learned is that a probability density function over all the possible outcomes is too

detailed and precise a depiction. Perhaps what climate scientists ought to deliver to the public, and to policymakers, is something coarser.

> In practice, coarser depictions of uncertainty are what we actually get from expert groups like the Intergovernmental Panel on Climate Change (IPCC). Even for GMST, IPCC uncertainty estimates reached on the basis of expert judgment assign only a portion of the probability mass and, moreover, in some cases assign it to a predictive range that extends significantly beyond that delineated by predictions from today's state-of-the-art models and/or by more formal probabilistic methods.[25]

She gives the following example from the IPCC report (see Fig. 13.1):

In Fig. 13.1, the gray bars give the ranges of values (for each emissions scenario), inside which the IPCC experts deemed that there was at least a 66% chance that the actual value of global mean surface temperature would fall. As we can see, this range is significantly larger than any of the formal methods for calculating probability would give us. This reflects their judgment, as Parker puts it, that "today's state-of-the-art models do not thoroughly or systematically sample existing uncertainty about how to adequately represent the climate system. More specifically, it reflects the judgment that these models are more likely to underestimate rather than overestimate changes in GMST and that, while they may well be off

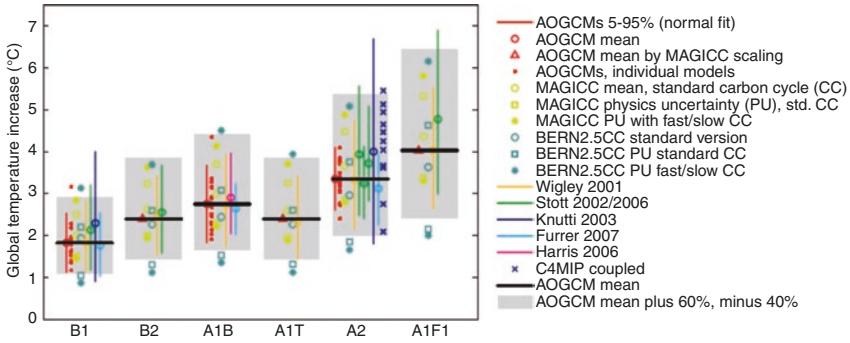

Fig. 13.1 Projections and uncertainties for global mean temperature increase in 2090–2099 (rel. to 1980–1999 avg.) for the six SRES marker scenarios (Source: IPCC AR4 WG1 2007)

by something like 50% in their projections for 2050, there is not a high probability that they are off by something like 500%" (Ibid., Parker).

I agree with Parker (and the IPCC!) that this is an excellent strategy. What I do not agree with is that it is a value-free strategy. Notice what Parker correctly notes is the justification for this report: that there is a (relatively) "high probability" that the models may be off by something like 50%, but a (relatively) "low probability" that they are off by something like 500%. I agree that these are the correct sorts of judgments to be making, but these are classic Rudnerian judgments—they reflect a balance of inductive risks. Deciding to omit some chunk of possibility space from covering a range of values because there is a sufficiently low (second order) probability that it belongs there is exactly the kind of judgment that Rudner was talking about—only elevated to the level of second order probabilities.[26] It can only be made with a combined judgment of the probability that the real value lies in that space, and of the moral, social, or political cost of being wrong. But this is exactly what the IPCC is doing when they leave off those tails on the grounds that they have a "low probability." It is a logical possibility, after all, that one might make the judgment that, even though the probability that the models are off by 500% is extremely small, that the seriousness ("in the typically ethical sense"—Rudner) of neglecting that possibility and having it actually be the case would outweigh that very small probability.

I would like to emphasize that I am not criticizing the IPCC here. I agree with Parker that this is the correct thing to do in light of the present situation with climate models, and in light of the situation that is likely to exist under any practicable state of affairs. But I insist that it is not value free. It is only a slight twist on the classic Rudnerian decision to decide that the (second order) probability that less than 66% of the (first order) probability lies in the gray bar is sufficiently low to be safely ignored. To decide that second order probabilities are sufficiently low to be ignored is to choose a balance of inductive risks. It reflects a judgment that the risk of sticking their necks out further and being wrong is equally balanced by the risk of not sticking it out far enough. As long as there is no principled PDF to be offered, some amount of neck sticking is required. And how far out one should stick one's neck is a classic balance of inductive risks.

If one is uncomfortable with second order probabilities, there are other ways to interpret what the IPCC is doing. But none of them change the conclusion. It is clear that the IPCC cannot be perfectly confident that exactly 66% of the probability mass lies precisely in the grey bars. If they were perfectly confident of this, than they would have a principled precise first order probability—and this is what we have argued, above, they cannot have. But this means that could have made a more precise estimate with less confidence, or a less precise estimate with more confidence. And choosing the right balance of precision and confidence here is a value judgment.

Of course, when values enter into the picture in *this* kind of way—when the experts at the IPCC make a determination about what kinds of minimum probabilities to report—the points I made earlier about the inscrutability of the values no longer apply. At this point in the process, one might even say that the values are being applied fairly self-consciously. And so vis-à-vis this part of the process, I think the ordinary lessons about the role of values in science (that scientists ought to be more self-conscious in their value choices, and that they ought to ensure that their values reflect those of the people they serve, etc.) do apply. And I have no reason to doubt that the IPCC does a reasonably good job of this. But we should not let this conceal the fact that the fundamental science on which IPCC bases its judgments (all the color-coded action inside the gray bars) conceals, in all the ways I described in the last section, an opaque, inscrutable tapestry of values.

Acknowledgements Thanks to Kevin Elliot, Rebecca Kukla, Elisabeth Lloyd, Wendy Parker, Isabelle Peschard, Bas van Fraassen, and Jessica Williams for helpful comments, criticisms, and suggestions as I worked on this manuscript. And thanks to all the participants at conferences and colloquia where I have presented earlier versions of this work, including at the Technical University Eindhoven, San Francisco State University, Georgetown University, the 2010 AGU meeting in San Francisco, and the University of South Florida, and at the 2011 Eastern APA Author Meets Critics session. Too many helpful suggestions, comments, and criticisms have been made to keep track of. Thanks to Justin Biddle and Johannes Lenhard for working with me on previous projects (see the bibliography) that have contributed immeasurably to my understanding of these topics.

Notes

1. Of course one might have worries about whether elected representatives generally represent the values of their constituents but that is the subject of a different discussion.
2. I variously use the expressions "social values," "ethical values," or "social and ethical values" which should not be read as flagging important philosophical differences.
3. See also (Frank 1954; Neurath 1913; Douglas 2000; Howard 2006; Longino 1990, 1996, 2002; Kourany 2003a, b; Solomon 2001; Wilholt 2009; Elliott 2011a, b).
4. Many discussions of UQ in climate science will also identify data uncertainty. In evaluating a particular climate model, including both its structure and parameters, we compare the model's output to real data. Climate modelers, for example, often compare the outputs of their models to records of past climate. These records can come from actual meteorological observations or from proxy data—inferences about past climate drawn from such sources as tree rings and ice core samples. Both of these sources of data, however, are prone to error, and so we are uncertain about the precise nature of the past climate. This, in turn, has consequences for our knowledge of the future climate. While data uncertainty is a significant source of uncertainty in climate modeling, I do not discuss this source of uncertainty here. For the purposes of this discussion, I make the crude assumption that the data against which climate models are evaluated are known with certainty. Notice, in any case, that data uncertainty is part of parameter uncertainty and structural uncertainty, since it acts by affecting our ability to judge the accuracy of our parameters and our model structures.
5. A parameter for a model is an input that is fixed for all time, while a variable takes a value that varies with time. A variable for a model is thus both an input for the model (the value the variable takes at some initial time) and an output (the value the variable takes at all subsequent times). A parameter is simply an input.
6. Some might argue that if we look at how the models perform on past data (for, say, mean global surface temperature), they often are distributed around the observations. But, first, these distributions do not display anything like random characteristics (i.e., normal distribution). And, second, this feature of one variable for past data (the data for which the models have been tuned) is a poor indicator that it might obtain for all variables and for future data.

7. Masson and Knutti (2011) discuss this phenomenon and its effects on multimodel sampling, in detail.
8. Shewhart (1939).
9. Which, inter alia, did much to bring the issue of "inductive risk" back into focus for contemporary philosophy of science and epistemology.
10. Whether they would do so in fact is not what is at issue here. Surely that would depend on features of their psychology and of the institutional structures they inhabit, about which we would have to have a great deal more empirical evidence before we could decide. What is at stake here is whether their social and ethical values would *necessarily* play a role in properly conducted science.
11. See, for example, Goldstein and Rougier (2006).
12. For an account of the controversies around early coupling, see Shackley et al. (1999); for a brief history of modeling advances, see Weart (2010).
13. As, for example, in the earth system modeling framework. See, e.g., Dickenson et al. (2002).
14. Because data are being continuously exchanged one can accurately describe the models as parallel rather than serial in the sense discussed in Winsberg (2006).
15. "Balance of approximations" is a term introduced by Lambert and Boer (2001) to indicate that climate models sometimes succeed precisely because the errors introduced by two different approximations cancel each other out.
16. There has been a move, in recent years, to eliminate "legacy code" from climate models. Even though this may have been achieved in some models (this claim is sometimes made about CM2), it is worth noting that there is a large difference between coding a model from scratch and building it from scratch, that is, devising and sanctioning from scratch all of the elements of a model.
17. See Rougier and Crucifix, this volume.
18. I do not have the space to talk about what "manageably small" might mean here. But see our discussion of "catch and toss" group authorship in the work mentioned in the next note.
19. One might reasonably wonder whether, in principle, a group could be an epistemic agent. In fact, this is the subject of a forthcoming paper by Bryce Huebner, Rebecca Kukla, and me. I would argue here, however, and hope that we will argue in more detail in that paper, that the analytic impenetrability of the models made by the groups involved here is an obstacle to these groups being agents with subjective degrees of belief.

20. One can think of the contribution to this volume by Rougier and Crucifix as a recognition of, and attempt to address, this problem: that complex climate models are too complex to help climate scientists develop subjective degrees of belief.
21. See especially Biddle and Winsberg (2009), and also Winsberg (2010, ch. 6).
22. Here, my point is very well supported by Elisabeth Lloyd's contribution to this volume. Her chapter chronicles in detail a very nice example of the kind of unforced methodological choice I am talking about: the choice of how to calibrate the relevant satellite data. The way Lloyd tells the story, the process involved a whole host of data-processing decisions and choices. I am simply adding to Lloyd's narrative the observation that each of the decisions and choices she chronicles can be understood as being underwritten by balances of inductive risk and prediction preferences.
23. One might complain that if the decisions do not reflect the explicit psychological motives or interests of the scientist, then they do not have a *systematic* effect on the content of science, and are hence no different than the uncontroversial examples of social values I mentioned in the introduction (such as attaching greater value to AIDS research than to algebraic quantum field theory). But though the effect of the values in the climate case might not have a *systematic* effect on the content of science, it is nonetheless an effect *internal* to science in a way that those other examples are not.
24. Again, Elisabeth Lloyd's contribution to this volume illustrates this point.
25. This comes from Parker's remarks at the 2011 meeting of the Eastern division of the American Philosophical Association during an author meets critic session for my (2010).
26. The probability that less than 66% of the probability mass lies inside the gray bar is a second order probability because it talks about the probability of a probability.

Works Cited

Allen, Myles. What Can Be Said About Future Climate? ClimatePrediction.net, June. Available at http://www.climateprediction.net/science/pubs/allen_Harvard2008.ppt. Accessed 3 July 2008.

Biddle, Justin, and Eric Winsberg. 2009. Value Judgments and the Estimation of Uncertainty in Climate Modeling. In *New Waves in the Philosophy of Science*, ed. P.D. Magnus and Jacob Busch. New York: Palgrave Macmillan.

Churchman, C. West. 1949. *Theory of Experimental Inference*. New York: Macmillan.

———. 1953. Science and Decision Making. *Philosophy of Science* 23 (3): 247–249.

Clark, Andy. 1987. The Kluge in the Machine. *Mind and Language* 2 (4): 277–300.

Dickenson, Robert E., Stephen E. Zebiak, Jeffery L. Anderson, et al. 2002. How Can We Advance Our Weather and Climate Models as a Community? *Bulletin of the American Meteorological Society* 83 (3): 431–434.

Douglas, Heather. 2000. Inductive Risk and Values in Science. *Philosophy of Science* 67 (4): 559–579.

Dunne, John. 2006. Towards Earth System Modelling: Bringing GFDL to Life. Paper Presented at the ACCESS 2006 BMRC Workshop. Available at http://www.cawcr.gov.au/bmrc/basic/wksp18/papers/Dunne_ESM.pdf. Accessed 11 Jan 2011.

Elliot, Kevin. 2011a. Direct and Indirect Roles for Values in Science. *Philosophy of Science* 78: 303–324.

———. 2011b. *Is a Little Pollution Good for You? Incorporating Societal Values in Environmental Research*. New York: Oxford University Press.

Frank, Philipp G. 1954. The Variety of Reasons for the Acceptance of Scientific Theories. In *The Validation of Scientific Theories*, ed. Phillipp Frank, 3–17. Boston: Beacon Press.

Gleckler, Peter J., Karl E. Taylor, and Charles Doutriaux. 2008. Performance Metrics for Climate Models. *Journal of Geophysical Research* 113: D06104. https://doi.org/10.1029/ 2007JD008972.

Goldstein, Matthew, and Jonathan C. Rougier. 2006. Bayes Linear Calibrated Prediction for Complex Systems. *Journal of the American Statistical Association* 101 (475): 1132–1143.

Howard, Don A. 2006. Lost Wanderers in the Forest of Knowledge: Some Thoughts on the Discovery-Justification Distinction. In *Revisiting Discovery and Justification: Historical and Philosophical Perspectives on the Context Distinction*, ed. Jutta Schickore and Friedrich Steinle, 3–22. New York: Springer.

IPCC (Intergovernmental Panel on Climate Change). 2001. *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third*

Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge: Cambridge University Press.

———. 2007. *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press.

Jeffrey, Richard C. 1956. Valuation and Acceptance of Scientific Hypotheses. *Philosophy of Science* 23: 237–246.

Kourany, Janet. 2003a. A Philosophy of Science for the Twenty-First Century. *Philosophy of Science* 70 (1): 1–14.

———. 2003b. Reply to Giere. *Philosophy of Science* 70 (1): 22–26.

Küppers, Günter, and Johannes Lenhard. 2006. Simulation and a Revolution in Modeling Style: From Hierarchical to Network-like Integration. In *Simulation: Pragmatic Construction of Reality, Sociology of the Sciences*, ed. Johannes Lenhard, Günter Küppers, and Terry Shinn, 89–106. Dordrecht: Springer.

Lambert, Steven, and George Boer. 2001. CMIP1 Evaluation and Intercomparison of Coupled Climate Models. *Climate Dynamics* 17 (2–3): 83–106.

Lenhard, Johannes, and Eric Winsberg. 2010. Holism, Entrenchment, and the Future of Climate Model Pluralism. *Studies in History and Philosophy of Modern Physics* 41: 253–262.

Lloyd, Elisabeth. 2012. The Role of 'Complex' Empiricism in the Debates About Satellite Data and Climate Models. *Studies in History and Philosophy of Science* 43: 390–401.

———. 2015. *Model Robustness* as a Confirmatory Virtue: The Case of Climate Science. *Studies in History and Philosophy of Science* 49: 58–68. https://doi.org/10.1016/j.shpsa.2014.12.002.

Longino, Helen. 1990. *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton: Princeton University Press.

———. 1996. Cognitive and Non-cognitive Values in Science: Rethinking the Dichotomy. In *Feminism, Science, and the Philosophy of Science*, ed. Lynn Hankinson Nelson and Jack Nelson, 39–58. Dordrecht: Kluwer.

———. 2002. *The Fate of Knowledge*. Princeton: Princeton University Press.

Masson, David, and Reto Knutti. 2011. Climate Model Genealogy. *Geophysical Research Letters* 38 (8): L08703. https://doi.org/10.1029/2011GL046864.

Neurath, Otto. 1913. Die Verirrten des Cartesius und das Auxiliarmotiv: Zur Psychologie des Entschlusses. In *Jahrbuch der Philosophischen Gesellschaft an der Universität Wien*, 45–59. Leipzig: Johann Ambrosius Barth.

Palmer, Tim N. 2001. A Nonlinear Dynamical Perspective on Model Error: A Proposal for Non-local Stochastic–Dynamic Parameterization in Weather and Climate Prediction Models. *Quarterly Journal of the Royal Meteorological Society* 127 (572): 279–304.

Parker, Wendy S. 2011. When Climate Models Agree: The Significance of Robust Model Predictions. *Philosophy of Science* 78 (4): 579–600.

Rudner, Richard. 1953. The Scientist *Qua* Scientist Makes Value Judgments. *Philosophy of Science* 20 (3): 1–6.

Shackley, Simon, James Risbey, Peter Stone, and Brian Wynne. 1999. Adjusting to Policy Expectations in Climate Change Science: An Interdisciplinary Study of Flux Adjustments in Coupled Atmosphere Ocean General Circulation Models. *Climatic Change* 43 (3): 413–454.

Shewhart, Walter A. 1939. *Statistical Method from the Viewpoint of Quality Control*. New York: Dover.

Smith, Leonard A. 2002. What Might We Learn from Climate Forecasts? *Proceedings of the National Academy of Sciences USA* 4 (99): 2487–2492.

Solomon, Miriam. 2001. *Social Empiricism*. Cambridge, MA: MIT Press.

Tebaldi, Claudia, and Reto Knutti. 2007. The Use of the Multimodel Ensemble in Probabilistic Climate Projections. *Philosophical Transactions of the Royal Society A* 365 (1857): 2053–2075.

Weart, Spencer. 2010. The Development of General Circulation Models of Climate. *Studies in History and Philosophy of Modern Physics* 41 (3): 208–217.

Wilholt, Torsten. 2009. Bias and Values in Scientific Research. *Studies in History and Philosophy of Science* 40: 92–101.

Wimsatt, William. 2007. *Re-engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*. Cambridge, MA: Harvard University Press.

Winsberg, Eric. 2006. Handshaking Your Way to the Top: Simulation at the Nanoscale. *Philosophy of Science* 73 (5): 582–594.

———. 2010. *Science in the Age of Computer Simulation*. Chicago: University of Chicago Press.

14

Modeling Climate Policies: The Social Cost of Carbon and Uncertainties in Climate Predictions

Mathias Frisch

14.1 Introduction

That anthropogenic climate change is occurring is a fact. It is scientifically well established that the average global temperature is increasing due to human influence and, in particular, due to the emission of greenhouse gases. But how do we decide on an appropriate policy response to the climate problem? While it is extremely well established through multiple lines of evidence *that* anthropogenic climate change is occurring, significant uncertainties remain concerning the question as to *how precisely* the climate system will be changing, especially on regional scales, and what precisely the effects of climate change on economic, social, and environmental systems will be. Thus, despite the fact that the basic mechanisms responsible for anthropogenic climate change are well

M. Frisch (✉)
Institut für Philosophie, Leibniz Universität Hannover, Hannover, Germany

E.A. Lloyd, E. Winsberg (eds.), *Climate Modelling*,
https://doi.org/10.1007/978-3-319-65058-6_14

understood, climate change presents us with a decision problem under deep uncertainty: what is an appropriate decision procedure under uncertainty in a context in which some of the negative outcomes we have to consider amount to existential threats to social and environmental systems?

In this chapter, I will critically compare two prominent approaches to climate policy: the framework of expected utility theory used in attempts to calculate the social cost of carbon and the precautionary approach underlying the 2 °C and 1.5 °C goals of the Paris Agreement. I will also engage critically with influential arguments that point to existing uncertainties and to problems with attempts to apply expected utility theory to climate change to argue that little or no action on climate change is needed. While I agree with some of the criticisms of the framework used to calculate the social cost of carbon, it does not follow that action on emissions reductions is not needed—quite to the contrary, precautionary considerations support much more urgent climate action than prominent cost-benefit analysis recommend.

I will proceed as follows. The next section provides an overview of the two strategies I will discuss and of the argumentative landscape. Then I will discuss several key uncertainties of the coupled economy-climate models used in calculations of the social cost of carbon. These uncertainties concern the climate sensitivity derived from sophisticated climate models (Sect. 14.3.1), the discount rate used in welfare models (Sect. 14.3.2), and the damage function representing damage to the climate system due to climate change (Sect. 14.3.3). The fact that these uncertainties are "deep" uncertainties that cannot be represented by precise probability distributions calls our ability to apply expected utility theory to the climate problem into question. In Sect. 14.4, I reconstruct and criticize decision strategies that some of the conservative critics of calculations of the social cost of carbon advocate. In their stead I argue for a precautionary approach. Yet significant questions remain. In particular, it is far from clear how to implement a precautionary strategy in a concrete decision principle that can be brought to bear on climate policy discussions.

14.2 A Two-Pronged Approach to Climate Policy

The 21st Conference of the Parties of the United Nations Framework Convention on Climate Change in Paris in 2015 agreed to hold "the increase in the global average temperature to well below 2°C above pre-industrial levels and to pursue efforts to limit the temperature increase to 1.5°C above pre-industrial levels, recognizing that this would significantly reduce the risks and impacts of climate change" (Article 2(a)). The Paris temperature goals are motivated by broadly precautionary thinking. Temperature increases of 2 °C or more would take us outside of the temperature band that we humans have experienced in our 200,000-year history (Jaeger et al. 2010). Moreover, there is evidence that many climate tipping points are located at around 2 °C above pre-industrial levels or slightly above 2 °C (Schellnhuber et al. 2016). Thus, limiting the temperature increase to well below 2 °C promises significantly to reduce the risk of catastrophic climate change.

Yet, it is not easy to make precise a chain of reasoning that can scientifically justify the temperature goals articulated in the Paris agreement. Indeed, the climate scientist Reto Knutti and coauthors write that "no scientific assessment ever defended or recommended a particular target" (Knutti et al. 2016). According to Knutti et al., the temperature target reflects a political consensus that takes into account scientific evidence but cannot be given a purely scientific rationale. Knutti et al. do not present their discussion as criticism of the Paris target, yet others have criticized the target for its purported lack of scientific foundation. The economist William Nordhaus said that "the scientific rationale for the 2°C target is not really very scientific" and the philosopher, economist, and IPCC AR5 lead author John Broome said that the number has "just been pulled out of the air" (in an interview with the *Sydney Morning Herald*, 12. Oct 2013). Thus, significant questions concerning the justification of the Paris Agreement remain. Moreover as Knutti et al. make clear, these questions are not merely academic, since determining the scientific rationale for the 2 °C target affects our understanding of that tar-

get. Does the target, as some seem to believe, constitute a "guard rail" and safe upper limit or, does it rather, as Knutti et al. argue, constitute a partly politically motivated anchor for climate policies, with significant potential risks to the climate system existing even at lower temperatures?

Critics of the precautionary approach underlying the 2 °C target, such as Nordhaus or Broome, point to expected utility theory or cost-benefit calculations as providing an alternative framework that allows for a formally more precise and mathematically well-founded assessment of different climate mitigation policies. While a broadly precautionary approach is playing a large and perhaps dominant role in international climate negotiations, cost-benefit calculations have been (and continue to be) influential as well and were an important component of the Obama administration's approach to climate policy. The core idea of the latter framework is that an optimal climate policy is one that maximizes intergenerational expected utility, taking into account the costs of mitigation policies as well as their future benefit of preventing damages that would otherwise have resulted from future GHG emissions. That is, a cost-benefit framework evaluates climate policies by comparing its costs to present generations against its future benefits in preventing climate change. In the United States, cost-benefit analyses are required by law for all regulatory actions and, since a 2008 decision by the Ninth Circuit Court of Appeals, these analyses are also required to include a consideration of climate benefits.

In order to determine the climate benefits of rule making, the Obama White House put into place an Interagency Working Group (IWG), which was charged with providing quantitative estimates for the social cost of carbon (SCC) (Interagency Working Group on Social Cost of Carbon 2010; 2013). The SCC is the cost associated with emitting an additional ton of carbon dioxide, or its equivalent, into the atmosphere. As the basis for its estimates the IWG used three prominent optimization integrated assessment models (IAMs), which practically implement an expected utility calculation by coupling a climate model with an economy model to calculate the emissions path that maximizes intergenerational utility or welfare.

Thus, the Obama administration pursued a two-pronged strategy to the climate problem. On the one hand, it adopted a cost-benefit approach toward rule making; on the other hand, the United States actively participated in the negotiations of the Paris agreement with its precautionary framework. The two prongs, however, are in tension with each other

not only conceptually but also as far as their policy recommendations are concerned. The expected utility framework adopted by the IWG presupposed that climate science and economics provides us with probabilities for various outcomes, which allow us to calculate the outcomes' expected utilities. The precautionary framework underlying the Paris agreement, by contrast, assumes that such probabilities are not available but that this is no reason to postpone action. This disagreement on foundations is mirrored by a disagreement on policy recommendations. The Paris agreement effectively commits its signatories to reduce carbon emissions to levels close to zero by mid century. The integrated assessment models considered by the IWG, by contrast, allow for much more modest reductions in emissions and, correspondingly, for much larger temperature increases as optimal.

However that may be, the Trump administration is moving quickly to dismantle both strategies. In an executive order signed March 28, 2017, Trump ordered a review of the Clean Power Plan, with the aim to "suspend, revise, or rescind" the Plan (Sec. 4a). He also called for a review of the social cost of carbon and ordered that the IWG be disbanded and all documents issued by the IWG pertaining to the calculations of the social cost of carbon "be withdrawn as no longer representative of governmental policy" (Sec. 5). And on June 1, 2017 Trump announced that the United States would withdraw from the Paris Agreement, claiming that "the Paris Accord is very unfair at the highest level to the United States." (Trump, 2017) With his executive order, Trump is following the advice of Thomas Pyle, head of the Department of Energy transition team for President Trump's administration, who has advocated that the use of the social cost of carbon in federal rulemaking be ended: "If the SCC were subjected to the latest science, it would certainly be much lower than what the Obama administration has been using."[1] Pyle here echoes a criticism made by Robert Murphy, a senior economist at the Institute for Energy Research (IER), who in written testimony before the Senate Committee on Environment and Public Works has argued "that the 'social cost of carbon' is not an objective empirical feature of the world, but is rather a very malleable figure dependent on subjective modeling assumptions, and can be made large, small, or even negative depending on parameter choices" (Murphy 2013). The basic criticisms Murphy and others make of the IWG's use of integrated assessment models to calculate a social cost of carbon are, first, that scientific uncertainties are too large to permit any reliable estimate of the

SCC and, second, that some of the assumptions going into the construction of the models are normative assumptions and, hence, that any model output cannot serve as an objective basis for climate policy decisions.

The use of cost-benefit analysis in climate policy has also been criticized by people who cannot be accused of being climate-change deniers. Thus Jonathan Masur and Eric Posner (2011) argue that evaluating climate damages involves inherently normative and political questions, which cannot be adequately incorporated into a cost-benefit analysis. By contrast, Michael Greenstone and Cass Sunstein, two of the architects of the Interagency Working Group, have recently defended the use of the SCC in regulatory analysis in an OpEd in *The New York Times* on December 15, 2016. Citing the central estimate of the IWG as $35 per ton of carbon dioxide they maintain that "this figure plays a central role in the cost-benefit analyses that agencies use in deciding whether to issue regulations to limit greenhouse gas emissions" and that "without it, such regulations would have no quantifiable benefits. For this reason, the social cost of carbon can be seen as the linchpin of national climate policy" (Greenstone and Sunstein 2016). Others share the view of the importance of the SCC. Richard Revesz et al. have argued that while the IWG's current estimates may underestimate the value of the SCC, these calculations are nevertheless useful for setting climate policy (Reversz et al. 2014). And the climate journalist Andrew Revkin wrote that "there's probably no more consequential and contentious a target for the incoming administration [as far as climate and energy policies are concerned] than an arcane metric called the 'social cost of carbon'" (Revkin 2017).

I want to argue here that Murphy's criticism is to some extent correct: climate-economy models contain deep uncertainties and they partly rely on normative assumptions. First, many of the uncertainties in the values of central parameters of integrated assessment models and climate models more generally are so-called deep or Knightian uncertainties for which no reliable probability distributions are known. Given the available evidence and the large uncertainties surrounding precise climate predictions, any assessment of climate risks, of climate tipping points and of potential damages will need to rely on expert judgment. Such expert assessment will, of course, be informed by predictions derived from climate models and from different lines of evidence more broadly. But the uncertainties in these predictions are too large to allow us simply to read off policy-relevant forecasts from the models in the manner proposed by

the IWG. The IAMs on which the calculations of the SCC are based feign precision where none exists. But without precise probability distributions no expected utility calculation is possible.

Second, some of the modeling assumptions—in particular those concerning the economic effects of climate change—are either overtly or implicitly normative. Hence IAMs cannot provide value-neutral recommendations as to which climate policy would maximize quantifiable or monetizable benefits. Moreover, some of the normatively loaded assumptions made in integrated assessment modeling have the consequence that potential harms to the populations most vulnerable to (and least responsible for) climate change are effectively ignored.

Yet the conclusions that some of the models' conservative critics want to draw do not follow. The models used by the IWG downplay uncertainties by making what are arguably unjustifiably optimistic assumptions about the values of certain key parameters. But instead of trying to correct for this error by broadening the class of assumptions under investigation, conservative critics, such as Murphy and Pyle, reinforce the error further by cherry-picking predictions that lie at the optimistic end of the spectrum found in the peer-reviewed literature.

Thus, a proper accounting of the uncertainties in our knowledge of how climate and economic systems interact and of the moral challenges of climate change puts us exactly in the epistemic and moral situation to which the Paris temperature targets are a response: in a situation in which we face scientifically plausible catastrophic harms, a precautionary approach is warranted, as embodied in the 2 °C or 1.5 °C target. What is more, when many of these potential harms fall in the first instance upon the poorest populations and will do so as a result of our own activities, we have a moral duty to cease these activities and adopt policies that protect the most vulnerable from catastrophic harms.

14.3 Optimization Integrated Assessment Models

According to expected utility theory, we should adopt the climate policy that maximizes expected utility. Calculating the expected utility associated with a climate policy requires as inputs, first, the costs and benefits

of different policy choices, including the economic costs of mitigation measures as well as the future benefits of reductions in temperature increases, and, second, a probability distribution over costs and benefits. In practice, the maximization calculation is performed with the help of so-called "optimization integrated assessment models" (IAMs). These models couple an economic general equilibrium model to an extremely simplified climate model with the aim of representing the impacts of climate change on human welfare, the impact of changes in economic activity on GHG emissions, and the effect of mitigation measures on economic growth. The IWG considered three widely discussed IAMs in its calculation of the social cost of carbon: William Nordhaus's DICE model (Nordhaus 2008), which is one of the earliest optimization IAMs and remains one of the most influential models; the PAGE model, which was used in the Stern Review (Stern 2007); and Richard Tol's FUND model (Tol 2002a, b).

The two core components of optimization IAMs—a climate model and an economy model—are coupled through two different channels: economic activity is assumed to affect climate change through the emission of greenhouse gases (GHGs); and economic activity is modeled as being affected by climate change through a so-called "damage function." Optimization IAMs are used to determine what the optimal emission abatement strategy would be by maximizing the present value of overall utility, which consists in an aggregate of utilities across time. Any such cross-temporal aggregation faces the problem as to what relative weight to assign to utilities at different times. It is common practice to discount future utilities with respect to the present. The choice of discount rates is one of the areas of criticism on which conservative critics focus. The IWG calculated values for the SCC for three different discount rates, 2.5%, 3%, and 5%. The resulting values for the SCC, reported on the EPA website, are $56, $36, and $11.[2] As is evident from these results, the choice of discount rate has a large influence on the value of the SCC. A large discount rate has the effect of minimizing the influence of costs or benefits far in the future on policy decisions today.

In what follows, I want to discuss three areas in which uncertainties arise—the calibration of an IAM's climate model, the discount rate chosen in the economy model, and the choice of damage function modeling

climate impacts. A full accounting would have to include many more uncertainties, but my discussion will serve to illustrate the general problem faced by treating climate change within the framework of expected utility theory.[3]

Climate Sensitivity

The climate model of an IAM consists of a small number of equations with parameters, whose values need to be calibrated with the help of more complex climate models. One central parameter is the so-called "equilibrium climate sensitivity," or *ECS*, which is defined as the equilibrium mean surface temperature response to a doubling in atmospheric CO_2. The IPCC report (AR5) provides probability density functions for the value of *ECS*, which are derived from complex climate models or paleo-climate data (Intergovernmental Panel on Climate Change 2014, Fig. 10.20a – see Fig. 14.1).

Now the first thing to note is that climate scientists know a lot about the value of the climate sensitivity. According to the (AR5), "there is *high confidence* that *ECS* is *extremely unlikely* less than 1°C and *medium confidence* that the ECS is *likely* between 1.5°C and 4.5°C and *very unlikely* greater than 6°C" (Intergovernmental Panel on Climate Change 2014, 10.8.2). Thus, we know a range of values within which the climate sensitivity is very likely to fall. The second thing to note is that the climate sensitivity is much better constrained on the lower end: it is reported to be, with high confidence, extremely unlikely to be less than 1 °C, but it is only *very likely* less than 6 °C and the latter statement is only made with medium confidence.[4] The third thing to note is that even though we know a lot about the value of the climate sensitivity, what we know is not enough to perform a cost-benefit analysis or welfare maximization analysis. In order to calculate the optimal emissions policy, we would need to know either with certainty what the consequences of different emissions policies would be or (at least) the probabilities with which different consequences would occur. But in order to derive a single probability distribution for *ECS*, we would have to know what the probabilistic dependencies between the different models are from which the IPCC distributions for *ECS* are

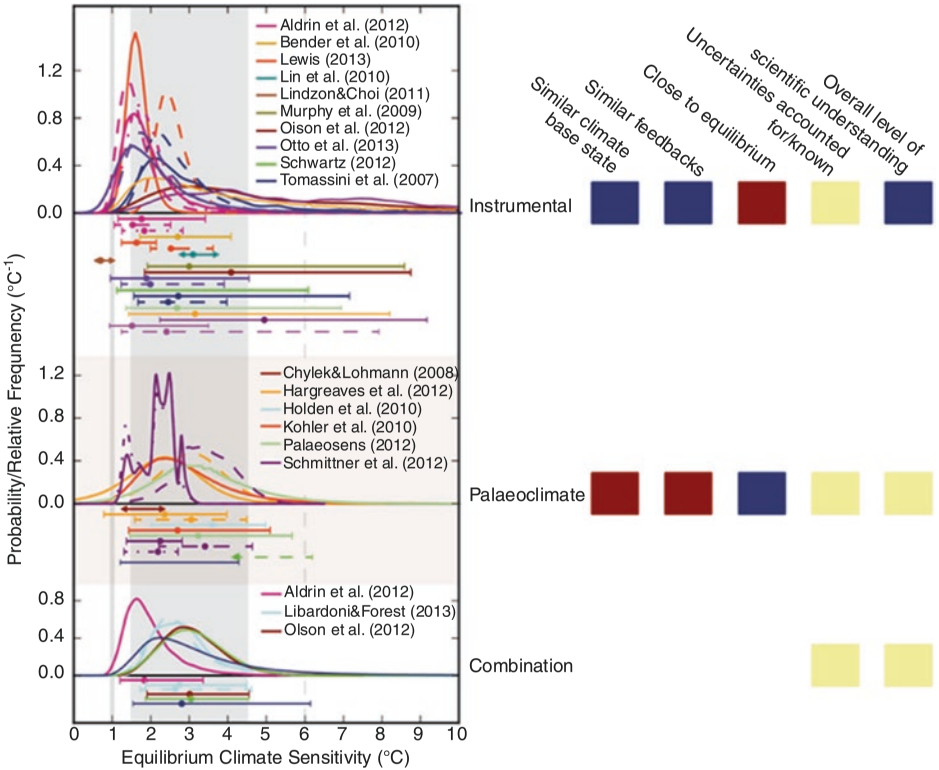

Fig. 14.1 Equilibrium Climate Sensitivity (ECS) estimated from observational constraints (Bindoff et al., Fig. 10.20b, IPCC AR5 WGI 2013, p. 925)

derived. And these dependencies are unknown. Thus, despite the fact that we know quite a bit about the value of *ECS*, we are in a situation of deep or "Knightian" uncertainty with respect to its value—that is, not only is the precise value of *ECS* unknown, but we do not even have grounds for associating a specific probability distribution with *ECS*.

IAMs "solve" this problem by simply positing a probability distribution that is peaked at the center of the IPCC range but no formal justification for this procedure is given. Moreover, the probability distributions posited by different IAMs are symmetric and ignore the possibility of extreme runaway climate change, which according to some models has a low but non-negligible probability of occurring. Some studies cited in (AR5) arrive at probability distributions for *ECS* with very "fat upper tails" that go to zero much more slowly than a normal distribution does and predict small but non-negligible probabilities for values of 6 °C or above for *ECS*. A study by the economist Martin Weitzmann (Weitzman 2012) suggests that the fat upper tails of the distribution have a significant effect on the optimal climate policy under uncertainty and that the expected size of the damages may be sensitive to the precise shape of the tails. But the shape of the tail of the distribution is even more difficult to determine than is its shape in the center, since by its very nature, there are many fewer climate model runs that are associated with tail events. The IAMs considered by the IWG simply ignore this complication by assuming a symmetric, thin-tailed distribution: a normal distribution in the case of Nordhaus's DICE and a triangular distribution (with no tails) in the case of the PAGE model. Hence, IAMs turn deep uncertainties into precise probabilities and do this in a manner that skews uncertainties towards less serious threats.

There is another problem with IAMs' treatment of the climate sensitivity. Most of the probabilities for *ECS* summarized by the IPCC are derived within the framework of different climate models. The probabilities are conditional only on those features of the climate system that are represented within a model but do not take into account any idealizations of the models or any factors that the models abstract away. That is, the IPCC probabilities for *ECS* are what one might call "model-based probabilities" and not yet decision-relevant probabilities. Sophisticated climate models offer an increasingly fine-grained representation of the

climate system and include more and more factors that are believed to be relevant to the overall state of the system. But many potentially important factors are still omitted. These include melting of the Greenland and Antarctic ice sheets; melting of the permafrost and large-scale release of methane from Siberian methane clathrates; release of seabed methane; and the release of carbon stored in the soil as a result of warming. Now it turns out that the factors not represented in the models tend to be *positive feedback factors* that exacerbate the rate of warming and could result in "tipping point" behavior. Indeed, Previdi et al. (2013) argue that *ECS* would be between 4 °C and 6 °C, if the ice sheet and vegetation albedo feedback were included. By contrast, climate scientists are not aware of large potential negative feedback factors that are left out in climate models. As Sir Nicholas Stern puts it, quoting Sir Brian Hoskins: climate models predict "the climate we get if we are very lucky" (Stern 2013, 842). Decision-relevant probabilities would also have to take into account the probabilities of futures in which we might not be so lucky—but these are probabilities that we do not know.

That climate models only provide us with model-based probabilities is not a criticism of climate models. Since the omitted factors cannot yet be adequately modeled, it is reasonable to exclude them. It is also not a criticism of integrated assessment models that take the values of their input parameters from climate models. Rather it is a criticism of the use of IAMs and climate models in climate policy discussions as directly offering us policy advice. The probabilities provided by climate models do not represent scientifically considered degrees of belief concerning future states of the climate system. Rather, they at best are probabilities under certain idealizing assumptions that we know to ignore significant climate risks. Moreover, there is no well-defined procedure for generating decision-relevant probabilities from the model-based probabilities, since there simply is not enough evidence that would allow us to tightly constrain probabilities for the various factors not included in the model. We simply do not know what the probabilistic risks associated with the omitted factors are.

This is not to say that we have no knowledge concerning these risks, just as the fact that there is deep uncertainty concerning the value of the climate sensitivity does not imply we have no knowledge of what its value

may be. For example, there exists a variety of evidence, including paleo-evidence, for temperature bands in which different climate tipping points are likely to occur (see Schellnhuber et al. 2016). The question, however, is how best to represent this kind of knowledge and our different levels of confidence in various claims and what an appropriate decision procedure is in light of existing deep uncertainties.

Summing up this brief survey of the use of climate models in integrated assessment modeling, I have pointed to two distinct problems. First, by ignoring uncertainties models feign precision where none can be had. Users of the models, such as the IWG, posit precise probability distributions for parameters, for which we at best have qualitatively weighted plausible ranges of values. Second, the probability distribution used in calculating the expected utility of various climate policies skew towards what in light of the existing evidence appear to be overly optimistic assumptions. Thus, the upper fat tails of model-based distributions are ignored, as is any adjustment of the model-based distributions to take into account positive feedbacks not modeled. Nordhaus justifies his choice of a normal distribution by saying that introducing a fat tailed "is highly speculative" (Nordhaus 2008, 106). But his own choice of distribution is of course no less speculative and suggests the rather optimistic decision rule that in situations of deep uncertainty we are entitled to limit the choice of models to those that are mathematically tractable and conservative in their damage estimates.

This second problem, it is worth pointing out, extends far beyond the use of IAMs and plagues discussions of the 2 °C goal as well. The total carbon budget that in policy discussions has come to be commonly associated with a 2 °C target, one trillion ton of CO_2 emissions, is calculated with the help of climate models, which give only a 66% chance of temperatures remaining within the target. This, it is important to emphasize, is a model-based probability. So even though the Paris target is best understood as being underwritten by broadly precautionary considerations, using purely model-based probabilities, which arguably underestimate climate risks, and focusing on a carbon budget that even under the models' idealized assumptions promises only a 66% probability of success may strike one as mixing precaution with a strong dose of recklessness.

Modeling Welfare

The second core component of an IAM is an economy model with a welfare function representing overall global welfare at a time. In principle, the concept of welfare equivalent consumption is meant to be very broad and include consumption and enjoyment of any good that we value, which includes goods that are not marketable and do not have a market price, such as, arguably, environmental goods and services. Thus, Nordhaus says: "Economic welfare should include everything that is of value to people, even if those things are not included in the market place" (Nordhaus 2008, 13). In practice, however, explicit claims to the contrary, non-marketable goods are simply ignored and welfare equivalent consumption is measured in terms of GDP.

To the extent that consumption of environmental services is included at all, the use of a single aggregate quantity of consumption as a measure of welfare implies that produced goods and environmental goods are treated as perfectly substitutable for each other (Sterner and Persson 67–68). That is, it is assumed that the rate of exchange between environmental and produced goods is not affected by the relative scarcity of goods of one type with respect to goods of the other and environmental damages can always be substituted for one-to-one by increases in material consumption. As Stephen Gardiner argues (Gardiner 2011), this assumption has the disturbing consequence that overall welfare could continue to increase, even if humans were eventually forced to live under artificial domes due to the negative consequences of climate change, as long as increases in the consumption of produced goods are large enough to make up for the loss in environmental goods.

It is important to distinguish two distinct criticisms in this context. The first, discussed by Gardiner, is the claim that an enjoyment of the environment cannot be monetized at all. Thus, no matter how high the growth in GDP is in the dome world, the loss in non-monetary welfare cannot be made up for by consumption of produced goods. Monetizable and non-monetizable goods are incommensurable, one might hold. Defenders of expected utility theory can respond to this challenge by arguing that as a matter of fact we do reach policy decisions by compar-

ing environmental and economic goods; thus these goods must be commensurable in practice.

The second criticism grants that environmental and economic goods are commensurable, but maintains that it is unrealistic to assume that they will be perfectly substitutable with produced goods. More plausibly, we should expect that as environmental goods become scarcer their relative price would go up, which will result in an increased importance of these goods in the overall economy. Indeed, as Sterner and Persson (2008) have shown for Nordhaus's DICE model, positing a two good economy with an environmental good that is not fully substitutable with an economic good results in a much more stringent climate policy as optimum than if we assume only a single good.

As we have seen, some critics of the IWG's calculation of the social cost of carbon criticize the IWG for appealing to value-laden assumptions. But both the question as to what goods to include in the welfare function and the question to what extent different goods are substitutable depend on our values. In asking what the costs and benefits of a given climate mitigation policy are, we are, of course, asking what the costs and benefits are *for us*. And, hence, any answer to that question essentially depends on what we value. Thus, determining what the optimal climate policy is, is ineliminably also a normative issue. One might reply that the task of regulatory analysis is more narrowly defined as that of calculating the economic costs and benefits of governmental rule making. But that does not avoid but merely postpones the need for a more comprehensive examination of potential impacts of climate policy on human welfare broadly understood and the question as to how we value different impacts.

Thus, it is no criticism of IAMs that they contain normative assumptions. Instead we should ask, first, if a model's normative presuppositions are made as explicit as is possible, and, second, whether a model's normative assumptions are the ones we ought to share.

Future Discounting

If we want an IAM to answer the question whether the costs associated with present-day mitigation measures are justified by their future bene-

fits, the model has to aggregate welfare or utility across different times. Any such cross-temporal aggregation faces the problem what relative weight to assign to utilities at different times. It is common practice to discount future utilities with respect to the present. But what is the correct discount rate to use? There is widespread disagreement in the literature not only on what choice of discount rate is appropriate but even on what the proper methodology for choosing a discount rate ought to be. What makes matters worse is, as we have seen, that predictions derived from IAMs are extremely sensitive to the choice of discount rate. Thus, the stark disagreement on optimal abatement measures between Nordhaus's DICE model and the PAGE model used in the Stern Review is to a significant extent due to the large difference in the presupposed discount rates between the two models: 5.5% in Nordhaus's case as opposed to 1.5% in the Stern Review. Indeed, it is often claimed (not entirely correctly) that whether an IAM recommends modest or stringent abatement measures is largely determined by the choice of discount rate.

There exists a large literature on the problem of future discounting. Here I want to focus on only one issue that highlights both how large the uncertainties affecting cost-benefit analyses are and also the normative consequences of certain modeling assumptions. One prominent source of disagreement is over the issue whether determining the discount rate is a normative or descriptive question (see, e.g., Posner and Weisbach 2010). Advocates of a descriptive approach argue that the discount rate reflects the opportunity costs of an investment and therefore can be determined empirically by examining existing rates of return on investments. Advocates of the normative approach hold that the discount rate depends on ethical judgments concerning our obligations toward the future. Yet, as Fleurbaey and Zuber (2012) have argued, the two sides of the debate aim at different targets.[6]

It is a *normative* question what our ethical obligations toward the future are. Comparing different rates of return on investment can only tell us whether a given investment is an efficient way of transferring wealth to the future, but it does not tell us *that* we should invest for the future. One might reply that expected utility theory answers this question for us: we should invest for the future, if that results in an increase in overall cross-temporal utility or welfare. But making this calculation pre-

supposes that we can compare and aggregate welfare across time. And this is what the discount rate allows us to do by enabling us to calculate the net present value of cross-temporal welfare.

There are two normative issues that affect the net present value of welfare at some time other than the present. The first issue is whether present welfare should be valued more highly just because it occurs in the present. How we decide this issue, is reflected in the choice of the so-called "pure rate of social time preference," which determines to what extent we take present welfare to possess an intrinsically higher value than welfare at other times. If we adopt a temporally egalitarian view, according to which welfare or utility is not valued higher just simply because of the time at which it exists, the pure rate of social time preference would be equal to zero.

The second normative issue concerns the shape of the utility function. It is common to adopt the law of diminishing marginal utility and assume that marginal utility declines. Thus, the discount rate is partly a measure of our (not inherently intertemporal) inequality aversion.

The two normative issues need to be distinguished from the descriptivists' concern. Once we have decided on normative grounds what our debt to the future ought to be, we can consider, for a given investment, what its opportunity costs will be, and thus ask whether a given policy or project provides us with the most efficient means are for transferring wealth from the present to the future. Here, comparison with market rates matter. That is, within the framework of expected utility theory, the discount rate (including the measure of inequality aversion) tells us whether and how much we should invest for the future. Market rates then tell us what the most efficient investment is. As Fleurbaey and Zuber point out, if we are only considering different investments with the same time profile, then there is no need to appeal to discount rates in this second step, since we can directly compare the different rates of return. However, if we consider investments with different time profiles—for example, shorter-term market investments with investments into a climate policy with very long-term benefits—we need to compute the net present value of the different investments to be able to compare their overall values.

If we set the pure rate of social time preference equal to zero, as many philosophers and economists argue we should,[7] the choice of discount

rate is an expression of inequality aversion. Standard IAMs consider a representative consumer at each time and consider the wealth transferred between representative consumers at different times. If we assume positive worldwide economic growth, then we are justified to discount future welfare simply because the future will be richer than we are today. But as Fleurbaey and Zuber (2012) show, this assumption changes dramatically, if we adopt a slightly more complex model that allows for different populations at each time with different levels of wealth. In a model that includes the transfer of wealth between different populations at different times (with different discount rates depending on the relative wealth of the respective populations), the overall discount rate will, in the long run, be mathematically dominated by the discount rate governing the transfer from worst-off populations in the present to worst-off populations in the future.

The effect of this is that discount rates may very well be negative. If the costs of a climate policy are carried by the high emitters of greenhouse gases, who also tend to be among the most affluent, and the beneficiaries include many of the future poor (as will arguably be the case), then such a policy should be evaluated with a negative discount rate. What is more, if we make the not implausible assumption that there will be climate change losers among the poorest populations, wealth transfer from the present poor to the future poor will be transfer from a comparatively richer population to even poorer populations in the future. Thus, even a climate policy that asks the present-day poor to contribute to mitigation costs could have a negative discount rate, as long as there are climate change losers among the presently poor populations. And this will be so even if we assume that global welfare will increase overall and, hence, that more highly idealized IAMs, which only posit a single representative consumer at each time, will posit a positive future discount rate.

As the IWG's analysis shows, the choice of discount rate has a large effect on the optimal climate policy and on the social cost of carbon. Moreover, since the choice of discount rate is a measure of inequality aversion (and also a measure of the degree of pure time preference for the present), it reflects ethical assumptions. In this sense, then, the criticism by the IER economist Murphy is correct: the overall SCC is "a very malleable figure dependent on subjective modeling assumptions, and can be

made large, small, or even negative depending on parameter choices." The choice of discount rate does not purely represent non-normative facts. Yet Murphy's characterization of the discount rate as a subjective modeling assumption makes it appear as if the choice was arbitrary. But that does not follow. Just as we can hope that our choice of climate parameters represent the climate system adequately (in a given context and for a given purpose), so we can hope that the normative parameters adequately represent either our actual normative preferences or the preferences we ought to have.

In fact, the problem with the IWG's calculation of the social cost of carbon is not, as Murphy charges, that it is based on moral or ethical assumptions. The climate problem is a moral problem just as much as it is a scientific problem and any adequate discussion of climate policy has to engage with the moral challenges raised by climate change. Rather the problem is that some of the modeling choices and even the choice of modeling framework, first, imply normative assumptions that are not made explicit and, second, restrict the type of ethical considerations that can be brought to bear on our policy choice.

By modeling each generation in terms of a single representative consumer standard IAMs ignore inequalities among each generation with significant consequences for the choice of discount rate. But the problem runs deeper than that. Not only do standard IAMs ignore wealth inequalities among each generation, but the utilitarian framework within which the IGW's calculation is performed is blind to considerations of justice and harm, which arguably are at the core of the moral problem posed by climate change.

There are strong intuitions, I take it, that Fleurbaey and Zuber's result gets at something important: a model allowing for different contemporaneous populations with an overall discount rate that is dominated by that between the worst-off at different times seems to get something right. I want to submit that this intuition has less to do with the fact that the model takes inequality aversion more comprehensively into account than models with a single representative consumer do. Rather, the intuitive appeal of the result lies in the fact that it ends up favoring those who *both* are the most threatened by climate change *and* are the least responsible for climate change. Fleurbaey and Zuber themselves hint at this when

they support their analysis by saying that "mitigation efforts, when they are well conceived, should put the burden on the high emitters who are typically among the affluent members of the present generation" (15). But that well-conceived mitigation measures should put the burden on high emitters does not fall out of a pure cost-benefit analysis, which at best can favor a redistribution of welfare from the rich to the poor. Instead, that high emitters should carry the main costs of a climate policy is suggested by principles of fairness or justice (see, e.g. Shue 2014). We, as high emitters, are harming future generations, and in particular the future poor in less developed countries, by threatening to deprive them of an environment in which they can support themselves and thrive. A cost-benefit analysis is blind to this fact, even if it can take the situation of the poorest populations into account "through the backdoor" through a negative discount rate.

The Damage Function

The last aspect of IAMs I want to comment on is the so-called "damage function" that represents the effects of temperature increase on economic systems. IAMs couple a climate model to a welfare function in two ways: first, consumption is taken to affect GHG emissions and, second, climate change is taken to affect consumption through a damage function.

Nordhaus's DICE model assumes that the effects of climate change on the economy can be represented in terms of a single temperature-dependent damage function of the form $C(T)=1/[1+(T/a)^b]$. Nordhaus chooses a quadratic function with $b=2$, a choice that has become particularly influential in IAM modeling. The PAGE and FUND models contain more complex treatments of damages, aggregating several sector and region specific damages. Climate damages are significantly more uncertain than the climate system's response to GHG emissions. While the treatment of damages by the three models is to some extent supported by climate impact studies, empirical constraints here are much weaker than in the case of climate models and damage models rely largely on older impact studies, some dating to the 1990s (National Academies of Sciences 2017). In addition, the different treatments of

damages are not independent of each other (with PAGE as the most recent model having been partly calibrated against FUND and DICE); thus the IWG's procedure of averaging the three different damage functions is not an adequate method for sampling the full space of potential damages.

Moreover, standard IAMs arguably severely underestimate possible climate damages. For a difficult-to-fathom 8 °C increase in global temperatures, the DICE and PAGE models predict "only" damages amounting to roughly 15% global GDP—a loss that would correspond to taking the 20,106 US GDP back to its 2004 level—while the FUND model predicts damages of roughly 6% GDP (i.e. less than a three-year delay for the United States in GDP growth as compared to no climate change).[8] Thus, one may worry that damage estimates by the main IAMs are extremely optimistic. This worry gains some support from a recent paper that reaches dramatically different conclusions about prospective climate damages: Burke et al. (2015) consider economic data from 166 countries in the years 1960–2010 to determine how changes in economic activity are coupled to annual fluctuations in temperature. Their study finds that under business as usual damages in 2100 will be an entire order of magnitude higher than those predicted by the IAMs used by the IWG, even when focusing on temperature effects alone and ignoring sea-level rise (see Sect. 14.2). In addition, Burke et al. find that damages will be extremely unequally distributed and will lead to an extreme widening of global income inequality. While climate damages in North America or Central Europe may be small enough to allow overall economic growth, Sub-Saharan Africa or South Asia are projected to see a catastrophic 75% decrease in GDP per capita (Fig. 14.2).

Thus, again we find that standard IAMs do not adequately represent the uncertainties associated with climate change and arguably underestimate the threats to the poorest and most vulnerable populations. Since the DICE model only has a single aggregate damage function, it cannot discriminate between different regional climate risks. But even the FUND model, which contains the most sophisticated treatment of damages, in addition to its overall rather optimistic view on damages ends up downplaying damages to the poorest countries by aggregating estimates of absolute values of damages. Thus, what may be devastating losses to GDP

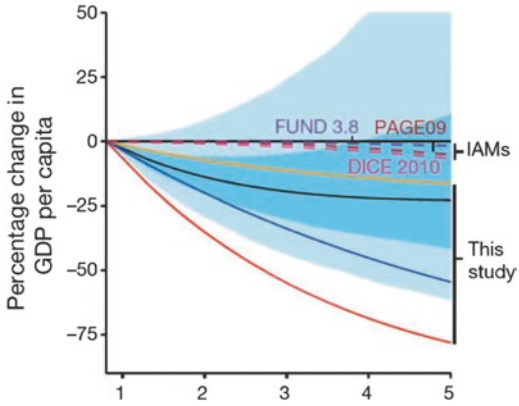

Fig. 14.2 Calculation of prospective damages from business-as-usual climate changes (From Fig. 5b, Burke et al. 2015, p. 4)

in some of the poorest countries will only show up as apparently relatively manageable reductions in global GDP.

14.4 Uncertainty and Precaution

Our brief survey of some of the core features of the IAMs used by the IWG to calculate the social cost of carbon suggests that some of the US conservatives' criticisms of the IWG's approach are warranted. In particular, the IWG posits precise values for some parameters and probability distributions for the values of other parameters even though the values of the parameters in question are deeply uncertain and we do not know what the appropriate probability distributions associated with their values would be. Thus, critics are right in contending that the IWG's estimates of the SCC depend on scientifically not sufficiently constrained modeling assumptions. The IWG treats climate damages as a problem of risk with known probabilities, which cannot be fully scientifically justified, rather than as a problem of deep uncertainties. In this sense, Murphy is right that the IWG's modeling assumptions are "subjective."

One might argue that the IWG is aware of this problem and responds to it by not just calculating a single value for the SCC but by proposing

three different values. In fact, one response to the decision problem under deep uncertainty faced by the IWG might be to run IAMs under a whole range of different assumptions exploring the space of scientifically plausible values for its key parameters and thereby derive a range of value for the SCC. But, first, this is not really what the IWG is doing, since the only quantity that they vary in their three estimates is the discount rate; and even here the IWG dramatically underestimates the range of plausible values. On the one hand, one of Murphy's complaints is that the IWG does not also calculate a SCC for a discount rate of 7%, even thought the US Office of Management and Budget instructs that a 7% rate should also be used for regulatory analysis in addition to a 3% rate. On the other hand, there are compelling arguments for using a negative discount rate, as we have seen above.

Second, if we use IAMs to generate a range of values for the SCC, we would have moved outside of expected utility theory since our calculations would no longer by themselves be able to deliver an answer to the question as to what climate policy to adopt. Rather the range of calculated values would have to be supplemented by an additional decision strategy that tells us how to decide on a value for the SCC given the calculated range of values.

What decisions strategy ought we then to adopt? In this section, I will discuss four influential responses to the problem of deep uncertainties in the context of climate change, ultimately arguing for a precautionary approach.

One response to the problems I have discussed is to argue that they do not provide a sufficient reason to abandon expected utility theory. Thus the philosopher John Broome writes:

> The lack of firm probabilities is not a reason to give up expected value theory. You might despair and adopt some other way of coping with the uncertainty […] That would be a mistake. Stick with expected value theory, since it is very well founded, and do your best with probabilities and values. (John Broome in *Climate Matters*)

Broome's answer to the lack of probabilities is that we simply should do the best we can with positing sharp probability distributions. And

arguably it is possible to address at least some of the problems we discussed within the framework of expected value theory. Models could more fully keep track of uncertainties and modelers could choose probabilities that are not unjustifiably optimistic and that do a better job at taking catastrophic risks into account than existing IAMs do. Moreover, we should develop models that provide representations of social and economic systems that are fine-grained enough to allow us to capture inequalities among different populations as far as threats and damages are concerned. Broome would argue that once we do much better with probabilities and values than existing IAMs do, then expected value theory, due to its formal foundation, does provide us with the best tool for deciding on a climate policy.

Positing precise probability distributions, one might hold, is no different in kind from the standard practice of making idealizing assumptions in science. Every scientific model will be idealized in some ways—expected utility theory with its probability distributions is no different. But while some of the uncertainties that we discussed are sufficiently constrained that we could plausibly be expected to trust models that do their "best with probabilities and values," as far as these aspects of the model are concerned, other uncertainties are so severe that any probability distribution will appear arbitrary and not scientifically sufficiently motivated. The value of the climate sensitivity ECS may be an example of the former kind. We may be able to settle on some reasonable asymmetric, somewhat fat-tailed distribution that does a reasonably good job at capturing all the probabilistic information about the value of ECS that climate models provide. It is much less clear, however, how to move from such a model-based distribution to a decision-relevant probability distribution, since it is unclear whether we have reasonably reliable estimates for the probabilities of the various factors ignored in our models. And what an appropriate probability distribution for globally aggregated climate damages ought to look like is even less constrained by available evidence.

We might nevertheless want to concur with Broome and argue that we just have to do the best we can in picking probability distributions, even though some of the probabilities will be to some extent arbitrary. Yet the danger with sticking with expected value theory is that there will have to

be some motivation underlying our choices—be it implicit values or purely subjective preferences—and this motivation will be buried underneath the purported mathematical precision of our calculation. The mathematically precise outputs of our models suggest a scientific well-foundedness that belies the deep uncertainties of the inputs.[9] In addition—and equally as troublesome—expected value theory is, as I have argued above, blind to central moral dimensions of the climate problem, such as considerations of justice.

If we do not follow Broome's advice and give up on expected value theory, which decision strategy ought we to follow? Before discussing a precautionary approach, I want to distinguish two different decision strategies in the face of deep uncertainty that, while perhaps not scientifically important, are politically quite influential and have more or less explicitly been endorsed by officials and policy advisors associated with the Trump administration. The first strategy, which has been advocated by Secretary of State Rex Tillerson, is that in situations of deep uncertainty we should ignore any deeply uncertain consequences of our policy decisions. Thus Tillerson has said: "It is a judgment of balance between future climatic events which could be catastrophic but are unknown, by the IPCC's own acknowledgement, and more immediate needs of humanity today to address poverty, starvation, broad-based disease control, and the quality of life that billions of people are living in today, which is unacceptable to many of us."[10]

The current poor, it is often claimed, are energy-poor. They rely on energy—and, hence presently still on carbon emissions—not only to escape poverty but as a means of survival. On this point, Tillerson will find broad agreement. But while some would argue that the energy-needs of the current poor are a constraint on climate policy just as much as the threats associated with unabated emissions are (see e.g. Shue 2014), Tillerson suggests that, because of the deep uncertainties in climate predictions, in deciding on whether or not to reduce carbon emissions we ought to focus exclusively on the needs of the present poor (and on the value associated with maintaining our "quality of life"). Faced with the need to help the present poor and uncertain future climate threats, the fact that the latter threats are infected by deep uncertainties allows us to ignore these threats—or at the very least, the deep uncertainties associ-

ated with precise climate predictions are sufficient to trump any concern about catastrophic climate change, no matter how catastrophic possible climate damages might be.

Now, one might think that it is obviously a mistake to ignore threats simply because they are associated with deep uncertainties. Yet there is an argument that suggests that not ignoring deeply uncertain threats can result in paralysis. Take a situation in which uncertain threats suggest a certain policy response, perhaps based on precautionary considerations. In many cases, we can imagine highly speculative and outlandish chains of events, according to which the proposed policy would itself result in catastrophic consequences. And while the chain of events we are imagining might be utterly implausible, the more outlandish the scenario we are imagining is, the less likely it might be that we can give a precise, albeit small probability for the scenario's occurrence. Now, if we could associate probabilities with various catastrophic outcomes, expected utility theory would tell us how serious we ought to take these outcomes. The worry is that under conditions of deep uncertainty—that is, without being in possession of probability distributions—we have to treat all catastrophic threats on a par. But then we may frequently find ourselves in situations of paralysis in which a decision strategy will counsel both for and against a given policy. In effect, Tillerson's strategy responds to the problem of paralysis by proposing that deeply uncertain threats ought to be ignored in decision making.

But do we really need to consider all uncertain threats? Just as considering uncertain purely speculative threats may lead to paralysis, ignoring threats just because we cannot associate precise probabilities with them may strike us as reckless. A more promising reply to the argument is the following. Allowing that some uncertainties cannot be represented in terms of precise probabilities does not imply that all deeply uncertain outcomes have to be treated equally. It may be possible to introduce yet more fine-grained non-probabilistic distinctions, but at the very least we should distinguish between threats that have at least some minimal plausibility and arise as the result of reasonably well-understood mechanisms and threats that are outlandish and purely speculative. We only have to consider the latter in decision problems. That is, threats only have to be taken into account if they do not violate what Henry Shue has called an "anti-paranoia requirement" (Shue 2015, 88).

Climate threats clearly satisfy Shue's requirement. Recall our discussion of the climate sensitivity above. The basic physical mechanisms behind anthropogenic climate change are well understood. Even though no precise probability distribution for ECS can be given, we know that it is likely to be between 1.5 °C and 4.5 °C (at least under climate models' idealizing assumptions). And while the science of climate damages is much less settled than basic climate science is and, hence, Burke et al.'s estimate of the damage function is a lot more uncertain than estimates of ECS, their study is based on a wealth of historical data and suggests a rather robust relationship between temperature changes and economic activity.

Tillerson's strategy is to completely ignore threats in situations characterized by deep uncertainty. A second decision strategy is implicit in the claims by the IER's Murphy and by Trump's advisor Perry that the SCC ought to be close to zero. Murphy and Perry's strategy is to cherry-pick data and modeling assumptions in the face of uncertain predictions. As Murphy and others correctly point out, there does exist a selective body of evidence that jointly appears to support the view that climate change poses no serious threat. Thus, there are some studies that arrive at a value for the climate sensitivity at the lower end of the IPCC range (e.g., Otto et al. 2013); the FUND model posits that the effects of climate change will be positive up to a temperature increase of 2.5 °C and, some argue, nevertheless overestimates damages (see Johnston 2016); and the OMB instructs that a discount rate of 7% be one of the values used in regulatory analysis. Combining these assumptions may indeed result in a SCC that is close to zero or even negative.

Yet Murphy and Perry's cherry-picking of assumptions appears even less justified than Tillerson's strategy. While Murphy is correct that some climate models predict a value for the ECS of only 1.5 °C, others allow for a value of up to 6 °C. Thus, Previdi et al. (2013) argue that ECS is significantly higher than the IPCC range, between 4 °C and 6 °C, if ice sheet and vegetation albedo feedbacks are included. While he is correct that the FUND model predicts that economies will be relatively resilient in the face of rising temperatures, Burke et al. (2015) predict damages an order of magnitude higher than standard IAMs and a catastrophic drop in economic activity in many regions of the world under a business-as-

usual scenario. And while Murphy is correct that the OMB instructs that a discount rate of 7% also be used, taking the threat to poorer populations seriously suggests that we use a negative discount rate. Now, there may be scientific reasons to trust some of these predictions more than others. But if, as Murphy himself argues, our knowledge of the relevant parameters ultimately remains deeply uncertain, proposing to base policy decisions on a combination of the most optimistic assumptions suggests a preference for an extremely risk-seeking decisions strategy.

There is, of course, another approach to decisions under conditions of severe uncertainty that provides an alternative to the strategies advocated by Tillerson and Murphy—an approach that in light of the severity of the threats posed by climate change seems significantly more prudent than its two rivals—and that is a precautionary approach. Henry Shue (2015) has identified three conditions under which prompt precautionary action is required. These conditions are: (i) we are facing the possibility of massive, catastrophic losses; (ii) the mechanisms by which these losses can occur are well understood and are scientifically plausible, even though we cannot give precise probabilities; and (iii) the costs for preventing these losses are not excessive. The second condition is the anti-paranoia requirement. The third condition, like the second condition, helps to prevent paralysis. If there were uncertain possible costs of preventing a catastrophe that are comparable to possible losses from the catastrophe, then a precautionary approach might recommend for and against taking preventive measures (see also Steel 2014). All three conditions are met in the case of climate change.

Now one might think that the choice between Murphy's or Tillerson's decision strategy and a precautionary approach is ultimately a matter of taste. Prudence might suggest a precautionary approach, particularly in light of the stakes involves, but those brave or reckless enough might favor a more risk-seeking strategy. But there is a further feature of the climate problem that implies that a precautionary approach may not only be prudent but is in fact morally required. What is morally odious about Tillerson's or Murphy's extreme risk-seeking approach to climate change is that they seek risks not for themselves, but that their strategy exposes future generations, and in particular the future poor, to grave threats. In deciding on a climate policy, we are in the main not considering threats

of possible harm to ourselves but threats to others—we are considering whether or not to engage in activities that threaten to expose others to grave harms. If future generations, and in particular future populations in less developed countries and in countries more exposed to climate risks, have a right to food, to water, to shelter, and, more generally, to an environment that sustains them and in which they can thrive, then we have a moral duty not to engage in activities that seriously put these rights in jeopardy. Moreover, as Shue has argued, our duty is not mitigated by the fact that the threats at issue are uncertain. As Shue says: "If I play Russian roulette with your head for my amusement as you doze and the hammer of the revolver falls on an empty chamber, I will have done you no physical harm. But I will have seriously wronged you by subjecting you to that unnecessary risk" (Shue, Deadly Delays, 152). We wrong future generations by exposing them to the risk of serious harms as a consequence of our actions.

I have argued here for a precautionary approach to climate policy. Yet I have not proposed a specific decision rule. Rather I have argued for precaution as a procedure rule or for, what Steel calls a "meta-precautionary principle" (Steel 2014, 9). Steel's meta-cautionary principle "asserts that uncertainty should not be a reason for inaction in the face of serious environmental threats" (Ibid.). One motivation for this principle is that it prevents paralysis. But Tillerson's and Murphy's decision rules satisfy this aspect of the meta-principle as well. Ignoring deep uncertainties or basing one's decision on the most optimistic scenarios also avoids paralysis. Thus, it is an important part of the principle that consideration of the threat enter into our decision making process.

Evaluating concrete decision rules embodying a precautionary approach is beyond the scope of this chapter, but I do want to end with a few remarks concerning conditions on such a principle if it is to be able to underwrite a climate goal, such as the Paris targets. Shue's conditions on precautionary action point to two important requirements for taking precautionary actions: that the losses we face are massive and much larger than the costs for preventing these losses; and that the threats pass some plausibility test. Now, the advantage of precautionary reasoning is that no more precise knowledge of probabilities seems required. Thus, several authors have proposed versions of a precautionary principle that require

only that we be able to compare precautionary strategies which do not lead to a catastrophe with strategies that do result in catastrophic outcomes under at least some scenario. (See, e.g. Gardiner (2006) and the critical discussion in Steel (2014, sec. 3.3).) For example, neither the maximin rule nor minimax regret requires more fine-grained information about the likelihood of various outcomes aside from an ability to distinguish plausible from radically implausible outcomes. And the same holds for a precautionary principle proposed by Steel in his discussion of climate change, which states: "If a scientifically plausible mechanism exists whereby an activity can lead to a catastrophe, then that activity should be phased out or significantly restricted" (Steel 2014, 30–31). As Steel argues, there are important differences between these different ways of spelling out a precautionary principle, but all require for their applicability only that an activity can plausibly lead to a catastrophic outcome and that there are situations in which the catastrophic losses can be prevented at relatively modest costs. No more fine-grained information on likelihoods is needed.

If our aim, however, is to decide on a quantitative policy target, such as the Paris temperature targets, then we need more detailed information, as scientific discussions of the Paris targets make clear. Deciding on target temperature range and deciding between a 2 °C and a 1.5 °C target requires evaluating probability ranges for various tipping points (see Schellnhuber et al. 2016) and probabilities for impacts such as the frequency of hot temperature extremes, changes in precipitation and precipitation extremes, changes in crop yields, and sea-level rise (see Schleussner et al. 2016). All of these impacts involve considerable uncertainties. But in order to be able to discriminate between different climate policies, we need to appeal to at least qualitative differences in severity or in likelihoods of impacts. And climate scientists do report probabilities for various tipping points and impacts. Now, many of these probabilities are be model-based probabilities derived from specific climate model runs. Other reported probabilities are based on expert judgment. How to think of the information conveyed by these various probabilities and how to incorporate this information into a (broadly precautionary) decision procedure under uncertainty remains an open question.

As (Knutti et al. 2016) argue, there is no purely scientific argument for the 2 °C temperature limit. Rather it is an anchoring device that was decided on based on a combination of scientific arguments, moral arguments, potential costs and feasibility. Deep uncertainties, some of which I have discussed in this chapter, imply that it is a mistake to think of 2 °C as a guardrail or as a safe upper limit. Purely precautionary considerations would arguably favor a more stringent target than 2 °C or even 1.5 °C. But the lower a proposed temperature target is, the more trade-offs between the target's potential benefits and its potential costs become important—that is, as we consider ever lower possible targets the kind of considerations lying at the heart of an expected utility analysis become relevant, even though, as we have seen, the probabilistic information required for such an analysis is not available. We are thus faced with the problem of finding a decision principle that does not require precise decision-relevant probabilities as input, as expected utility theory does, but is sensitive to more fine-grained distinctions among likelihoods or levels of plausibility than paradigmatic applications of precautionary reasoning are.[11]

14.5 Conclusion

In this chapter, I discussed two main approaches to climate policy making that have dominated US and international climate policy discussions in recent years: expected utility calculations and a precautionary approach. The former approach provided the formal framework for the Obama administration's attempts to calculate a value for the social cost of carbon. The latter approach has provided the guiding principle for the United Nations Conference of Parties since the Rio Declaration in 1992 and is one important motivation for the Paris Agreement.

I have argued that the deep uncertainties characterizing our knowledge of future states of the climate system and of climate damages make the exercise of trying to calculate a single well-supported value for the social cost of carbon impossible. Moreover, the framework of cost-benefit analysis used by the IWG is blind to important moral dimensions of the cli-

mate problem. At best, then, the type of calculations performed by the IWG could provide us with a range of possible climate costs associated with our emissions, which might inform but cannot determine a specific choice of climate policy.

The only morally acceptable alternative framework for climate policy decisions is provided by a broadly precautionary approach: unless we want to gamble immorally and recklessly with the lives and the wellbeing of future populations existing uncertainties require of us to embrace some type of precautionary approach. While the Paris Agreement appears to be to some extent motivated by a broadly precautionary approach, a precautionary strategy would arguably have resulted in a more stringent target. Yet it is an open question to what extent a precautionary approach can result in such specific policy recommendations and deliver more than a general call for urgent mitigation measures.

Notes

1. In a memo leaked to the press. See https://cleantechnica.com/2016/12/08/leaked-transition-team-memo-outlines-trumps-catastrophic-energy-agenda/ accessed Feb 17, 2017, 1 pm EST.
2. https://www3.epa.gov/climatechange/Downloads/EPAactivities/social-cost-carbon.pdf accessed on 2/25/2017.
3. For other critical discussions of integrated assessment models, see Ackerman et al. (2009); Frisch (2013); Pindyck (2013).
4. For an explanation of the two axis along with the IPCC report expresses confidence, see the IPCC Guidance Note (Mastrandrea et al. 2010).
5. See also the discussion of discount rates in Posner and Weisbach (2010) and the criticism thereof in Frisch (2012).
6. Here is how Frank Ramsey put the issue: "it is assumed that we do not discount later enjoyments in comparison with earlier ones, a practice which is ethically indefensible and arises merely from the weakness of the imagination" (quoted in Weitzman 2012).
7. See Interagency Working Group on Social Cost of Carbon (2010).
8. Just how much uncertainty remains, especially as far as regional predictions of changes to the climate system are concerned, has recently been

underscored by World Climate Research Programme Director David Carlson of the WMO, who in the context of discussing "heat waves" in the Arctic in the winter of 2016/17 said: "Even without a strong El Niño in 2017, we are seeing other remarkable changes across the planet that are challenging the limits of our understanding of the climate system. We are now in truly uncharted territory," (https://public.wmo.int/en/media/press-release/climate-breaks-multiple-records-2016-global-impacts, accessed on March 21, 2017).

9. https://thinkprogress.org/exxons-ceo-just-won-his-shareholders-rejected-climate-change-proposals-573d12dde5e7#.h7xpxfs4x. Accessed 2/28/2017.
10. While there are several formal frameworks for representing reasoning under conditions of severe uncertainty (for an overview, see Kunreuther et al. 2014), it remains to be seen to what extent these frameworks can provide tools for policy decisions.

References

Ackerman, Frank, Stephen J. DeCanio, Richard B. Howarth, and Kristen Sheeran. 2009. Limitations of Integrated Assessment Models of Climate Change. *Climatic Change* 95 (3–4): 297–315. https://doi.org/10.1007/s10584-009-9570-x.

Bindoff, Nathaniel, Peter Stott, Krishna AchutaRao, Myles Allen, Nathan Gillett, David Gutzler, Kabumbwe Hansingo, et al. 2013. Chapter 10 – Detection and Attribution of Climate Change: From Global to Regional. In *Climate Change 2013: The Physical Science Basis. IPCC Working Group I Contribution to AR5*. Cambridge: Cambridge University Press.

Broome, John. 2012. *Climate Matters: Ethics in a Warming World*, Norton Global Ethics Series. New York: W. W. Norton & Company.

Burke, Marshall, Solomon M. Hsiang, and Edward Miguel. 2015. Global Nonlinear Effect of Temperature on Economic Production. *Nature* 527 (7577): 235–239. https://doi.org/10.1038/nature15725.

Fleurbaey, Marc, and Stéphane Zuber. 2012. *Climate Policies Deserve a Negative Discount Rate*. Working s, HAL.

Frisch, Mathias. 2012. Climate Change Justice. *Philosophy & Public Affairs* 40 (3): 225–253. https://doi.org/10.1111/papa.12002.

———. 2013. Modeling Climate Policies: A Critical Look at Integrated Assessment Models. *Philosophy & Technology* 26 (2): 117–137. https://doi.org/10.1007/s13347-013-0099-6.

Gardiner, Stephen M. 2006. A Core Precautionary Principle*. *Journal of Political Philosophy*14(1):33–60.https://doi.org/10.1111/j.1467-9760.2006.00237.x.

———. 2011. *A Perfect Moral Storm: The Ethical Tragedy of Climate Change*. Cary: Oxford University Press.

Greenstone, Michael, and Cass R. Sunstein. 2016. Donald Trump Should Know: This Is What Climate Change Costs Us. *The New York Times*, December 15.

IPCC. 2010. *Appendix 15A. Social Cost of Carbon for Regulatory Impact Analysis under Executive Order 12866*. Final Rule Technical Support Document (TSD): Energy Efficiency Program for Commercial and Industrial Equipment: Small Electric Motors. U.S. Department of Energy. U.S. Department of Energy, Washington, DC.

———. 2014a. Evaluation of Climate Models. In *Climate Change 2013 – The Physical Science Basis*, 741–866. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781107415324.020.

———. 2014b. Integrated Risk and Uncertainty Assessment of Climate Change Response Policies. In *Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press.

———. 2014c. Social, Economic and Ethical Concepts and Methods. In *Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge/New York: Cambridge University Press.

Johnston, Jason Scott. 2016. The Social Cost of Carbon. *Regulation* 39: 36.

Knutti, Reto, Joeri Rogelj, Jan Sedláček, and Erich M. Fischer. 2016. A Scientific Critique of the Two-Degree Climate Change Target. *Nature Geoscience* 9 (1): 13–18. https://doi.org/10.1038/ngeo2595.

Kunreuther, Howard, Gupta Shreekant, V. Bosetti, R. Cooke, V. Dutt, M. Ha-Duong, H. Held, et al. 2014. Integrated Risk and Uncertainty Assessment of Climate Change Response Policies. In *Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, ed. O. Edenhofer, R. Pichs-Madruga, Y. Sokona, E. Farahani, S. Kadner,

K. Seyboth, A. Adler, et al. Cambridge/New York: Cambridge University Press.

Mastrandrea, Michael D., Christopher B. Field, Thomas F. Stocker, Ottmar Edenhofer, Kristie L. Ebi, David J. Frame, Hermann Held, et al. 2010. Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties.

National Academies of Sciences, Engineering. 2017. *Valuing Climate Damages: Updating Estimation of the Social Cost of Carbon Dioxide*. Washington, DC: The National Academies Press.

Otto, Alexander, Friederike E.L. Otto, Olivier Boucher, John Church, Gabi Hegerl, Piers M. Forster, Nathan P. Gillett, et al. 2013. Energy Budget Constraints on Climate Response. *Nature Geoscience* 6 (6): 415–416. https://doi.org/10.1038/ngeo1836.

Pindyck, Robert S. 2013. Climate Change Policy: What Do the Models Tell Us? *Journal of Economic Literature* 51 (3): 860–872.

Posner, Eric A., and David Weisbach. 2010. *Climate Change Justice*. Princeton: Princeton University Press.

Previdi, M., B.G. Liepert, D. Peteet, J. Hansen, D.J. Beerling, A.J. Broccoli, S. Frolking, et al. 2013. Climate Sensitivity in the Anthropocene. *Quarterly Journal of the Royal Meteorological Society* 139 (674): 1121–1131. https://doi.org/10.1002/qj.2165.

Revkin, Andrew. 2017. Will Trump's Climate Team Accept Any 'Social Cost of Carbon'? *ProPublica*, January 11.

Schellnhuber, Hans Joachim, Stefan Rahmstorf, and Ricarda Winkelmann. 2016. Why the Right Climate Target Was Agreed in Paris. *Nature Climate Change* 6 (7): 649–653. https://doi.org/10.1038/nclimate3013.

Schleussner, Carl-Friedrich, Joeri Rogelj, Michiel Schaeffer, Tabea Lissner, Rachel Licker, Erich M. Fischer, Reto Knutti, Anders Levermann, Katja Frieler, and William Hare. 2016. Science and Policy Characteristics of the Paris Agreement Temperature Goal. *Nature Climate Change* 6: 827–835.

Shue, Henry. 2014. *Climate Justice: Vulnerability and Protection*. New York: Oxford University Press.

———. 2015. Uncertainty as the Reason for Action: Last Opportunity and Future Climate Disaster. *Global Justice: Theory Practice Rhetoric* 8 (2). 10.21248/gjn.8.2.89.

Steel, Daniel. 2014. *Philosophy and the Precautionary Principle: Science, Evidence, and Environmental Policy*. Cambridge: Cambridge University Press.

Stern, Nicholas. 2007. *The Economics of Climate Change: The Stern Review*. Cambridge: Cambridge University Press.

———. 2013. The Structure of Economic Modeling of the Potential Impacts of Climate Change: Grafting Gross Underestimation of Risk onto Already Narrow Science Models. *Journal of Economic Literature* 51 (3): 838–859.

Sterner, Thomas, and U. Martin Persson. 2008. An Even Sterner Review: Introducing Relative Prices into the Discounting Debate. *Review of Environmental Economics and Policy* 2 (1): 61–76.

Tol, Richard S.J. 2002a. Estimates of the Damage Costs of Climate Change. Part 1: Benchmark Estimates. *Environmental and Resource Economics* 21 (1): 47–73.

———. 2002b. Estimates of the Damage Costs of Climate Change, Part II. Dynamic Estimates. *Environmental and Resource Economics* 21 (2): 135–160.

Weitzman, Martin L. 2012. GHG Targets as Insurance Against Catastrophic Climate Damages. *Journal of Public Economic Theory* 14 (2): 221–244.

15

Modeling Mitigation and Adaptation Policies to Predict Their Effectiveness: The Limits of Randomized Controlled Trials

Alexandre Marcellesi and Nancy Cartwright

15.1 Climate Policies: Mitigation and Adaptation

The negative effects of anthropogenic global warming[1] on natural and social systems promise to be diverse and important: melting of glaciers and of the polar ice caps (IPCC 2007a, 356–360) contributing to a rise of sea-levels (op. cit., 418); increase in the frequency and intensity of extreme weather events like droughts, heat waves, or floods (IPCC 2012); decrease in crop productivity resulting in increased risk of hunger (IPCC 2007b, 298); increased risk of extinction for a great number of plant and animal species (op. cit., 792); and so on. Most of these negative effects are

A. Marcellesi
New York University School of Law, New York, NY, USA

N. Cartwright (✉)
Department of Philosophy, Durham University, Durham, UK

University of California, San Diego, CA, USA

E.A. Lloyd, E. Winsberg (eds.), *Climate Modelling*,
https://doi.org/10.1007/978-3-319-65058-6_15

expected to occur regardless of the way emissions of greenhouse gases (GHGs) evolve in the future, and some of them are already being observed.

It is not, however, too late for policy makers to act. First, though many of the effects of global warming will inevitably occur, their intensity depends on how large the rise in average temperature turns out to be. Reducing emissions of GHGs, the cause of anthropogenic global warming, can thus help moderate the intensity of these effects. Second, because most of the effects of global warming will inevitably occur, policies for adapting to these effects and limiting their harmful consequences are necessary.[2]

This chapter is about some of the serious problems we can expect to face in modeling the effects of climate change policies—in evaluating the effectiveness of policies that have been implemented and in predicting the results of polices that are proposed. The difficulties we will discuss are shared with other kinds of social and economic policies, but they can be particularly problematic for climate change policies, as we will show below. Policies for addressing climate change are commonly divided into two categories, mitigation and adaptation, corresponding to the two levels at which policy makers can address climate change.[3] The Intergovernmental Panel on Climate Change (IPCC) defines a mitigation policy as "A human intervention to reduce the sources or enhance the sinks of greenhouse gases" (IPCC 2007a, 949) and an adaptation policy as an "Adjustment in natural or human systems in response to actual or expected climatic stimuli or their effects, which moderates harm or exploits beneficial opportunities" (IPCC 2007b, 869). One can put the distinction between mitigation and adaptation in causal terms by saying that while mitigation policies are designed to reduce the causes of global warming, adaptation policies are designed to moderate its harmful effects on natural and human (or social) systems.

15.2 Evidence-Based Climate Policies

Agencies which fund mitigation and adaptation policies typically want 'their money's worth'; they want to fund policies 'that work', that is policies that produce the effects they are designed to produce

where and when they are implemented.[4] Claims that a given policy 'works', moreover, should be based on evidence. This idea, which is at the root of the widespread evidence-based policy movement, seems natural enough: A policy should be funded, and implemented, only if there is reasonable evidence that it will produce the desired effect in the specific location and at the specific time at which it is implemented.

In order to produce such evidence, organizations implementing policies are invited to conduct 'impact evaluations'. Impact evaluations (IEs) are studies measuring the effects of policy interventions. They are, by definition, retrospective: A policy must have been implemented for its effects to be measured. These IEs have two main functions: First, when an IE establishes that the policy had the effect it was designed to have, it thereby provides a post hoc justification for the decision to fund and implement the policy. Second, the results of IEs are supposed to inform subsequent policy decisions by providing evidence supporting predictions about the effectiveness of policies.

Both functions are important, and this is why many of the agencies that fund policies devote part of their resources to IEs. An example in the domain of climate policies is the Global Environment Facility (GEF). The GEF, an intergovernmental agency which funds many mitigation and adaptation policies, has its own evaluation office, which produces guidelines for conducting IEs.[5]

As we mentioned above, the aim of IEs is to measure the effects of policy interventions. This is essentially an issue of causal inference. Teams of researchers that carry out IEs are, in the words of statistician Paul Holland, in the business of "measuring the effects of causes" (Holland 1986, 945). The extensive literature on causal inference in statistics and related disciplines (e.g., econometrics or epidemiology) provides policy makers with many different methods, experimental and observational, for conducting IEs.

Indeed, the counterfactual approach to causal inference (Rubin 1974; Holland 1986) which is prominent in statistics has had a palpable influence on the field of evaluation. According to the World Bank's guide to impact evaluation, for instance,

> To be able to estimate the causal effect or impact of a program on outcomes, any method chosen must estimate the so-called *counterfactual*, that is, what the outcome would have been for program participants if they had not participated in the program. (Gertler et al. 2011, 8, emphasis added)[6]

As this quotation hints, the idea at the root of the counterfactual approach is that the size of the contribution of a putative cause C to an effect E among program participants is identical to the difference between the value of E for those participants in a situation in which C is present and the value which *E would* take in a situation in which C is absent, all else being equal. If this difference is equal to zero, then C is not a cause of E in that population; if it is greater than zero, then C is a positive cause of E, and if it is smaller than zero, then C is a negative cause of E. According to the counterfactual approach to causal inference, answering the question 'What is the effect of C on E in a given population?' thus requires answering the following counterfactual queries 'What value would E take for individuals in that population exposed to C were C absent, all else being equal?' and 'What value would E take for individuals not exposed to C were C present, all else being equal?'

This commitment to a counterfactual approach goes together with a strong preference for experimental methods, and for randomized controlled trials (RCTs) in particular, over observational methods. According to their advocates,[7] RCTs yield the most trustworthy or, as development economists Esther Duflo and Michael Kremer put it (Duflo and Kremer 2005), "credible" estimates of the mean effect of C on E in a given population. RCTs are, to use a common expression, the 'gold standard' of causal inference.[8]

15.3 What Are RCTs, and Why Are they Considered the 'Gold Standard'?

RCTs are experiments in which individuals in a sample drawn from the population of interest are randomly assigned either to be exposed or not exposed to the cause C, where an individual can be anything

from a single student to a single village to a hospital to a single country or region. Individuals who are exposed to C form the 'treatment' group while individuals who are not exposed form the 'control' group.[9] Random assignment does, in ideal circumstances and along with a sufficiently large sample, make it probable that the treatment and control groups are homogeneous with respect to the net effect of all causes of E besides C. And the homogeneity of the two groups with respect to causes of E other than C enables one to answer the counterfactual question 'What would be the mean value of E for individuals (in the study population) exposed to C were C absent, all else being equal?' by citing the mean value taken by E for individuals not actually exposed to C.[10] In other words, ideally conducted RCTs make it likely, by their very design,[11] that all else is indeed equal between the treatment and control groups, and thus that the actual mean value of E for the control group can be identified with the mean value which E would take for the treatment group were individuals in this group not exposed to C (and vice versa for the control group). This is in turn enables one to estimate the mean of the difference between the effect an individual would have were they subject to C versus were they not—often called the *causal* or *treatment effect* of C on E—in the sample, or study population, accurately.[12]

Here is a different way to put it. Assume that the effect of interest E is represented by a continuous variable Y_i and that the putative cause C is represented by a binary variable X_i taking value 1 when individual i is exposed to the cause and 0 when it is not. Assume also that the relationship between X_i and Y_i in the study population is governed by the following linear causal principle:

$$(\mathrm{CP})\ Y_i = a + b_i X_i + W_i$$

Here W_i is a continuous variable which represents factors that are relevant to the value of Y_i besides X_i. And coefficient b_i represents the effect of X_i on Y_i for i. Since b_i represents the individual-level effect of X_i on Y_i, the population-level mean effect of X_i on Y_i is by definition equal to $\mathrm{Exp}[b_i]$, where Exp[.] is the expectation operator.[13]

Randomly assigning individuals to the treatment and control groups in principle guarantees the probabilistic independence of X_i from both b_i and W_i, and this in turn enables one to accurately estimate Exp[b_i] from the difference between the expected value of the effect in the treatment group and its expected value in the control group.[14] This difference is equal to:

$$\exp\left[Y_i|X_i=1\right]-\exp\left[Y_i|X_i=0\right]=\left(a+\exp\left[b_i|X_i=1\right]+\exp\left[W_i|X_i=1\right]\right)$$
$$-\left(a+\exp\left[b_i|X_i=0\right]+\exp\left[W_i|X_i=0\right]\right)$$

In the ideal case in which assignment of individuals to either treatment or control genuinely is independent of b_i and W_i, this difference is the mean treatment effect—often referred to as just the 'treatment effect'—and can be estimated from the observed outcome frequencies. It is equal to[15]

$$\exp\left[Y_i|X_i=1\right]-\exp\left[Y_i|X_i=0\right]=\exp\left[b_i\right]$$

So the mean treatment effect is non-zero just in case Exp[b_i] is non-zero, which can happen only if b_i is non-zero for some i in the population, which means that for that individual X_i does contribute to the value of Y_i: X_i causes Y_i in that *i*.

Experimental and observational studies in which assignment to the treatment and control groups is non-random are widely considered less desirable than RCTs because their designs, unlike that of RCTs, do not in principle make the causal homogeneity of the two groups (regarding causes of E other than C) probable, even in large samples, or, alternatively, their designs do not guarantee the probabilistic independence of X_i from b_i and W_i. This is why RCTs are considered the 'gold standard' by a large number of social and policy scientists.

If RCTs are the 'gold standard' for measuring the effects of causes, and if the aim of IEs is to measure the effects of policy interventions, then it seems legitimate to conclude that IEs should be designed as RCTs when-

ever possible. Indeed, this is the view advocated by a variety of policy scientists, for instance members of the Jameel Poverty Action Lab (J-PAL) such as Esther Duflo. J-PAL funds and carries out IEs that use RCTs, at the exclusion of any other evaluation methodology.[16] The view that RCTs provide the best evidence regarding the effects of policies is also embraced by the Grading of Recommendations Assessment, Development and Evaluation (GRADE) working group, a group of health scientists that produces standards for rating the quality of evidence. According to GRADE's evidence-ranking scheme, adopted by many agencies worldwide including the World Health Organization, results from RCTs are rated as having 'high quality' while results from observational studies receive a 'low quality' rating (Balshem et al. 2011, 404, table 3). The views of these organizations about RCTs are echoed in hundreds of other agencies dedicated to vetting policy evaluations around the Anglophone world in areas from education to crime to aging to climate change.

So are RCTs a "silver bullet" for policy evaluation, to use an expression from Jones (2009)? How relevant to policy making is the evidence they generate? Should the evidence base for mitigation and adaptation policies be improved by conducting RCT-based IEs? We will argue below that RCTs have important limitations and that the emphasis put on them contributes to obscuring questions that must be answered for the effectiveness of policy interventions to be reliably predicted. In Sects. 15.4 and 15.5 we will show, first in theory and then in practice—using a particular family of mitigation policies as a concrete example, that even if we agree that an RCT is necessary, results from RCTs provide only a small part of the evidence needed to support effectiveness predictions. Then, in Sect. 15.6, we will show that RCTs are ill-suited to evaluate the effects of most adaptation policies. Our main aim is to underline some particular methodological problems that face the use of RCTs to evaluate mitigation and adaptation policies. We use particular policy examples to illustrate these problems. But we do not aim to offer an exhaustive treatment of these particular policies nor of the full range of challenges that arise in evaluating the effectiveness of mitigation and adaptation policies in general.

15.4 The Limited Relevance of RCTs to Effectiveness Predictions

Internal and External Validity

It is common, in the social and policy sciences, to distinguish between the internal and external validity of studies seeking to measure the effects of causes. According to the standard view, a study is internally valid when it produces results that are trustworthy, and externally valid when its results hold in contexts other than that of the study itself.[17] Because RCTs in principle are supposed to yield the most trustworthy estimates of treatment effects, they are also considered to have the highest degree of internal validity.[18]

It is possible for a study to have a high degree of internal validity while having a very low degree of external validity. A particular RCT, for instance, might yield conclusions that are highly trustworthy but which only hold of the study population involved in the RCT and not of any other population. Results from a study are useful for the purpose of predicting the effectiveness of policy interventions only if they are both internally and externally valid. If IEs are to be useful to policy makers, then, they must produce results that have a high degree of external validity, in addition to being internally valid.

What does it take for a study result to be externally valid? It is often said that, for a study result to hold in contexts other than that of the study itself, the circumstances considered must be 'similar' to that of the study.[19] But what makes a set of circumstances 'similar' to some other set of circumstances? We briefly describe a framework, fully developed in Cartwright and Hardie (2012), that enables one to address questions of external validity in a rigorous and fruitful manner.

Causal Roles, Causal Principles, and Support Factors

Causes do not produce their effects willy-nilly, at least not where it is possible to predict these effects. Rather, the effect of C on E in a given population is governed by *causal principles* that hold in that population. These

causal principles can, without real loss of generality, be represented in the form of (CP) above, where C is represented by X_i and E is represented by Y_i.[20] C *plays a causal role* in (CP) just in case it genuinely appears in the equation, i.e., just in case there are values of b_i such that $b_i(X_i = 1) \neq 0$ for some i in the given population. But C does not work alone to produce a contribution to E: It works together with what we call *support factors*. These support factors are represented by b_i in (CP).[21]

The idea that causes work together with support factors derives from the view that causes are INUS conditions in the sense of Mackie (1965). To say that C is an INUS condition for E is to say that it is an *I*nsufficient but *N*ecessary part of an *U*nnecessary but *S*ufficient condition for the production of a contribution to E.[22] Mackie's classic example is that of a fire caused by a short circuit. The short circuit is not individually sufficient to produce a contribution to the fire; other factors, which we call 'support factors', are required: the presence of flammable material, the presence of oxygen, the absence of sprinklers, and so on. These support factors, together with the short circuit, are jointly sufficient to produce a contribution to the fire. But they are not jointly necessary: There are other ways to contribute to a fire, i.e., there are other sets of factors—e.g., sets that have lit cigarettes instead of short circuits—that are also jointly sufficient to produce a contribution to the fire.[23]

Policies are causes, and as such are INUS conditions. They generally cannot produce a contribution to the effect they are designed to address by themselves: They need support factors. And the distribution of these support factors will differ from situation to situation. We can even expect considerable variation in which factors *are* support factors, that is, which factors are needed to obtain a given effect often varies with context. Consider again Mackie's example as an illustration of this point: The short circuit may not require the absence of sprinklers in houses that are not connected to the water supply system in order to produce a contribution to the fire, though it may require the presence of a particularly large amount of flammable material in houses whose walls have been painted using fire-resistant paint in order to produce the same contribution to the fire. There is no 'one size fits all' set of support factors that, together with the cause of interest, will produce the same contribution to the effect in every context. What matters is the presence of the 'right mix'

of support factors, i.e., the presence of the right support factors in the right proportions, and what the 'right mix' consists in often differs from context to context.

The framework briefly sketched above enables one to frame questions about external validity in more precise terms than does the claim that external validity is a matter of how 'similar' sets of circumstances are. To ask whether a trustworthy result from a particular study regarding the mean effect of C on E will hold in a population other than the study population is to ask:

- Does C play the same causal role in the target population as in the study population?
- Are the support factors required for C to produce a contribution to E present in the right proportions in the target population?

When both questions have positive answers, C will make a positive contribution in the target population if it does so in the study population. If either has a negative answer, it is still possible that C will make a positive contribution but the RCT result is irrelevant to predicting whether it will or not—it provides no warrant for such a prediction.

Which Questions Do RCTs Answer?

An ideal RCT for the effect of C on E will give you an unbiased estimate of $\mathrm{Exp}[b_i]$, the mean value of b_i over individuals in the study population, or treatment effect. If true value estimated is larger than 0, then you know that C makes a positive contribution to E for at least some individuals in the study population. And if this value is smaller than 0, then you know that C makes a negative contribution to E for at least some individuals in the study population.[24]

An ideal RCT may thus get you started on your external validity inference by providing you with some trustworthy information about the causal role C plays with respect to E in at least one population, the study population. But it gets you nowhere at all towards learning what you need to know about support factors: An ideal RCT will not tell you what the support factors are (i.e., what b_i represents) nor about individual val-

ues of b_i, i.e., about the effect of C on E for particular individuals, nor for what proportion of the study population C plays a positive, or negative, role.[25]

How much further can an ideal RCT can take you on the way to a reliable external validity inference? The short answer is: not much further. The framework introduced above makes it clear why. First, an ideal RCT will not tell you what the causal principle governing the relationship between C and E in the study population looks like.[26] Second, an ideal RCT will not tell you what the support factors required for C to produce a contribution to E in the study population are, nor how they are distributed. Third, an ideal RCT will not tell you whether C plays the same causal role in the principles governing the production of E in the target population as in the study population. Fourth, an ideal RCT will not give you information about the support factors required for C to produce a contribution to E in the target population, nor about whether the support factors needed in the target population are the same as in the study population (which, very often, is not the case). And you need these pieces of information to produce a reliable prediction about the effectiveness of a policy.

Advocates of RCTs often reply that what is needed to overcome these limitations is more RCTs, but RCTs carried out in different locations.[27] The reasoning underlying this rejoinder seems to be the following: If RCTs conducted in locations A, B, and C all yield positive results regarding the effects of a policy, then you have strong evidence that this policy will produce the same effects when you implement it in a fourth location, call it D. This reasoning, however, is problematic insofar as it assumes without justification that the policy can play the same causal role in D as it does in A, B, or C. Since the RCTs in A, B, and C cannot individually tell you what causal principle is at work in each of these locations, their conjunction cannot, a fortiori, tell you what causal principle is at work in D. And if you don't know what causal principle is at work in D, then you also don't know whether the policy can play there the causal role you want it to play.[28]

Inferring from results in three—or even a dozen or two dozen—different locations, no matter how different they are, to the next one is a notoriously bad method of inference. It is induction by simple enumeration.

Swan 1 is white, swan 2 is white, swan 3 is white…. So the next swan will be white. Of course science does make credible inductions all the time. But their credibility depends on having good reason to think that the individuals considered are the same in the relevant way, that is, in the underlying respects responsible for the predicted feature. In the case of causal inference from RCT populations that means that they are the same with respect to the causal role C plays and with respect to having the right mix of the right support factors.

Policy scientists writing about mitigation and adaptation policies often lament the current state of the evidence base and, naturally, call for its "strengthening" via rigorous IEs (Prowse and Snilstveit 2010, 228). So should agencies which fund and implement mitigation and adaptation policies carry out RCTs? Should the GEF, as a report of its Scientific and Technical Advisory Panel urges (STAP 2010), start designing its policies *as experiments*, and preferably RCTs, in order to improve the evidence base for climate change policies? The discussion above should make it clear that we think that RCTs are of limited relevance when it comes to producing evidence that's relevant for predicting the effectiveness of policies. We illustrate this point in the next section by examining a particular family of mitigation policies.

15.5 Predicting the Effectiveness of Mitigation Policies

Mitigation Via Payments for Environmental Services

Payment for Environmental Services (PES) programs are policies that seek to conserve the environment by paying landowners to change the way they use their land. Environmental, or ecosystem, services (ESs) are loosely defined as "the benefits people obtain from ecosystems" (MEA 2005, 26). PES policies involve a buyer, the user of the ES or a third-party acting on her behalf, and a seller, the provider of the ES.[29]

Thus a person who owns a forest and uses it for a timber activity may provide ESs by stopping this activity and by replanting trees that were cut

down. In this case, the ESs provided consist in the protection of currently existing carbon stocks, via avoided deforestation, and the improvement of carbon sequestration, via the planting of new trees. Both of these ESs are directly relevant to climate change mitigation, though not all PES programs target ESs that are relevant to climate change mitigation. Many PES programs are designed with the conservation of biodiversity as their main aim.[30]

In order to stop her timber activity, the landowner described above must have an incentive to do so. Why stop her timber activity if this means a loss of earnings, and why replant trees if this means a cost without a benefit? This is where PES programs come in: They are supposed to create the incentives necessary for landowners to change the way they use their land and provide an ES. As Engel et al. put it: "The goal of PES programs is to make privatively unprofitable but socially-desirable practices become profitable to individual land users, thus leading them to adopt them" (Engel et al. 2008, 670).[31]

Governmental and intergovernmental agencies see PES programs targeting deforestation as offering a major opportunity for mitigating climate change. A significant portion of the total emissions of GHGs, and CO_2 in particular, comes from deforestation.[32] If PES programs can create incentives to reduce deforestation, especially in developing tropical countries in which deforestation is a major concern, then they can contribute to a reduction in emissions of GHGs, and thus to a moderation of global warming and of its negative effects.[33]

PES programs are modeled after existing conditional cash transfer programs in domains such as development, for instance, the Mexican *Oportunidades* program.[34] There are numerous IEs, including ones that take the form of RCTs, measuring the effects of conditional cash transfer programs that target poverty-reduction and education. This is particularly true for the *Oportunidades* program, first implemented in 1997 (see, e.g., Parker and Teruel 2005). This is not the case for PES programs and, in particular, for those PES programs that are relevant to climate change mitigation. There are few IEs measuring the effects of PES programs on, e.g., deforestation. And there are no completed IEs of PES programs that take the form of an RCT.

The current state of the evidence base for PES programs is deplored by Pattanayak et al., who "see an urgent need for quantitative causal analyses of PES effectiveness" (Pattanayak et al. 2010, 267). "Such analyses", they add, "would deliver the hard numbers needed to give policy makers greater confidence in scaling up PES" (ibid). In this spirit, the report to the GEF mentioned above (STAP 2011) urges the intergovernmental organization to design its policies—including PES programs—as experiments as much as is possible, and this in order to facilitate the evaluation of their effects.

What Will RCTs Add to the Evidence Base for PES Programs?

Responding to the call for an improvement of the evidence base for the effectiveness of PES programs in securing environmental services, MIT's J-PAL, in collaboration with the International Initiative for Impact Evaluation (3ie) and Innovations for Poverty Action (IPA), is currently carrying out an RCT aimed at measuring the effectiveness of a PES program in reducing deforestation and biodiversity loss in the Hoima and Kibaale districts of Western Uganda.[35] Deforestation rates are particularly high in these two districts, where landowners "often cut trees to clear land for growing cash crops such as tobacco and rice or to sell the trees as timber or for charcoal production" (Jayachandran 2013a).

The design of J-PAL's RCT is as follows (Jayachandran 2013b, 311). First, 1245 private forest owners—spread over 136 villages—were identified. They form the RCT's study population. A survey was then conducted to record several of their characteristics: number of hectares of land owned, past tree-cutting behavior, attitude toward the environment, access to credit, and so on. Sixty-five out of the 136 villages—representing 610 landowners—were then randomly assigned to the treatment group, the remaining villages being assigned to the control group. Landowners residing in villages in the treatment group were called into meetings by a local nongovernmental organization (NGO), the Chimpanzee Sanctuary & Wildlife Conservation Trust (CSWCT), to receive information about the program as well as contract forms. The

'treatment' that is randomly assigned in this RCT can thus be described as 'being offered the opportunity to sign a PES contract with CSWCT'. One of the aims pursued by J-PAL's scientists here is to estimate the effect of this treatment on deforestation and biodiversity loss.

Landowners who chose to participate in the program (or take up the 'treatment') then signed contracts with the local NGO. As Jayachandran (2013b, 311) reports,

> The contract specifies that the forest owner will conserve his entire existing forest, plus has the option to dedicate additional land to reforestation. Under the program, individuals may not cut down medium-sized trees and may only cut selected mature trees, determined by the number of mature trees per species in a given forest patch. Participants are allowed to cut small trees for home use and to gather firewood from fallen trees.

Compliance with the contract is monitored via spot checks by CSWCT staff. Landowners who comply receive $33/hectare of forest preserved annually, an amount that was selected because it is assumed to be greater than what landowners would earn from cutting down and selling trees (other than those specified by the PES contract) for timber or charcoal, or from clearing land to grow cash crops (e.g., tobacco). As we indicated above, the assumption guiding the design of this and other PES programs is that agents will modify their behavior—here, will stop cutting down trees—if they are given the right monetary incentives to do so.

This RCT, as the official project description states, is justified by the fact that "although many PES schemes have been undertaken globally, there has not been concrete proof, emanating from scientific empirical data collected from real life PES schemes, that they are effective" (GEF 2010, 6). Note, furthermore, that this study is funded by the GEF, whose administration thus seems to be sensitive to the call for RCT-based IEs of PES programs that can deliver "hard numbers" and give "concrete proof" based on "scientific empirical data" of the effectiveness of "real life" PES programs.

As the project description indicates, one of the aims of the study is to generate, develop and disseminate a "replicable PES model based on lessons learned and best practices" (GEF 2010, 3). The aim of this RCT

thus is not simply to demonstrate the effectiveness of the specific PES programs implemented in the Hoima and Kibaale districts in producing ESs. The explicit aim is to show that PES programs aimed at reducing deforestation and biodiversity loss are effective *in general*, and to develop a PES model that can be scaled up and applied in locations besides select districts in Western Uganda.

Is the RCT currently carried out by J-PAL likely to achieve the result sought? Is it likely to provide strong evidence that PES programs work in general? How much evidence can it provide for this conclusion? If you are a policy maker contemplating the implementation of a PES program, is the RCT likely to provide reasonably strong evidence that such a program will work in the location you are targeting? We do not believe so, for reasons that were advanced in their theoretical form in Sect. 15.4.3. The J-PAL RCT, if it is carried out according to the script, will deliver an unbiased estimate of the mean effect of the PES program on deforestation and biodiversity loss in the study population.

But it will not reveal the causal principle governing the relationship between the PES program and the reduction of deforestation and biodiversity loss in the study population.[36] It also won't tell you what support factors are needed for the PES program to play a positive causal role in the study population, nor how these factors are distributed in this population. The J-PAL RCT will not, a fortiori, tell you where the causal principle at work in the study population also holds in the population you are targeting. And it won't tell you what the support factors required for the PES program to play a positive causal role in the target population are, nor how they will be distributed.

One needs these essential additional pieces of information, regarding causal principles and support factors, in order to predict at all reliably whether the PES program will play the same causal role when it is implemented in other locations, e.g., when it is scaled up to other districts in Western Uganda, or when it is implemented in Eastern Uganda, or when it is implemented in other countries in sub-Saharan Africa, and so on. One cannot arrive at a "replicable PES model", i.e., at a PES model that will work in many locations, without a detailed understanding of how the PES program works in the original study population. Nor is it clear

that there is a reliable "replicable PES model" that works 'in general' to be found. It is not obvious that one can formulate substantial and useful generalizations about PES programs across settings (cultural, political, economic, religious, etc.) and, especially, across types of ESs (Can one generalize results obtained in a context in which the ES is avoided deforestation to a context in which the ES is the preservation of water resources?). The framework introduced above is designed to help you think about how a policy works when it does, and about what it would take for it to work in a different location.

We are obviously not claiming that nothing will have been learned during the four years of the J-PAL project described above, besides an estimate of some treatment effect. The policy scientists carrying out J-PAL's RCTs are neither blind nor stupid. They will gain a wealth of new knowledge regarding the local institutional and social context, the way landowners respond to the PES program, differences between villages that are relevant to the effect of the program, and so on. Note, however, that this context-specific knowledge (1) may well have been acquired even if enrollment in the PES program had not been randomly offered to landowners, (2) is just as important as is knowledge of the treatment effects to predicting the effectiveness of subsequent PES programs, and (3) is likely to be overshadowed by the "hard numbers", i.e., the estimates of treatment effects. The framework introduced above, and fully developed in (Cartwright and Hardie 2012), shows why this context-specific knowledge is essential to predicting the effectiveness of policies. And it also gives you the tools to articulate this knowledge in ways that make it relevant to effectiveness predictions.

The bottom line, here, is that if you are a policy maker contemplating the implementation of a PES program for reducing deforestation and biodiversity loss in a particular location, the results from J-PAL's RCT will offer you some guidance, but not much. You need knowledge about the causal principles at work and the support factors required for the PES program to produce a positive contribution in the location you are targeting. Let us further illustrate the importance of support factors by looking at five hypothesized support factors needed by PES programs in some locations.

Some of the Support Factors (Sometimes) Needed by PES Programs

We briefly list below five of the factors identified in the literature as playing a role in determining the effectiveness of PES programs in reducing deforestation and biodiversity loss.[37] As we noted above (Sect. 15.4.2), a policy might require different support factors in different contexts in order to produce the intended contribution to the effect of interest. These five factors, therefore, may be support factors for PES programs in some contexts, but not in others. The second factor—the low cost of enforcing PES programs—for instance, may not be a required support factor in contexts in which the sellers of the ES tend to abide by contracts for cultural or religious reasons.

Our framework makes it plain why these factors matter and why having evidence about their presence and distribution is crucial. If we make the unrealistic assumption that these factors are support factors always required by PES programs then, for your effectiveness prediction regarding a PES program to be properly supported by evidence, you must have evidence that these factors are present, and distributed in just the right way, in the location in which the program is to be implemented.[38] Below we list the five factors we have seen cited in the literatures about PES programs and some of the questions they immediately give rise to. But behind these there are bigger questions that need answering: 'Are these necessary in all cases?', 'What else is necessary in any particular case?', 'Will the necessary factors be in place, or can they be put in place, in the new place?', and very importantly, 'What kinds of study can help us find out the answers to these bigger questions?'

1. *Strong property rights*. A PES program, it is argued, can only be effective if there exist property rights and the means to enforce them in the location in which the program is to be implemented. There is no landowner for the ES buyer to sign a contract with if there is no landowner to start with. But how strong do these property rights need to be, and do they need to be guaranteed by a government? Where are property rights strong enough, and where are they too weak for PES programs to be effective?

2. *Low cost of monitoring and enforcing PES contracts*. If the economic and political cost of monitoring and enforcing PES contracts is high then there is an incentive for the buyer not to do so, and thus for the seller to breach the contract. These costs must be low for PES programs to be effective. But how low must they be? And how does one assess these costs?
3. *Sustainable and flexible funding source*. PES programs can only be effective, it is argued, if they are funded on the long term and if the funding source is flexible enough to allow for re-negotiation of PES contracts. If the price of timber rises, then the payment for forest conservation provided to a forest owner must rise for the incentives to stay the same, and for the forest owner to keep providing an ES. Can NGOs provide sustainable and flexible funding? What about governmental agencies in countries that are politically unstable?
4. *Absence of leakage*. If a forest owner agrees to stop her timber activity on a parcel she owns and for which the PES contract was signed, but then goes on to use the extra earnings from the contract to buy a similarly sized parcel nearby and resume her timber activity on that parcel, then the PES program is not effective in reducing deforestation and biodiversity loss. Opportunities for 'leakage' must be limited for the PES program to play the expected causal role. How does one assess opportunities for leakage?
5. *Access to credit*. If a forest owner cannot easily borrow money to cover emergency expenses (e.g., medical bills), then she might cut down and sell trees instead, even if she signed a PES contract covering those trees. An easy access to credit might thus lower the chances that forest resources will be used as a 'safety net' and thus have a bearing on the effectiveness of the PES program. But how exactly does one measure 'access to credit', and how easy must access to credit be in order for the resources covered by the PES contract to stop being a 'safety net'?

We emphasize that these are just five among the numerous factors that may be support factors required for a PES program to produce a contribution to the reduction of deforestation. The point we want to illustrate here is that J-PAL's RCT will not tell you whether these are support factors required in the location you are targeting, nor whether they are actu-

ally present there, nor how they are distributed. Unfortunately, you need this information in order to accurately predict whether a PES program will play the causal role you want it to play in the location in which you are contemplating its implementation.

15.6 Evaluating the Effects of Adaptation Policies: The Limits of RCTs

Remember that adaptation policies seek to modify natural or human systems in order to reduce their vulnerability to weather-related events due to climate change. The term 'vulnerability' has a precise meaning in this context. According to the IPCC's definition, the vulnerability of a system (usually some geographical unit, e.g., a city) to climate change is the "degree to which [it] is susceptible to, and unable to cope with, adverse effects of climate change, including climate variability and extremes" (IPCC 2007b, 883). More precisely, the vulnerability of a system is "a function of the character, magnitude, and rate of climate change and variation to which [it] is exposed, its sensitivity, and its adaptive capacity" (ibid). An adaptation policy is designed to reduce the vulnerability of a system by reducing its sensitivity—i.e., the extent to which it is harmed by climate change—or by enhancing its adaptive capacity—i.e., its ability to adjust to moderate the harmful effects of climate change. A distinction is often drawn between environmental vulnerability—as measured for instance by the country-level Environmental Vulnerability Index (EVI)—and social vulnerability—as measured for instance by one of the Social Vulnerability Indices (SoVi).[39]

There are various obstacles to the use of RCT-based IEs to evaluate the effects of adaptation policies. First, adaptation policies take a wide variety of forms, many of which simply do not lend themselves to randomization. Consider for instance the 'Adaptation to Climate Change through Effective Water Governance' policy under implementation in Ecuador that aims to improve the country's adaptive capacity by mainstreaming "climate change risks into water management practices…" (GEF 2007, 2). This policy will change water management practices in Ecuador, e.g.,

by incorporating climate risks in the country's National Water Plan. How is one to evaluate the extent to which such a policy will improve Ecuador's adaptive capacity and thus reduce its vulnerability, both environmental and social, to climate change? RCTs are no help here, given that the policy is implemented at the level of an entire country. One cannot, for a variety of reasons (political, practical, etc.), randomly assign countries to particular policy regimes.

The same point applies to the many adaptation policies that aim to improve some country's adaptive capacity, and thus reduce its vulnerability, by modifying its institutions. Here is another example. The government of Bhutan is, with the help of the United Nations Development Programme (UNDP), implementing the Reducing Climate Change-Induced Risks and Vulnerabilities from Glacial Lake Outburst Floods [GLOFs] policy which, among other things, aims to integrate the risk of GLOFs due to climate change occurring in the Punakha-Wangdi and Chamkhar valleys in Bhutan's national disaster management framework.[40] Such policies, because they target country-level institutions, cannot in practice be evaluated using RCT-based IEs. The problem here is that a vast number of adaptation policies fall into this category. Note also that such policies, by their very nature, are tailored to the institutions of a particular country and so may not be implementable in any other country. A policy that improves Bhutan's adaptive capacity, for instance, may not be applicable, and a fortiori may not have the same beneficial effects, in a country which faces similar risks but has a different institutional structure (e.g. Canada, which, unlike Bhutan, is a federal state).

Second, for many adaptation policies, RCT-based IEs are superfluous. Consider for instance the Kiribati Adaptation Program (Phase II) implemented between 2006 and 2010 that included the construction of a 500 meters long seawall to protect the country's main road, a coastal road around Christmas Island.[41] One does not need an RCT in order to determine whether this seawall is helping protect the road and reduce beach erosion (inside this wall). The physical configuration of seawalls guarantees that they will reduce the sensitivity of the systems inside them to the consequences of climate change (e.g., to rising sea levels, erosion, and extreme weather events). One might argue that an RCT would enable one to determine *by how much* the Kiribati seawall reduces the sensitivity

of the systems it helps protects, i.e. would enable one to estimate the size of the effect of this seawall on sensitivity. In this case, as with most adaptation policies, however, the need for an immediate reduction in sensitivity trumps the need for estimates of treatment effects.

One could have conducted an RCT in which the coastline along the Christmas Island road is divided into *n* sections, half of them randomly assigned to the 'seawall' group and half of them to the 'no seawall' group, and compared the condition of the road and the extent of beach erosion between sections in the 'seawall' group and those in the 'no seawall' after a year, for instance. This would have provided one with estimates of the effect of seawalls on road condition and beach erosion on Kiribati's Christmas Island (assuming both road condition and beach erosion can be reliably measured). Conducting such an RCT would make little sense for Kiribati's policy makers, however. Roads are useful only if they enable you to get somewhere, and they can only do so if they are uninterrupted and in good condition rather than irreversibly damaged at random intervals. The aim of this hypothetical example is not to caricature the position of those who, like members of the GEF's Scientific and Technical Advisory Panel (STAP 2010), call for more RCT-based IEs of adaptation and mitigation policies. It is simply to illustrate that such calls sometimes conflict with the goals the policies that are to be evaluated are supposed to achieve. What matters in the end is that these policies produce the beneficial effects they were designed to produce, not that we have highly trustworthy point estimates of the size of these effects.

This is not to say that there are no adaptation policies the effects of which can be evaluated using RCT-based IEs. Policies which offer farmers rainfall index insurance, i.e., policies that insure farmers against both deficits and excesses in rainfall, can be considered adaptation policies, and their effects on the vulnerability of particular study populations to climate change can in principle be evaluated using RCTs, even though no such RCT has been conducted to date.[42] This is true in general of adaptation policies that do not seek to reduce a country's vulnerability by modifying its institutions (e.g., by incorporating climate risks into its planning tools) or its infrastructures (e.g., by building seawalls) but rather target units (e.g., individual farmers or villages) that can more easily be randomly assigned to some treatment group. The mistake here would be

to think that such policies should occupy a privileged position in the portfolio of policies available to policy makers preoccupied with adapting to climate change simply because they can be evaluated using RCT-based IEs. As we showed in Sect. 15.5 for PES policies aiming at mitigation, the fact that a policy lends itself to randomization does not imply that it can more easily be generalized beyond the study population. And it also does not imply that this policy is more effective than other policies that cannot be similarly evaluated. A policy that offered Ugandan farmers the possibility of using drought-resistant seeds might lend itself to an RCT-based IE more easily than does Uganda's national irrigation master plan,[43] but this obviously does not mean that the former is more effective than the latter at reducing the sensitivity of Ugandan farmers to droughts due to climate change.

We showed in Sect. 15.5 that results from RCT-based IEs of mitigation policies such as PES programs provide only a small part of the total evidence needed to support effectiveness predictions. The situation is more challenging even in the case of adaptation policies, since many of these cannot be evaluated using RCTs in the first place. The lesson of this section thus is that, both for evaluating past adaptation policies and for supporting predictions regarding the effectiveness of future adaptation policies, we need more than RCTs. Nor is it especially the issue of random assignment that raises difficulties. We face here rather problems that are endemic with comparative group studies: They are often not possible and they tell us only a little of what we need to know to make use of their own results.

15.7 Conclusion

Should J-PAL scientists pack their bags and cancel the RCT they are currently carrying out in Western Uganda? No. Are RCTs a bad tool for causal inference? No. Are estimates of treatment effects irrelevant for policy making in the domain of climate change policies? No.

We want to emphasize that our criticisms are not directed at RCTs per se. Criticizing RCTs in principle makes little more sense than criticizing hammers in principle. Both RCTs and hammers are well-designed tools.

One can criticize their instances: There are bad hammers and poorly conducted RCTs. And one can criticize the use to which they are put. It is the use to which RCTs are frequently put that we target and criticize.

Calling for more and more RCTs in order to strengthen the evidence base for mitigation policies such as PES programs is a bit like calling for the use of more and more hammers in order to carve a statue out of a block of marble. What one needs is not more and more hammers, but hammers and chisels, i.e. tools of a different kind. In the policy case, what one needs is not more estimates of treatment effects produced by more RCTs. If one starts with an RCT, what one needs is evidence of a different kind, evidence that is relevant to external validity inferences, and so to prediction about the effectiveness of particular policies implemented in particular contexts. The framework sketched above in Sect. 15.4.2 tells you what kind of evidence is needed, namely evidence about causal principles and support factors.

What we advocate corresponds, to some extent, to what Pattanayak et al. (2010, 6) call "economic archeology", i.e., the qualitative evaluation of existing policies in order to reveal the contextual factors that are relevant to their effectiveness. What we argue is that calls for an improvement of the evidence base for PES programs, and mitigation and adaptation policies in general, should emphasize the need for more "economic archeology" just as much, or even more, than they emphasize the need for estimates of treatment effects generated by RCTs. This is particularly true for adaptation policies since, as we showed in Sect. 15.6, these often cannot be evaluated using RCTs. The "hard numbers" produced by RCTs—when and where they are available—are of little use for policy without knowledge of the networks of factors that give rise to these numbers, and without models of these networks (see Cartwright forthcoming). The framework sketched here, and fully developed in Cartwright and Hardie (2012), provides one with the means to do "economic archeology" where RCTs are involved in a rigorous and fruitful manner.

But it is important to stress that we do not need to start with RCTs in order to pursue economic archeology. The issue of course is how to do economic archeology in anything like a rigorous and reliable way. This involves understanding how best we can provide evidence about causal relations in the single case. So, besides a call for more and more RCTs,

surely there should be an equally urgent call for more systematic study of what counts as evidence for causality in the single case.

Acknowledgements Both authors would like to thank the Templeton Foundation's project 'God's Order, Man's Order and the Order of Nature', the UCSD Faculty Senate, and the AHRC project 'Choices of evidence: tacit philosophical assumptions in debates on evidence-based practice in children's welfare services' for support for the research and writing of this chapter. Nancy Cartwright would in addition like to thank the Grantham Research Institute on Climate Change and the Environment at LSE.

Notes

1. We use the expressions 'anthropogenic global warming' and 'climate change' interchangeably in this paper.
2. Global warming is expected to have limited positive effects, in the short run and in some regions, for instance in the domain of timber productivity (IPCC 2007b, 289). It is also the task of policy makers to design policies for taking advantages of these positive effects.
3. This distinction is reflected in the Fourth IPCC Assessment Report. This report treats of mitigation and adaptation in two distinct parts, though it contains a chapter on the relations between them (IPCC 2007b, chapter 18).
4. They also want policies that have large benefit/cost ratios. We leave aside issues related to cost-benefit analysis itself in what follows, and focus on the preliminary step to any such analysis: the evaluation of the likelihood that a policy will yield the intended benefit.
5. See http://www.thegef.org/gef/eo_office. Other funding agencies such the World Bank (http://ieg.worldbankgroup.org/), the International Monetary Fund (http://www.ieo-imf.org), or the US Food and Drug Administration (http://www.fao.org/evaluation/) also have their own evaluation offices. There are also organizations, such as the International Initiative for Impact Evaluation (3ie, http://www.3ieimpact.org/), whose sole role is to fund and carry out IEs. The multiplication of evaluation offices results in the multiplication of guidelines and methodologies for conducting IEs.
6. It is widely assumed, and not just by the World Bank, that answering a causal question about the effect of a policy just is to answer some coun-

terfactual question about what would have happened in the absence of the policy. Thus Duflo and Kremer, both members of the influential Jameel Poverty Action Lab at MIT, claim that "Any impact evaluation attempts to answer an essentially counterfactual question: how would individuals who participated in the program have fared in the absence of the program?" (Duflo and Kremer 2005, 3). And Prowse and Snilstveit, in a review of IEs of climate policies, claim that "IE is structured to answer the [counterfactual] question: how would participants' welfare have altered if the intervention had not taken place?" (Prowse and Snilstveit 2010, 233).

7. Who are sometimes called 'randomistas' as in, e.g., Ravallion et al. (2009).
8. See, e.g., Rubin (2008).
9. The terminology comes from clinical trials.
10. It also enables one to answer the question 'What would be the mean value of E for individuals (in the study population) not exposed to C were C present, all else being equal?' by citing the mean value taken by E for individuals actually exposed to C. Note that we are here talking about mean values of E over the treatment and control groups respectively and over an extended run of repeated randomizations on the study population. RCTs enable one to estimate the mean causal effect of C on E in a given population, not the individual causal effect of C on E for any specific individual in this population.
11. 'Ideal' RCTs (ones for which balance of other causes is actually achieved) are, in the words of Cartwright Hardie (2012, §I.B.5.3), 'self-validating', i.e., their very design guarantees the satisfaction of the assumptions that must be satisfied in order for the causal conclusions they yield to be true.
12. For more on RCTs and on the way they establish their conclusions, see Cartwright and Hardie (2012, §I.B.5) and Cartwright (2010).
13. We treat 'mean', 'expectation', and 'expected value' as synonyms here.
14. The probabilistic independence of X_i from b_i guarantees that the size of the effect of C on E for i is causally unrelated to whether i is assigned to the treatment or the control group. And the probabilistic independence of X_i from W_i guarantees that whether i is assigned to the treatment or control group is causally unrelated to the causes of E that do not appear in (CP).
15. For the full proof see e.g., Holland and Rubin (1987, 209–210). Essentially the same results as these hold for more complicated functional forms for (CP); we choose the linear form for ease of illustration.

16. Though this does not mean that J-PAL members only work on RCTs, it does mean that all the IEs sponsored and conducted by J-PAL take the form of RCTs.
17. There is a lot to be said about the standard view and why the labels 'internal validity' and 'external validity' are both vague and misleading. Given limitations of space, however, these issues cannot be discussed here. For more, see Cartwright and Hardie (2012, §I.B.6.3).
18. The hedge 'in principle' is important. Poorly executed RCTs will not produce unbiased estimates of treatments effects.
19. See Cartwright and Hardie (2012, op. cit.) for a concrete example of an appeal to similarity. See also http://blogs.worldbank.org/impactevaluations/impactevaluations/why-similarity-wrong-concept-external-validity
20. All the conclusions we draw below apply mutatis mutandis when the relevant causal principles take more complex forms than that of (CP) (e.g., non-linear forms).
21. You may be used to thinking of b_i as the size of the effect of X_i on Y_i. Indeed, this is the way we described it above when introducing (CP). But because, as we explain below, causes are INUS conditions, the two descriptions are equivalent: The effect of C on E just is what happens to E when C is present along with all of its required support factors.
22. Each term in an equation like (CP) represents a contribution to the effect. Mackie's original theory does not mention 'contributions' because he only consider binary 'yes-no' variables. Our presentation is more general in that it encompasses both cases in which the cause and effect variables are binary, and more common cases in which they are not.
23. As the 'short circuit' example makes evident, the distinction between policies and support factors is a pragmatic one. Both a policy and its support factors are causes, and so both are INUS conditions. Some factor is usually singled out as the policy because it is practical, ethically acceptable, or cost-efficient to manipulate it. Note also that we claim that all causes are INUS conditions, but not that all INUS conditions are causes.
24. If this estimate is equal to 0, or very close to 0, then you cannot directly draw any conclusion about the causal role played by C in the study population because you do not know whether C is ineffective or, alternatively, its positive and its negative effects balance out. We leave this case aside here.
25. See Heckman (1991) for a further critique of the limitations of RCTs when it comes to estimating parameters that are of interest for policy making.

26. Apart from giving you a trustworthy estimate of the value of Exp[b_i].
27. Banerjee and Duflo, for instance, make the following claim: "A single experiment does not provide a final answer on whether a program would universally 'work'. But we can conduct a series of experiments, differing in […] the kind of location in which they are conducted…" (Banerjee and Duflo 2012, 14). They add that "This allows us to […] verify the robustness of our conclusions (Does what works in Kenya also work in Madagascar?)…" (ibid).
28. You may think this is an uncharitable reconstruction of the argument advanced by advocates of RCTs. But the claims they sometimes make, e.g., Banerjee and Duflo's claim, quoted in note 27, regarding the need for several RCTs in order to establish that a policy works "universally", seem to invite reconstructions that are far less charitable. One could thus see advocates of RCTs as advancing an argument of the form 'If RCTs produce conclusive results in A, B, and C, then the policy works "universally", and it will therefore work in D'. This construal seems less charitable in that it attributes to advocate of RCTs a claim (the conditional in the previous sentence) that's highly likely to be false.
29. In the case of mitigation-relevant PES program, the buyer of the ES often is an intergovernmental agency, e.g., the GEF, acting as a third party on behalf of users of the ES. When the GEF is the buyer of the ES, the users it represents are the citizens of states that are members of the UN.
30. Of course, many PES programs that target biodiversity also results in the protection of carbon stocks and, conversely, many PES programs that target climate change mitigation also result in the conservation of biodiversity.
31. The theory behind PES programs comes from the work of Ronald Coase on social cost (Coase 1960). But see Muradian et al. (2010) for an alternative theoretical framework within which to understand PES programs.
32. 20 percent according to IPCC (2007a), 12 percent according to van der Werf et al. (2009).
33. The UN, for instance, is developing a program called 'REDD+' that relies on PES-type programs in order to reduce deforestation. Note that 'REDD' is an acronym for 'Reduction of (carbon) Emissions from Deforestation and forest Degradation'.
34. In the *Oportunidades* (originally PROGRESA) program, parents receive conditional payments for activities that improve human capital, e.g., enrolling their children to school. The idea is to reduce poverty both in

the short term, via the cash payments, and in the long run, by improving human capital. The payments in this program, as well as in PES programs, are conditional in that they are made only if the service (e.g. an ES) is actually provided: They are not one-time payments that are made upfront.

35. The project is supposed to last for four years, from April 2010 through April 2014.
36. And it won't tell you whether the same causal principle is at work in those parts of the study populations composed of landowners from the Hoima district and those parts composed of landowners the Kibaale districts.
37. See e.g., Pattanayak et al. (2010), Pirard et al. (2010), Alix-Garcia et al. (2009), GEF (2010, 35), or Jayachandran (2013b).
38. And if the assumption that these factors are always required is dropped, then you also need evidence that these factors are indeed support factors needed for the PES program to produce the intended contribution to the effect in the location you are targeting.
39. See http://www.vulnerabilityindex.net/ for the EVI and http://webra.cas.sc.edu/hvri/ for the US county-level SoVI. Note two difficulties with using these indices to evaluate the effects of adaptation policies. First, they are measures of vulnerability to environmental hazards in general, whether or not they are due to climate change. Second, there is no wide consensus as to how to measure overall vulnerability (at various geographical scales), and neither is there a consensus regarding how to measure an important component of vulnerability, namely adaptive capacity.
40. Seehttp://www.adaptationlearning.net/bhutan-reducing-climate-change-induced-risks-and-vulnerabilities-glacial-lake-outburst-floods-punakh
41. Seehttp://www.thegef.org/gef/greenline/july-2012/preparation-adaptation-and-awareness-kiribati%E2%80%99s-climate-challenge
42. RCTs conducted about weather insurance usually attempt to estimate the effects of such insurance on investment decisions (see e.g., Giné and Yang 2009) or to understand the causes of weather insurance take-up (see e.g., Cole et al. 2013). See de Nicola (2015) for a non-randomized evaluation of the effects of rainfall index insurance on the welfare of farmers and so on their adaptive capacity.
43. See www.mwe.go.ug

References

Alix-Garcia, Jennifer, Alain De Janvry, Elisabeth Sadoulet, and Juan Manuel. 2009. Lessons Learned from Mexico's Payment for Environmental Services Program. In *Payment for Environmental Services in Agricultural Landscapes*, 163–188. Rome: Springer.

Balshem, Howard, Mark Helfand, Holger J. Schünemann, Andrew D. Oxman, Regina Kunz, Jan Brozek, Gunn E. Vist, et al. 2011. GRADE Guidelines: 3. Rating the Quality of Evidence. *Journal of Clinical Epidemiology* 64 (4): 401–406. https://doi.org/10.1016/j.jclinepi.2010.07.015.

Banerjee, Abhijit, and Esther Duflo. 2012. *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*. New York: PublicAffairs.

Cartwright, Nancy. 2010. What Are Randomised Controlled Trials Good For? *Philosophical Studies* 147 (1): 59–70.

———. Forthcoming. Will Your Policy Work? Experiments vs. Models. In *The Experimental Side of Modeling*, ed. I.F. Peschard and B.C. van Frassen.

Cartwright, Nancy, and Jeremy Hardie. 2012. *Evidence-Based Policy: A Practical Guide to Doing It Better*. Oxford: Oxford University Press.

Coase, Ronald Harry. 1960. The Problem of Social Cost. *The Journal of Law and Economics* 3: 1–44.

Cole, Shawn, Xavier Giné, Jeremy Tobacman, Petia Topalova, Robert Townsend, and James Vickery. 2013. Barriers to Household Risk Management: Evidence from India. *American Economic Journal: Applied Economics* 5 (1): 104–135.

de Nicola, Francesca. 2015. The Impact of Weather Insurance on Consumption, Investment, and Welfare. *Quantitative Economics* 6 (3): 637–661.

der Werf, Van, R. Guido, Douglas C. Morton, Ruth S. DeFries, Jos G.J. Olivier, Prasad S. Kasibhatla, Robert B. Jackson, G. James Collatz, and James T. Randerson. 2009. CO_2 Emissions from Forest Loss. *Nature Geoscience* 2: 737–738.

Duflo, Esther, and Michael Kremer. 2005. Use of Randomization in the Evaluation of Development Effectiveness. *Evaluating Development Effectiveness* 7: 205–231.

Engel, Stefanie, Stefano Pagiola, and Sven Wunder. 2008. Designing Payments for Environmental Services in Theory and Practice: An Overview of the Issues. *Ecological Economics* 65 (4): 663–674.

GEF. 2007. Adaptation to Climate Change through Effective Water Governance in Ecuador. Project Executive Summary.

———. 2010. *Developing an Experimental Methodology for Testing the Effectiveness of Payments for Ecosystem Services to Enhance Conservation in Productive Landscapes in Uganda (Request for CEO Endorsement/Approval)*. Washington, DC: Global Environment Facility.

———. 2016. Developing an Experimental Methodology for Testing the Effectiveness of Payments for Ecosystem Services to Enhance Conservation in Productive Landscapes in Uganda. Washington, DC: Global Environmental Facility, June 4. https://www.thegef.org/project/developing-experimental-methodology-testing-effectiveness-payments-ecosystem-services

Giné, Xavier, and Dean Yang. 2009. Insurance, Credit, and Technology Adoption: Field Experimental Evidence from Malawi. *Journal of Development Economics* 89 (1): 1–11.

Heckman, James J. 1991. *Randomization and Social Policy Evaluation*. National Bureau of Economic Research Cambridge, MA. http://www.nber.org/papers/t0107

Holland, Paul W. 1986. Statistics and Causal Inference. *Journal of the American Statistical Association* 81 (396): 945–960.

Holland, Paul W., and Donald B. Rubin. 1987. Causal Inference in Retrospective Studies. *ETS Research Report Series* 1987 (1): 203–231. https://doi.org/10.1002/j.2330-8516.1987.tb00211.x.

IPCC. 2007a. *Climate Change 2007: The Physical Science Basis*. New York: Intergovernmental Panel on Climate Change.

———. 2007b. *Impacts, Adaptation and Vulnerability*. New York: Intergovernmental Panel on Climate Change.

———. 2012. *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation: Special Report of the Intergovernmental Panel on Climate Change*. New York: Cambridge University Press.

Jayachandran, S. 2013a. Evaluating a Payments for Ecosystem Services Program in Uganda, April 22. http://www.climate-eval.org/?q=print/2235

Jayachandran, Seema. 2013b. Liquidity Constraints and Deforestation: The Limitations of Payments for Ecosystem Services. *The American Economic Review* 103 (3): 309–313.

Jones, Harry. 2009. *The 'Gold Standard' is Not a Silver Bullet for Evaluation*. Overseas Development Institute London. http://www.alnap.org/pool/files/3695.pdf

Mackie, J. 1965. Causes and Conditions. *American Philosophical Quarterly* 2: 245–264.

MEA. 2005. *Ecosystems and Human Well-Being: Current State and Trends*. Washington, DC: Millennium Ecosystem Assessment.

Muradian, R., E. Corbera, U. Pascual, N. Kosoy, and P.H. May. 2010. Reconciling Theory and Practice: An Alternative Conceptual Framework for Understanding Payments for Environmental Services. *Ecological Economics* 69 (6): 1202–1208.

Parker, Susan W., and Graciela M. Teruel. 2005. Randomization and Social Program Evaluation: The Case of Progress. *The Annals of the American Academy of Political and Social Science* 599 (1): 199–219. https://doi.org/10.1177/0002716205274515.

Pattanayak, Subhrendu K., Sven Wunder, and Paul J. Ferraro. 2010. Show Me the Money: Do Payments Supply Environmental Services in Developing Countries? *Review of Environmental Economics and Policy* 4 (2): 254–274. https://doi.org/10.1093/reep/req006.

Pirard, Romain, Raphaël Billé, and Thomas Sembrés. 2010. Questioning the Theory of Payments for Ecosystem Services (PES) in Light of Emerging Experience and Plausible Developments. *Institut Pour Le Développement Durable et Les Relations Internationales (IDDRI)*, Analyses, 4, no. 2010/06/10, 5–22.

Prowse, Martin, and Birte Snilstveit. 2010. Impact Evaluation and Interventions to Address Climate Change: A Scoping Study. *Journal of Development Effectiveness* 2 (2): 228–262.

Ravallion, Martin, et al. 2009. Should the Randomistas Rule? *Economists' Voice* 6 (2): 1–5.

Rubin, Donald B. 1974. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology* 66 (5): 688–701.

———. 2008. Comment: The Design and Analysis of Gold Standard Randomized Experiments. *Journal of the American Statistical Association* 103: 1350–1353.

STAP. 2010. *Payments for Environmental Services and the Global Environment Facility: A STAP Advisory Document*. Washington, DC: Scientific and Technical Advisory Panel. https://www.thegef.org/sites/default/files/publications/STAP_PES_2010_1.pdf.

STAP. n.d. *Payments for Environmental Services and the Global Environment Facility: A STAP Advisory Document*. Washington, DC: Scientific and Technical Advisory Panel.

Index[1]

[1] Note: Page number followed by 'n' refers to notes.

E.A. Lloyd, E. Winsberg (eds.), *Climate Modelling*,
https://doi.org/10.1007/978-3-319-65058-6

D